Formalisation des raisonnements éthiques : modélisation des processus en éthique et modélisation, représentation et automatisation du raisonnement causal

Camilo Sarmiento Lozano

École doctorale nº 130 : Informatique, Télécommunications et Électronique

Doctorat Sorbonne Université

THÈSE

pour obtenir le grade de docteur délivré par

Sorbonne Université

Spécialité doctorale "Sciences et technologies de l'information et de la communication"

soutenue publiquement par

Camilo SARMIENTO LOZANO

le 26 avril 2024

Formalisation des raisonnements éthiques : modélisation des processus en éthique et modélisation, représentation et automatisation du raisonnement causal

Directeur de thèse : **Dr. HDR. Gauvain BOURGNE**
Directeur de thèse : **Pr. Jean-Gabriel GANASCIA**

Composition du jury

Dr. HDR. Gregory BONNET,	Université de Caen Normandie	Rapporteur
Dr. HDR. Florence DUPIN DE SAINT-CYR,	Université Paul Sabatier	Rapporteur
Pr. Isabelle BLOCH,	Sorbonne Université	Présidente
Pr. Jérôme LANG,	Université Paris-Dauphine	Examinateur
Pr. Catherine TESSIER,	ONERA	Examinateur
Dr. HDR. Gauvain BOURGNE,	Sorbonne Université	Directeur
Pr. Jean-Gabriel GANASCIA,	Sorbonne Université	Directeur

Sorbonne Université, Faculté de Sciences
Laboratoire de recherche en informatique de Sorbonne Université
UMR7606 CNRS, 4 Pl. Jussieu, 75005 Paris France

Résumé

Cette thèse s'inscrit dans le domaine de l'éthique computationnelle dont le but est de formaliser le raisonnement éthique. Autrement dit, ce travail fait partie du domaine qui cherche à reproduire notre capacité en tant qu'êtres rationnels à évaluer moralement une action. Deux objectifs sont recherchés au travers de la formalisation de ce raisonnement : mieux le comprendre et l'intégrer dans des systèmes informatiques de sorte à assurer que les décisions prises soient conformes à des principes moraux choisis.
Cette thèse contribue à ce domaine de deux façons. Un premier groupe de contributions est lié à la proposition d'un cadre commun permettant de formaliser de façon fidèle les principes moraux les plus courants dans la philosophie occidentale. Pour reprendre les termes du titre, ce premier groupe de contributions peut être résumé en quelques mots comme la « modélisation des processus en éthique ». Le deuxième groupe de contributions est lié à la proposition d'une formalisation du raisonnement causal. En plus de permettre une meilleure compréhension de ce raisonnement, cette formalisation rend possible de l'intégrer dans des systèmes informatiques de sorte à pouvoir établir des relations causales complexes. Cette capacité intervient dans la formalisation d'un grand nombre de principes moraux. Ayant pour objectif que notre proposition puisse être utilisée dans la formalisation de tous ces principes moraux, nous l'avons conçue de sorte à ce qu'elle satisfasse un certain nombre de conditions. Tout d'abord, notre formalisation repose sur un formalisme permettant de rendre explicites la plupart des subtilités des problèmes, aussi bien pour le raisonnement causal que le raisonnement éthique. Ensuite, la définition de causalité intégrée dans notre formalisme est dépourvue de toute confusion avec la notion de responsabilité. Sans cela, elle ne pourrait pas être commune à la formalisation de tous les principes moraux. Finalement, notre proposition est en capacité de traiter tous les cas causaux, dont les plus complexes. Pour reprendre les termes du titre, ce deuxième groupe de contributions peut être décrit comme la « modélisation, représentation et automatisation du raisonnement causal ». Les contributions principales de cette thèse appartiennent à ce deuxième groupe.

Abstract

This thesis is in the field of computational ethics, which aims to formalise ethical reasoning. In other words, this work is part of the field that seeks to emulate our capacity as rational beings to morally evaluate an action. The formalisation of this reasoning has two objectives : to better understand it and to integrate it into computer systems to ensure that decisions made comply with chosen moral principles.
This thesis makes a contribution to the field in two ways. Firstly, it proposes a common framework for formalising faithfully the most common moral principles in Western philosophy. This first contribution can be summarised as 'modelling ethical processes'. The second set of contributions pertains to the proposal for formalising causal reasoning. This formalisation not only enhances our comprehension of this reasoning but also enables its integration into computer systems, facilitating the establishment of complex causal relationships. This capability is crucial for formalising a wide range of moral principles. To ensure that our proposal can formalise all these moral principles, we have designed it to satisfy a number of conditions. Firstly, our formalisation is based on a formalism that explicitly addresses the subtleties of problems related to both causal and ethical reasoning. Secondly, our formalism's definition of causality free of any confusion with the notion of responsibility. Otherwise, it would not be common to formalise all moral principles. Finally, our proposal can handle all causal cases, including the most complex. The second group of contributions focuses on 'modelling, representing and automating causal reasoning'. The main contributions of this thesis belong to this second group.

Table des matières

Introduction

> « *[...] the greater the freedom of a machine, the more it will need moral standards.* »
>
> PICARD [1997]

Il n'est pas rare d'entendre que les titres de thèse sont abscons et qu'ils ne sont compréhensibles que pour l'auteur du document. Notre objectif est qu'à la fin de cette introduction le titre de cette thèse n'ait plus aucun secret pour le lecteur. Mieux encore, une fois la lecture de ces quelques pages terminée, l'idée est que le lecteur comprenne comment cette thèse est structurée et qu'il ait une idée précise des arguments qui y sont développés. Pour ce faire, nous avons décomposé le titre de cette thèse en trois parties. Chacune de ces parties correspond au titre d'une des sections de l'introduction. Dans celles-ci nous expliquons les termes utilisés et les principaux arguments développés dans cette thèse.

Formalisation des raisonnements éthiques

Plus qu'au travail réalisé, la première partie du titre fait référence au cadre dans lequel s'inscrit cette thèse. Elle permet d'indiquer le but dans lequel ce travail est réalisé et le domaine de recherche dans lequel il s'inscrit. En choisissant les termes « formalisation des raisonnements éthiques », nous avons voulu nous placer dans le domaine qui cherche à formaliser notre capacité en tant qu'êtres rationnels à évaluer moralement une action. Nous appelons ce domaine l'éthique computationnelle. Étudions plus en détail les termes choisis.

Formalisation Lorsque nous utilisons le terme « formalisation », nous adoptons la définition donnée par SAINT-CYR et collab. [2014]. Il est question d'un processus en trois étapes : modéliser, représenter et automatiser. L'étape de modélisation consiste à définir rigoureusement les concepts et les processus d'un point de vue mathématique. L'étape de représentation consiste à indiquer comment les concepts doivent être codés pour être traités par un ordinateur. Finalement, l'étape d'automatisation consiste à indiquer comment les processus sont reproduits par des algorithmes. Comme nous le verrons par la suite, la structure adoptée pour présenter les contributions de cette thèse s'appuie sur cette décomposition.

Deux objectifs peuvent être recherchés lors de la formalisation d'un processus : analyser ce dernier afin de mieux le comprendre, ou simplement le reproduire pour qu'il puisse être réalisé par un ordinateur. Nous sommes concernés aussi bien par le premier que le deuxième objectif lorsque nous travaillons à formaliser le raisonnement éthique.

Raisonnement éthique Par raisonnement éthique nous entendons la capacité que nous avons à déterminer s'il est juste ou non qu'une décision soit prise par un agent. Cette évaluation n'est pas triviale, la façon dont il faut procéder a fait, et fait toujours, l'objet de nombreux débats au sein de la discipline philosophique qu'est l'éthique.
Tout au long de l'histoire de la philosophie, de nombreux penseurs ont proposé différentes théories de comment cette évaluation devait être faite. Toutes ces théories dites « morales » partagent un même objectif : procurer une procédure fiable de décision permettant aux agents rationnels et bien informés de produire des verdicts moraux corrects. Concrètement, évaluer moralement une décision revient à déterminer son statut déontique, i.e. si l'action est juste ou non. Si elles partagent le même objectif, déterminer ce qu'il est juste de faire, ces théories ne sont par contre pas d'accord sur le moyen d'y parvenir. Ce qui différencie les théories morales entre elles est la façon dont sont structurées les notions de juste et de bien. Certaines définissent directement ce qui est juste sans faire intervenir les notions de bien. D'autres définissent en premier ce qui est bien, donc ce qui est considéré comme ayant de la valeur intrinsèquement, puis à partir de cela définissent ce qui est juste. Dans le chapitre 1 nous décrivons plus en détail les différentes structures qui existent et nous présentons quelques théories morales. Le choix des théories morales présentées en détail dans le chapitre a été fait de façon à donner l'aperçu le plus large possible des structures existantes.

Pour bien comprendre certains choix réalisés dans cette thèse, il est important de soulever un dernier point qu'ont en commun toutes les théories morales : elles sont toutes sensibles au contexte. Autrement dit, le statut déontique d'une décision dépend en partie de faits non moraux reliés au contexte. Ces faits non moraux peuvent aussi bien être reliés aux agents qu'aux circonstances. Par conséquent, à l'heure d'être appliquées, toutes les théories morales requièrent que l'agent dispose d'une bonne capacité à identifier le contexte propre à chaque cas, autrement dit, les subtilités du problème éthique.

Éthique computationnelle L'éthique computationnelle est le sous-domaine de l'intelligence artificielle dont l'objectif est de formaliser le raisonnement éthique. Plus précisément, il se préoccupe d'assurer que des systèmes informatiques qui automatisent au moins une partie d'un processus décisionnel prennent des décisions qui puissent être vues comme éthiques par les humains.
L'émergence et le développement de ce domaine est principalement dû à la prolifération d'outils du quotidien utilisant des résultats obtenus dans la discipline scientifique qu'est l'intelligence artificielle. Molnar et Gill [2018] décrivent les systèmes de décision comme un type particulier de technologie visant à assister ou à remplacer le jugement d'un être humain. La création de ces systèmes se fait grâce à des disciplines comme les statistiques, la linguistique et l'informatique, et s'appuie sur des techniques telles que les systèmes à base de règles, la régression et l'apprentissage machine, dont les réseaux de neurones et l'apprentissage profond. Comme l'explique O'Neil [2018], une partie de ces techniques mathématiques, auparavant utilisées principalement sur les marchés financiers, sont de plus en plus utilisées dans des applications directement liées à l'individu. Qui plus est, Buolamwini et Gebru [2018] soulèvent que ces applications sont utilisées dans des processus de décision impliquant de plus en plus de responsabilités. En l'occurrence, ces dernières années, nous avons pu voir l'apparition d'algorithmes déterminant le candidat le plus adapté à un poste, l'attribution ou non d'un prêt à un individu, le meilleur traitement pour un pa-

tient, la tendance de récidive d'un individu condamné par la justice et nous avons même vu en Chine la mise en place d'un système de crédit social qui, pour résumer, permet de classer les habitants comme « bons » ou « mauvais » citoyens au moyen d'un déploiement massif de dispositifs de surveillance [CITRON et PASQUALE, 2014]. Cette diffusion, couplée à des discours avertissant sur certains dangers associés à leur utilisation, a suscité une demande croissante pour des outils « dignes de confiance », i.e. offrant des garanties selon lesquelles les décisions prises sont transparentes, explicables et éthiques. L'éthique computationnelle a pour objectif de répondre à cette dernière garantie : que les décisions puissent être considérées comme éthiques.

Pour accomplir cette tâche, l'éthique computationnelle cherche à intégrer dans la prise de décision les principes moraux sur lesquels reposent les différentes théories morales en philosophie. En quelque sorte, il est question de concevoir un superviseur éthique qui assure que les décisions prises par un agent soient conformes aux principes moraux choisis. Dans cette thèse nous utilisons le terme agent pour faire référence à un objet technique qui prend en compte l'information qu'il reçoit de son environnement et qui peut agir pour le modifier ; il n'est pas utilisé dans le sens de la philosophie de l'action [SCHLOSSER, 2019] où il est lié au principe d'intentionnalité et de libre arbitre. MOOR [2006] présente quatre niveaux permettant de classifier les agents selon leur intégration de facteurs éthiques :

1. agents avec impact éthique : dont les décisions peuvent être évaluées éthiquement ;
2. agents éthiques implicites : auxquels des considérations éthiques ont été intégrées lors de leur conception en prenant en compte l'environnement où ils sont censés évoluer ;
3. agents éthiques explicites : dotés d'une capacité à reproduire le raisonnement éthique ;
4. agents pleinement éthiques : capables de réaliser des jugements moraux explicites et de les justifier. Ce niveau requiert que les agents soient dotés d'une conscience, d'intentionnalité et de libre arbitre.

Seuls les trois premiers niveaux décrivent un agent comme nous l'entendons dans cette thèse. L'éthique computationnelle peut être vue comme le sous-domaine de l'intelligence artificielle qui cherche à concevoir des agents qui atteignent le troisième niveau, i.e. des agents éthiques explicites.

Pour concevoir des agents éthiques explicites trois types d'approches peuvent être utilisées : descendantes, ascendantes et hybrides. Les approches descendantes sont celles qui correspondent le plus à notre volonté de formaliser le raisonnement éthique, aussi bien pour le reproduire que pour mieux le comprendre. De ce fait, dans cette thèse nous allons exclusivement nous intéresser aux approches descendantes. Plus exactement, nous allons principalement étudier des approches qui formalisent des théories morales. Nous parlons alors d'un sous-domaine de l'éthique computationnelle que nous appelons éthique computationnelle normative. Nous détaillons ce choix et donnons quelques raisons supplémentaires dans le chapitre 1.

Parmi les approches en éthique computationnelle normative existantes, le travail que nous proposons ici s'inscrit dans l'esprit des travaux de BERREBY et collab. [2015, 2017, 2018]; BOURGNE et collab. [2021] qui ont proposé le superviseur éthique modulaire ACE (Action-Causality-Ethics) dont la structure est illustrée sur la figure 1. Toutefois, la contribution de cette thèse n'est pas de proposer un superviseur éthique. La contribution exacte de cette thèse est l'objet des deux sections suivantes.

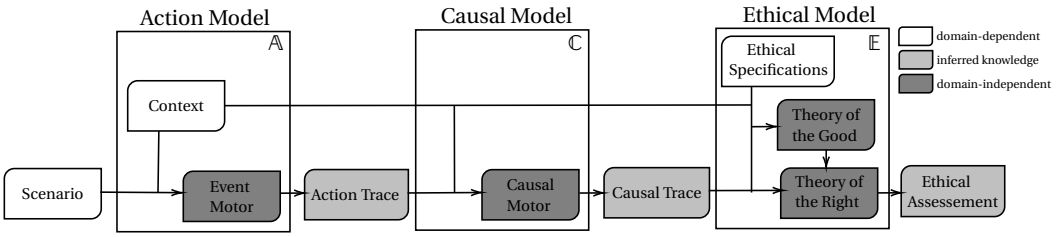


FIGURE 1 – Cadre modulaire 𝔸ℂ𝔼 pour formaliser le raisonnement éthique [BOURGNE et collab., 2021].

Modélisation des processus en éthique

Les contributions de cette thèse peuvent être classées en deux groupes : celles intrinsèquement reliées à l'éthique computationnelle et celles qui y contribuent mais qui peuvent aller au delà de ce domaine. La deuxième partie du titre fait référence au premier groupe de contributions. L'objectif de ces contributions est d'apporter des clarifications qui pourront servir à l'avancement du domaine. En quelques mots, nous avons étudié les théories morales proposées en philosophie pour identifier les différents processus qui sont utilisés dans le raisonnement éthique. En choisissant les termes « modélisation des processus en éthique », nous avons voulu appuyer sur cette idée de clarification. Par « modélisation » nous entendons la définition rigoureuse des processus en éthique d'un point de vue mathématique. Plus que de proposer un superviseur éthique, ce qui reviendrait à formaliser chacun de ces processus et donc d'en proposer une interprétation personnelle, l'objectif de cette thèse est de fournir un cadre commun permettant de faire cette formalisation tout en restant le plus fidèle possible à la théorie morale qui est formalisée. De par notre volonté de faire de ce cadre un outil commun, l'étape de modélisation est idéale car elle ne demande pas de faire des choix sur le langage informatique à utiliser. Nous restons dans un langage mathématique commun à toutes les représentations.

Comme introduit précédemment, l'éthique computationnelle normative a pour objectif de formaliser le raisonnement éthique en vue d'intégrer des principes éthiques dans des systèmes de prise de décision. Ces principes sont généralement issus de la philosophie. Toutefois, le langage utilisé en philosophie n'est pas le même qu'en informatique. De fait, intégrer ces principes philosophiques nécessite de les traduire vers une représentation informatique formelle. Cette traduction n'a d'intérêt que si elle reste fidèle à l'essence du principe que nous cherchons à traduire, mais cette fidélité est difficile à évaluer.

Le cadre que nous proposons a pour objectif de dépasser la difficulté d'obtenir une traduction fidèle. Sa conception peut être décomposée en trois étapes. La première est de nous plonger dans la littérature philosophique et d'en extraire ce qui est essentiel pour pouvoir formaliser les théories morales de façon fidèle. La deuxième consiste à modéliser les concepts communs à toutes les théories morales de façon suffisamment générale pour englober les formalisations déjà proposées dans le domaine. Finalement, la troisième étape a pour rôle de modéliser les différentes théories morales en utilisant une architecture modulaire qui permette d'identifier clairement tous les processus y intervenant. Nous distinguons d'un côté le squelette clairement défini des théories, de l'autre les processus qui ne le sont pas. De cette façon, pour s'assurer d'être fidèle à une théorie donnée, toute formalisation peut reprendre la modélisation du squelette que nous proposons, puis proposer une version

des processus indéfinis. En plus d'aider à la fidélité de la formalisation, nous souhaitons que ce cadre facilite la comparaison entre les différentes formalisations d'une théorie. Nous proposons une première étude comparative de travaux en éthique computationnelle en nous appuyant sur notre cadre pour la structurer. Les trois étapes nécessaires à la conception de ce cadre sont distribuées en deux chapitres. Le chapitre 1 contient les résultats de la première étape, celle consistant à extraire l'essentiel dans la littérature philosophique. Le chapitre 4 réuni la présentation de la deuxième et de la troisième étape, i.e. la modélisation des concepts communs et des différentes théories, ainsi que l'étude comparative réalisée.

L'architecture modulaire par processus adoptée pour la modélisation des théories morales fait apparaître que parmi les nombreux processus intervenant dans le raisonnement éthique, certains se retrouvent dans plusieurs théories morales. En faisant l'étude comparative, il ressort que parmi ces processus, celui permettant de réaliser un raisonnement causal est celui utilisé dans le plus de formalisations différentes de théories morales. Par conséquent, s'intéresser à la causalité est nécessaire à la formalisation de la plupart des théories morales, et donc à l'éthique computationnelle.

L'importance de la causalité paraît évidente lorsqu'il est question de modéliser des théories morales conséquentialistes, i.e. qui adhèrent à l'idée que la valeur des conséquences associées à une décision est la seule chose à prendre en compte pour déterminer si une action est juste. Mais, la causalité peut également être nécessaire lorsqu'il s'agit de déterminer si une conséquence est un moyen pour une fin, comme dans la deuxième formulation de l'Impératif Catégorique de Kant, ou pour déterminer si une conséquence est intentionnelle, comme dans la doctrine du double effet essentielle pour la théorie du droit naturel. Finalement, bien que le lien semble moins intuitif, la causalité peut intervenir dans des théories qui utilisent des codes de conduite car au sein de ces codes il est possible de trouver des normes morales de tout type. En particulier, il est tout à fait envisageable d'avoir des normes qui demandent de considérer les conséquences des décisions.

Toutefois, il n'est pas suffisant de prendre n'importe quelle approche pour traiter la causalité, il est nécessaire de pouvoir la traiter dans toute sa complexité. En effet, une partie des limites actuelles des approches en éthique computationnelle est due à l'absence systématique d'un processus permettant d'établir des relations causales complexes. Une telle absence a deux conséquences principales sur les propositions faites jusqu'à présent dans le domaine : la simplification excessive de la représentation des problèmes qui peut fausser le raisonnement éthique ; l'incapacité à traiter des problèmes complexes causalement. Ainsi, les approches en éthique computationnelle ont besoin de pouvoir établir des relations causales complexes. Ce besoin est discuté plus en détail dans les chapitres 3 et 4. Étant donné l'importance de ce processus, nous avons décidé d'en faire le cœur de cette thèse. Le deuxième groupe de contributions de cette thèse lui est dédié.

Modélisation, représentation et automatisation du raisonnement causal

La troisième et dernière partie du titre fait référence au deuxième groupe de contributions. L'objectif de ces contributions est de proposer une approche causale qui puisse servir de base pour tout raisonnement causal en éthique computationnelle. Autrement dit, nous proposons une formalisation du processus permettant d'établir des relations causales complexes. En choisissant les termes « modélisation, représentation et automatisation du rai-

sonnement causal », nous avons voulu insister sur la différence avec le premier groupe de contributions. Il aurait été possible de simplement parler de « formalisation du raisonnement causal ». Cependant, en faisant apparaître les trois étapes de la formalisation nous insistons sur le fait que le raisonnement causal est bien le cœur de cette thèse. Nous allons bien plus loin dans la formalisation du raisonnement causal que celle du raisonnement éthique. En effet, nous faisons des choix importants sur la modélisation, la représentation et l'automatisation du raisonnement causal. Ces choix répondent à plusieurs problématiques que nous avons identifiées et que nous présentons dans la suite de cette section. Ces problématiques proviennent aussi bien du domaine qui s'intéresse au raisonnement causal qu'à celui de l'éthique computationnelle. Avant de rentrer dans ces détails, clarifions ce à quoi nous faisons référence lorsque nous parlons de causalité.

Raisonnement causal Lorsque dans cette thèse nous parlerons de causalité, nous ferons référence à la causalité effective et non à la causalité générale, toutes deux détaillées ci-dessous. Puis, lorsque nous ferons référence au domaine de la causalité, il s'agira de l'ensemble de la communauté pluridisciplinaire s'intéressant à la causalité. Cela inclut des philosophes, des juristes, des psychologues, des mathématiciens, des physiciens et des informaticiens.
La causalité générale peut être vue comme la découverte de lois régissant le monde dans lequel nous vivons. C'est en quelque sorte ce que la science cherche à déterminer. Par exemple, la seconde loi de Newton permet d'établir la relation qui existe entre l'accélération d'un objet et sa masse. Une fois établies, ces lois nous permettent non seulement de mieux comprendre le monde, mais aussi de prendre des décisions en aillant une meilleure idée des conséquences que celles-ci peuvent avoir. La connaissance de ces lois ouvre la porte à un raisonnement a priori. Dans le cadre de cette thèse nous sommes des utilisateurs de la causalité générale car elle est indispensable pour représenter le monde et son évolution. Toutefois, nous n'y contribuons pas. Tout au long de cette thèse nous considérons que les lois causales qui régissent le monde sont déterministes et que la rétrocausalité n'est pas possible [Hall et Paul, 2003], i.e. nous considérons qu'une cause précède toujours sa conséquence.
La causalité effective peut être vue comme la détermination des facteurs présents dans une situation donnée et qui ont produit une conséquence. Autrement dit, il s'agit de déterminer qu'elle partie d'une loi causale a été appliquée dans un cas précis. Par exemple, le fait qu'une pierre ait percuté une bouteille posée sur une table peut être reliée à la façon dont Suzy l'a lancée, utilisant en partie la seconde loi de Newton. Pouvoir établir une relation causale dans des situations précises nous permet par exemple d'évaluer la responsabilité juridique d'un individu par rapport à un préjudice, ou le statut déontique d'une décision dans certaines théories morales. Cette capacité peut être assimilée à un raisonnement a posteriori. Lorsque nous nous fixons de proposer une approche causale qui puisse servir comme base pour tout raisonnement causal en éthique computationnelle, nous parlons de causalité effective.
Aussi surprenant que cela puisse paraître, malgré l'omniprésence de cette notion dans notre quotidien et le nombre de travaux pluridisciplinaires s'étant penchés sur la question, il n'y a pas de consensus sur une définition de ce qu'est une cause effective. Deux types d'approches principales existent aujourd'hui, toutes deux ont pour origine la conception de la causalité telle qu'introduite par Hume [1748]. D'un côté, nous trouvons des approches dites

« contrefactuelles » qui placent la nécessité de la cause par rapport à la conséquence comme l'élément central [MENZIES et BEEBEE, 2020] ; pour ces approches une cause est quelque chose qui fait la différence. De l'autre, nous trouvons des approches dites « par régularité » ou « par inférence » qui font intervenir la nécessité, mais la placent au second plan, derrière la suffisance [ANDREAS et GUENTHER, 2021] ; pour ces approches une cause est surtout quelque chose qui a participé à amener une conséquence. Nous ne rentrerons pas plus dans les détails dans cette introduction, et le laissons pour le chapitre 3.

Problématiques

Revenons à présent aux problématiques qui ont encouragé et guidé la conception de notre approche causale pouvant servir comme base à tout raisonnement causal dans le domaine de l'éthique computationnelle. Nous parlerons d'« approche commune » pour nous y référer plus succinctement. Comme mentionné précédemment, ces problématiques proviennent aussi bien du domaine qui s'intéresse au raisonnement causal qu'à celui de l'éthique computationnelle. Aucune approche existante en causalité effective ne permet de répondre à ces problématiques dans leur l'intégralité. L'objectif principal de cette thèse est de proposer une approche qui le fasse.

Représentation Commençons par une problématique qui est commune à la causalité et à l'éthique computationnelle, le formalisme choisi pour représenter le monde et son évolution. Plus tôt dans l'introduction, nous avons mentionné que l'éthique était sensible au contexte, i.e. que le statut déontique d'une action dépend en partie de faits non moraux reliés au contexte. Formaliser le raisonnement éthique en vue de concevoir un agent éthique explicite demande alors nécessairement de le doter d'une représentation du contexte et de la façon dont celui-ci évolue en fonction des décisions qu'il prend [GIPS, 1995]. Cette représentation doit respecter certaines exigences propres à l'éthique. En effet, les subtilités de chaque problème peuvent être décisives dans l'évaluation éthique. Pour que la formalisation du raisonnement éthique puisse être réellement utile, il est nécessaire de pouvoir représenter ces subtilités. Mais cette sensibilité au contexte n'est pas une spécificité de l'éthique, le raisonnement causal est aussi sensible à la façon dont le problème est représenté. Il est donc indispensable que l'approche commune que nous voulons concevoir utilise un formalisme qui permette de rendre explicites toutes ces subtilités, aussi bien pour le raisonnement causal que le raisonnement éthique.

Nous considérons que toute approche causale peut être décomposée en deux parties : la définition de la causalité et le formalisme utilisé pour représenter cette définition. Cette séparation nous permet de ne pas commettre l'erreur de confondre les problématiques qui relèvent de la définition de causalité avec celles qui relèvent du formalisme utilisé. Les formalismes classiques utilisés pour représenter le contexte dans le raisonnement causal ne sont pas très expressifs, ils ne permettent pas de saisir les subtilités nécessaires au raisonnement qui nous intéresse. Ils ne sont donc pas satisfaisants pour notre approche commune. Ce point fait l'objet d'une discussion plus approfondie dans le chapitre 3.
Pour trouver des formalismes plus expressifs il faut s'intéresser au domaine de la représentation de l'action et du changement, un sous-domaine de l'intelligence artificielle. Décrire des changements causés par l'exécution d'actions est un des premiers problèmes que les chercheurs en intelligence artificielle ont essayé de résoudre. Le chapitre 2 y est dédié. Des tra-

vaux récents ont proposé des approches causales reposant sur des représentations venant de ce sous-domaine [BATUSOV et SOUTCHANSKI, 2018; BERREBY et collab., 2018; BOURGNE et collab., 2021; LEBLANC et collab., 2019]. Toutefois, comme nous le montrons dans le chapitre 3, la plupart des formalismes utilisés n'ont pas l'expressivité suffisante permettant de traiter certains cas de causalité, et ceux qui l'ont sont difficilement automatisables.

En supposant que nous arrivions à trouver un formalisme dans le juste milieu, un autre problème se pose. Ces formalismes intègrent bien des notions de causalité, mais il s'agit de causalité générale. En effet, ils utilisent la notion pour déterminer les effets des décisions et donc pouvoir simuler l'évolution du monde. Cependant, comme mentionné précédemment, en éthique comme en droit nous avons besoin d'un autre raisonnement, nous cherchons les causes d'un état du monde, un raisonnement a posteriori. Nous avons donc besoin d'intégrer dans ces formalismes un raisonnement permettant de faire de la causalité effective. L'approche commune proposée pour cela doit être en capacité de gérer toutes les situations, y compris les plus complexes causalement. Les trois problématiques suivantes correspondent chacune à un de ces problèmes complexes à traiter en causalité.

Surdétermination Les cas de surdétermination sont des situations particulièrement complexes du fait qu'il s'agit de cas où plus d'une cause est suffisante à elle seule à produire une conséquence, aucune n'est donc nécessaire.
Deux raisons principales expliquent l'importance de pouvoir traiter de tels cas. La première est que de nombreuses situations dont il est important de connaître les causes sont presque inévitablement des cas de surdétermination. Mentionnons par exemple le réchauffement climatique, la pollution d'un site protégé, la délocalisation d'un site industriel, l'inégalité de représentation en politique entre les femmes et les hommes, l'existence de déserts médicaux, une perte économique pour une filière agricole ou le suicide d'une personne. La deuxième raison est qu'une partie importante des débats encore ouverts sur la notion de causalité ont à voir de près ou de loin avec la surdétermination, que ce soit en philosophie, psychologie, droit, mathématiques ou informatique. Lorsqu'une nouvelle définition de causalité est proposée, elle est la plupart du temps validée en étant confrontée à des exemples complexes. Ces exemples sont tous des exemples de surdétermination.

Les approches qui placent la nécessité de la cause par rapport à la conséquence comme l'élément central ont par nature plus de difficultés à traiter ces cas. Elles sont obligées de faire appel à des processus complexes pour y remédier. Quant aux approches par inférence, le fait de placer la nécessité au second plan leur permet par nature de gérer plus facilement ces cas. Nous discutons plus en détail de ces cas et de la façon dont ils sont gérés par les différentes approches existantes dans le chapitre 3.

Causalité négative Les cas de causalité négative sont des cas où il y a une négation dans la relation causale. Ces cas sont à l'origine de nombreux débats dans le domaine de la causalité car ils suscitent parfois des intuitions différentes aux cas sans négation.
Très simplement, une relation causale peut être vue comme une relation binaire qui lie une cause à une conséquence. Lorsque nous parlons de négation dans la relation causale, deux cas sont envisageables. Le premier est celui où la négation est dans la conséquence. Le traiter revient à pouvoir déterminer les causes pour lesquelles une conséquence ne s'est pas produite. Ces causes ont une importance en éthique computationnelle, nous voulons être en mesure de pouvoir raisonner sur le fait qu'un agent ait agi pour empêcher le mal ou le

bien de se produire. Le deuxième cas est celui où la négation est dans la cause. Le traiter revient à pouvoir déterminer les conséquences du fait qu'un évènement n'ait pas eu lieu. De telles relations causales sont aussi pertinentes en éthique computationnelle, par exemple, lorsque nous voulons être en mesure de prendre en compte le fait qu'un agent aurait pu agir pour empêcher le mal ou causer le bien, mais a omis de le faire.
Une partie importante des débats encore ouverts sur la notion de causalité gravite autour de ces deux cas de causalité négative, notamment lorsqu'il est en même temps question de surdétermination. Dans la mesure où l'intuition pour résoudre les cas de surdétermination sans négation semble différente de l'intuition pour résoudre ceux avec, nous avons décidé de séparer cette problématique de celle de la surdétermination. Comme nous le verrons dans la discussion plus détaillée du chapitre 3, la nature causale des relations contenant une négation peut être questionnée.

Transitivité Les cas impliquant une transitivité de la causalité sont des situations où il est nécessaire d'évaluer la portée que peut avoir une action, ce qui demande de pouvoir la relier à ses conséquences indirectes.
D'un côté, la nécessité d'avoir une relation causale transitive semble indispensable. En effet, cela permet de construire des chemins causaux qui paraissent nécessaires au moins pour deux aspects. Premièrement, les termes utilisés pour décrire les cas de surdétermination renvoient à l'existence de chemins causaux. Comprendre ces cas impose de passer par un raisonnement sur ces chemins. Deuxièmement, les agents éthiques explicites qui nous intéressent évoluent dans un monde dans lequel d'autres agents sont présents et peuvent agir. Il est donc nécessaire pour toute question d'imputabilité d'avoir un moyen de relier chaque action à ses conséquences directes comme indirectes. D'un autre côté, il existe des exemples dans le domaine de la causalité qui montrent qu'avoir une relation transitive donne des résultats qui semblent contre-intuitifs. Nous discutons plus en détail de ces exemples dans le chapitre 3.

Distinction avec la responsabilité Cette cinquième et dernière problématique est commune à la causalité et à l'éthique computationnelle. Il s'agit de l'importance d'établir une distinction claire entre causalité et responsabilité, où la première notion est purement factuelle alors que la deuxième est subjective. L'observation de WRIGHT [1985, 1988, 2011] est que ces notions sont trop souvent confondues car la frontière qu'elles partagent est mince. Très souvent dans le langage commun, lorsque nous entendons « la cause » nous pensons en réalité à « la cause responsable ». Cette confusion est quelque part compréhensible, dans la vie quotidienne nous sommes plus souvent confrontés à raisonner en termes de responsabilité car nous vivons dans des sociétés où les normes sont omniprésentes, qu'elles soient formelles comme les normes institutionnelles ou étatiques, ou informelles comme les normes sociales. Établir une distinction claire entre causalité et responsabilité est indispensable aussi bien pour le raisonnement causal que le raisonnement éthique.

En droit, WRIGHT [1985] défend l'idée que la distinction entre causalité et responsabilité est fondamentale. Bien séparer ces concepts est la clé de voûte qui lui permet de proposer une approche qui traite de façon satisfaisante les cas de surdétermination.
WRIGHT [1985] décompose le processus pour déterminer si un individu est responsable juridiquement d'un préjudice en trois étapes. De ces trois étapes, il montre que la deuxième, i.e. l'enquête causale, doit impérativement être factuelle et donc indépendante de tout as-

pect normatif. Elle permet de déterminer si une action est à l'origine du préjudice. Les deux autres étapes sont soumises à des considérations normatives qui déterminent les causes qui donneront lieu à une responsabilité juridique. Celles-ci font intervenir d'autres considérations subjectives, inévitables dans le cadre plus large de la responsabilité. Ces trois étapes seront détaillées et illustrées par un exemple dans le chapitre 3.

Le besoin de séparer clairement causalité et responsabilité existe également lorsque le raisonnement causal est utilisé pour l'éthique computationnelle. Comme en droit, en éthique la causalité n'est qu'une étape pour arriver au résultat final. Si d'un côté la causalité est une étape pour déterminer s'il existe une responsabilité juridique, de l'autre c'est une étape pour déterminer le statut déontique d'une action. Selon la théorie morale, la causalité joue un rôle plus ou moins important. Comme pour les aspects normatifs en responsabilité juridique, le choix de la théorie morale fait en sorte que certaines causes sont pertinentes ou non dans l'évaluation éthique. Chaque théorie morale défend donc des aspects normatifs qui lui sont propres, ces choix découlent des fondements de la théorie. Mélanger causalité et responsabilité revient à intégrer dans l'enquête causale des aspects normatifs propre à une vision particulière. Si cette enquête causale est utilisée dans la formalisation d'une théorie morale ne partageant pas cette vision, alors la formalisation obtenue ne correspondra pas à la théorie souhaitée. Une telle formalisation ne serait pas inutilisable, mais elle ne correspondrait pas réellement à la théorie morale dont elle est censée être la formalisation. Si nous voulons avoir une approche de causalité positive commune adaptée à l'éthique computationnelle, celle-ci doit impérativement respecter la séparation entre causalité et responsabilité.

Pour résumer, l'approche causale pouvant servir de base pour tout raisonnement causal en éthique computationnelle que nous cherchons à concevoir doit satisfaire un certain nombre de conditions. Tout d'abord, elle doit reposer sur un formalisme qui permette de rendre explicites toutes les subtilités des problèmes, aussi bien pour le raisonnement causal que le raisonnement éthique. Puis, la définition de causalité effective choisie pour cette approche doit être dépourvue de toute confusion avec la notion de responsabilité. Sans cela, elle ne pourrait pas être commune à la formalisation de toutes les théories morales. Pour cela, elle se doit d'être purement factuelle et donc laisser de côté toute considération normative. Finalement, l'approche proposée doit pouvoir traiter tous les cas causaux, dont les plus complexes. Cela inclut les cas de surdétermination, les cas de causalité négative et les cas débattus de transitivité.

Nos réponses à ces problématiques

Nous terminons cette introduction en présentant dans les grandes lignes les contributions appartenant au deuxième groupe. Pour rappel, il s'agit de contributions au domaine de l'éthique computationnelle mais pouvant aller au delà de ce domaine. Ces contributions visent à répondre aux problématiques mentionnées jusqu'ici.

Modélisation et représentation de l'action et du changement La première problématique à laquelle nous répondons est le choix du formalisme pour représenter le monde et son évolution. Nous avons mentionné qu'il est indispensable que l'approche commune que nous voulons concevoir utilise un formalisme qui permette de rendre explicites toutes ces subtilités, aussi bien pour le raisonnement causal que le raisonnement éthique. Pour commencer,

nous avons décidé de modéliser le monde et son évolution comme un système de transition d'états. Autrement dit, nous considérons une modélisation discrète où nous avons d'un côté les descriptions des états possibles du monde, et de l'autre les transitions qui nous permettent de passer d'un état à un autre. Cette façon de voir le monde et son évolution est la sémantique derrière notre formalisme. Dans le chapitre 2, nous montrons qu'un grand nombre des formalismes proposés par le sous-domaine de l'intelligence artificielle qui s'est concentré sur la représentation de l'action et du changement partagent cette même sémantique. Puis, nous avons décidé d'utiliser un langage de description d'action car cette représentation nous a semblé être une alternative prometteuse pour répondre à la problématique de la représentation. Dans le chapitre 2, nous expliquons que ces langages peuvent être vus comme une syntaxe permettant de décrire un système de transition d'états.

Dans le chapitre 6, nous présentons le langage de description d'action que nous utilisons pour notre approche commune. Celui-ci a été conçu pour répondre à la nécessité de rendre explicites les nuances, aussi bien de la causalité que de l'éthique. Cette représentation de l'action et du changement peut par exemple venir remplacer $\mathbb{A}$ dans le cadre modulaire $\mathbb{ACE}$ illustré sur la figure 1. Comme mentionné précédemment, ce langage intègre certaines notions causales, mais est dénué d'un raisonnement permettant de faire de la causalité effective. C'est ce à quoi nous nous attelons ensuite.

Le choix de modéliser le monde et son évolution comme un système de transition d'états a tout de suite de nombreux avantages par rapport aux formalismes classiques en causalité qui ne font pas de distinction entre évènements et états du monde, comme les équations structurelles. Le plus important, de notre point de vue, en plus de permettre de saisir les subtilités de chaque problème, est qu'il permet d'étudier séparément différentes problématiques propres à la causalité. Comme mentionné précédemment, une relation causale peut être vue très simplement comme une relation binaire qui lie une cause à une conséquence. Le tableau 1 illustre les quatre relations élémentaires qu'il est possible de trouver dans un système de transition d'états. Contrairement aux formalismes classiques utilisés en causalité, dans les systèmes de transition d'états il existe une vrai différence sémantique entre ces relations. Il est alors possible de proposer une définition pour « causer par l'action », par exemple, sans que celle-ci ne s'applique également aux cas de « causer par l'omission ». Nous verrons dans cette thèse à quel point cela est indispensable.

		Conséquence	
		+	-
Cause	+	*Causer par l'action*	*Empêcher par l'action*
	-	*Causer par l'omission*	*Empêcher par l'omission*

TABLEAU 1 – Relations « causales » élémentaires qu'il peut y avoir dans un système de transition.

La présentation de l'ensemble de notre approche causale commune est distribuée entre le chapitre 6 et le chapitre 7. Nous la décomposons en nous appuyant sur les quatre relations qui apparaissent dans le tableau 1. L'ensemble de notre approche peut être vue comme un escalier à quatre marches où chaque marche prend appui sur les précédentes.
Le cas où aussi bien la cause que la conséquence sont positives correspond à la première marche de notre approche. C'est ce qui apparaît dans le tableau comme « causer par l'action », et que nous traitons dans le chapitre 6. Les deux cas où la cause est négative correspondent à la deuxième marche de notre approche. C'est ce qui apparaît dans le tableau

comme « causer par l'omission » et « empêcher par l'omission », et que nous traitons dans le chapitre 7. Le cas où la cause est positive et la conséquence négative correspond à la troisième marche de notre approche. C'est ce qui apparaît dans le tableau comme « empêcher par l'action », et que nous traitons également dans le chapitre 7. La question de la transitivité correspond à la quatrième et dernière marche de notre approche. Nous la traitons également dans le chapitre 7. Tous ces éléments ensemble forment notre approche commune. Cette proposition peut par exemple venir remplacer $\mathbb{C}$ dans le cadre modulaire $\mathbb{ACE}$.

Modélisation de la surdétermination Avant de nous lancer dans la proposition de notre approche commune, nous proposons un travail de clarification de ce qu'est la surdétermination. Comme nous l'avons mentionné précédemment, les cas de surdétermination sont ceux qui posent encore le plus de problèmes au domaine de la causalité. La clarification que nous proposons consiste en deux étapes : une définition formelle de ce qu'est la surdétermination et une typologie formelle des différents cas de surdétermination. Pour ce faire, nous utilisons un système de transition d'états très général. Ce choix est fait pour deux raisons : tout d'abord son côté très expressif permet de saisir toutes les subtilités des différents cas de surdétermination ; ensuite, sa généralité permet une utilisation des résultats présentés ici par un plus grand nombre d'approches existantes. La clarification que nous apportons n'est pas uniquement utile à notre approche commune, elle peut bénéficier au domaine de la causalité dans son ensemble. Nous faisons ce travail pour la relation « causer par l'action » dans le chapitre 5 et pour la relation « empêcher par l'action » dans le chapitre 7.

Grâce à la typologie que nous proposons, nous montrons qu'il est possible d'améliorer considérablement la comparaison entre différentes approches. Ayant des définitions claires des différents types de cas de surdétermination, il est possible d'établir des propriétés qui caractérisent la façon dont une définition de causalité va considérer un type de surdétermination. Si pour un premier exemple une définition de causalité considère que A est une cause de C et pas B, alors, pour un autre exemple classé comme étant du même type que le premier par notre typologie, cette définition devrait donner la même réponse. Plutôt que de confronter les différentes approches à des exemples pas nécessairement représentatifs de l'ensemble des cas possibles, nous pouvons à présent prouver quelles seront les causes trouvées par une définition, et cela pour tous les exemples d'un même type.
La comparaison des typologies pour la relation « causer par l'action » et pour la relation « empêcher par l'action » montre qu'il existe une différence entre ces deux relations. Des types de surdétermination possibles en causalité positive, ne le sont pas dans le deuxième cas. Il s'agit d'un premier résultat montrant l'importance de pouvoir étudier séparément les différentes relations du tableau 1.

Modélisation, représentation et automatisation de la causalité positive La première marche de notre approche commune est le cœur de celle-ci, elle représente les fondations essentielles au développement de toute autre marche. Cette première marche consiste en une définition pour le cas où aussi bien la cause que la conséquence sont positives. C'est ce qui apparaît dans le tableau comme « causer par l'action ». Nous la présentons dans le chapitre 6.

La définition que nous adoptons est le test NESS (« Necessary Element for the Sufficiency of a Set ») proposé par le juriste WRIGHT [2011]. Notre première contribution peut être décrite comme une représentation du test NESS dans le langage de description d'action que nous utilisons.

Notre approche commune repose donc majoritairement sur des travaux dans le droit, et plus spécifiquement dans le droit pénal. De toutes les disciplines où la causalité a fait l'objet d'études, c'est celle où les travaux ont le plus de liens avec les besoins de l'éthique computationnelle. En effet, il y est question d'actions réalisées par des individus dans des situations bien spécifiques et qui rendent potentiellement l'individu qui les a réalisées responsable d'un préjudice. Nous avons choisi spécifiquement le test NESS pour ses capacités à gérer les cas de surdétermination de façon factuelle, sans passer par des raisonnements hypothétiques. Par conséquent, elle assure une séparation entre causalité et responsabilité qui font de cette définition celle qui semble convenir le mieux comme base pour une approche de causalité positive commune adaptée à l'éthique computationnelle.

S'agissant du cœur de notre approche, nous avons voulu aller jusqu'à l'étape d'automatisation pour cette première marche. Nous proposons une implémentation complète et correcte en Answer Set Programming de la définition de causalité positive adoptée. Nous la présentons dans le chapitre 6. Le programme logique obtenu permet de raisonner sur des situations causales complexes. Un tel programme permet à l'éthique computationnelle de pouvoir traiter des cas qui ne l'étaient pas auparavant. Mais l'utilité de cette implémentation ne se limite pas à l'éthique computationnelle. Celle-ci permet l'exploration de relations causales complexes dans un cadre de prise de décision. Le chapitre 8 est un exemple d'application de cette contribution pour une utilisation autre que l'éthique computationnelle. Plus exactement, il présente les bénéfices de son utilisation pour l'explicabilité dans le domaine de l'argumentation abstraite.

Modélisation et représentation de la négation dans la relation causale et de la transitivité La dernière partie de notre contribution concerne les trois autres marches, que nous traitons dans le chapitre 7. Pour rappel, celles-ci concernent la question de la transitivité et les trois relations du tableau 1 restantes, à savoir « causer par l'omission », « empêcher par l'omission » et « empêcher par l'action ». Nous montrons que l'essentiel de ces marches est au-delà du cadre purement causal, il est de l'autre côté de la frontière qui sépare causalité effective et responsabilité. Chaque théorie morale peut avoir une version différente de comment ces relations doivent être définies. De ce fait, il n'est pas possible de les inclure dans notre approche commune. Toutefois, nous montrons que la première marche construite est une base factuelle propice à formaliser différentes versions existantes de comment ces relations doivent être définies.

La deuxième marche traite les deux relations où la cause est négative, i.e. « causer par l'omission » et « empêcher par l'omission » dans le tableau 1. Nous montrons que dans un système de transition d'états l'omission ne peut pas avoir un statut causal, il s'agit d'une question de responsabilité. De fait, derrière toute volonté d'attribuer un statut causal à une omission il y a un raisonnement hypothétique. Une omission ne produit rien, le monde reste inchangé. Mais, si un agent avait la possibilité d'agir et que dans un monde hypothétique cet agir aurait changé l'issue des choses, le fait qu'il ne l'ait pas fait le rend responsable du fait que le monde soit resté tel qu'il est. Nous montrons que c'est dans le choix des mondes qu'il est possible d'envisager dans le raisonnement hypothétique, que repose la décision de quelles conséquences pourront être attribuées à l'omission. S'agissant d'un choix normatif dépendant d'un grand nombre de facteurs en dehors de la causalité, nous montrons que l'omission est essentiellement une question de responsabilité. Il n'est donc pas possible de trouver une vision qui convienne à toutes les théories morales, même si nous acceptions de

sortir du cadre purement causal en faisant un pas vers la responsabilité. Pour cette raison, proposer une définition de comment prendre en compte l'omission est au-delà du cadre de cette contribution.

La troisième marche concerne la relation « empêcher par l'action » dans le tableau 1. Nous montrons que, contrairement aux cas d'omission de la deuxième marche, certains cas qui impliquent des conséquences négatives peuvent être traités factuellement. Nous proposons une extension de notre approche de causalité positive pour les traiter. Malgré l'existence de cette solution factuelle, celle-ci ne correspond pas tout à fait à la notion d'empêcher qui vient intuitivement à l'esprit, en tout cas dans un type de cas en particulier. Nous montrons que la notion d'empêcher peut être décomposée en deux niveaux de raisonnement, un factuel et un normatif. Dans les cas qui incombent au premier niveau, l'approche factuelle que nous proposons est satisfaisante. Les cas qui posent problème font intervenir le deuxième. Il s'agit de problèmes qui demandent nécessairement d'établir des relations traitées par le deuxième marche, et donc qui sont en dehors du cadre purement causal.

La quatrième et dernière marche concerne la transitivité de la relation causale. Nous montrons que les débats atour de cette notion sont également liés à des questions de responsabilité. Comme pour les autres marches, nous montrons que notre approche factuelle peut servir de base pour traiter différentes visions possibles. Pour cela nous présentons les éléments à ajouter à notre approche afin que ce soit le cas.

Première partie

État de l'art

Chapitre 1

État de l'art : théorie morale

« *Moral theory is the study of substantive moral conceptions, that is, the study of how the basic notions of the right, the good, and moral worth may be arranged to form different moral structures.* »

RAWLS [1974]

Sommaire

Ce chapitre se veut une introduction à la *théorie morale*, une composante de l'éthique qui s'intéresse à ce qu'est le juste, le bien et le vertueux, où le « juste » doit être compris comme conforme à la morale (« right » en anglais). Une étude de cette composante de l'éthique est indispensable étant donné que cette thèse s'inscrit dans le domaine de l'éthique computationnelle qui cherche à modéliser notre capacité en tant qu'êtres rationnels à évaluer moralement une action. Ce chapitre est construit principalement à partir des éléments présentés dans *Moral Theory : An Introduction* [TIMMONS, 2012], un ouvrage qui peut être vu comme une cartographie approfondie de la théorie morale. De ce fait, à moins qu'il n'en soit spécifié autrement, les connaissances qui sont présentées dans ce chapitre sont le résultat d'une synthèse de cet ouvrage consistant à en extraire les éléments essentiels pour cette thèse.

Avant de se plonger dans cette composante spécifique de l'éthique qu'est la théorie morale, faisons un bref point sur la structure de cette discipline philosophique qu'est l'éthique. Également appelée philosophie morale, l'*éthique* est la discipline philosophique s'intéressant à la morale. Deux composantes principales structurent l'éthique.
D'un côté, la *méthaéthique* regroupe différentes branches qui s'intéressent aux aspects non moraux autour de l'étude de la morale. Par exemple, des questions sémantiques comme ce que veut dire « juste » ou « bien », des questions métaphysiques comme l'aspect universel ou non des jugements moraux, et des questions épistémologiques comme la justification des connaissances sur l'aspect juste ou non d'une action.
D'un autre côté, l'*éthique normative* regroupe deux branches qui cherchent à répondre à des questions morales. La première branche est la théorie morale qui, comme mentionné précédemment, s'intéresse à ce qu'est le juste et le bien. C'est la branche à laquelle nous nous intéressons ici. La deuxième branche est l'*éthique appliquée*, son but est d'étudier une pratique ou action spécifique. L'expérimentation animale, la peine de mort et l'inaction face au réchauffement climatique sont des exemples de sujets que traite cette branche. Pour reprendre l'analogie faite par TIMMONS [2012], l'éthique appliquée est à la théorie morale ce que l'ingénierie est à la science.

Ce chapitre est divisé en quatre sections. La section 1.1 introduit les concepts de base nécessaires à la compréhension de la théorie morale. Puis, une fois les fondations posées, les sections suivantes présenteront quelques unes des théories morales occidentales les plus importantes. Celles-ci seront regroupées en trois familles, chacune correspondant à une section. La section 1.2 regroupe les théories axées sur le devoir. La section 1.3 regroupe les théories axées sur la valeur. La section 1.4 regroupe les théories axées sur la vertu. La section 1.5 introduit brièvement l'éthique computationnelle, fait le lien avec la branche de la théorie morale et établit le périmètre éthique de cette thèse. Pour illustrer les concepts introduits, nous faisons référence dans ce chapitre à l'opéra Tosca de Puccini. Nous faisons principalement référence à trois personnages : Floria Tosca, célèbre cantatrice ; Mario Cavaradossi, peintre, amant de Tosca et sympathisant républicain ; le Baron Scarpia, chef de la police de Rome.

1.1 Concepts de base en théorie morale

Cette première section est une présentation des concepts fondamentaux en théorie morale. Les sections 1.1.1 et 1.1.2 introduisent comment le juste, le bien et la vertu structurent cette composante de l'éthique. La section 1.1.3 décrit la structure adoptée pour la présentation des différentes théories morales dans les sections suivantes.

1.1.1 Trois notions de base en théorie morale

Comme l'indique la citation de RAWLS [1974] en début de chapitre, le juste, le bien et la vertu sont trois notions de base en théorie morale. La notion de *juste* peut être rapprochée de la notion de conduite juste ou de devoir. En effet, évaluer moralement une action consiste à déterminer son *statut déontique*, i.e. si l'action est juste ou non. Plus exactement, il s'agit de dire si une action est requise, optionnelle ou interdite, où une action requise ou optionnelle est une action juste alors qu'une action interdite ne l'est pas. Dans l'opéra Tosca, le Baron Scarpia fait exécuter Mario Cavaradossi car il est un sympathisant républicain. Évaluer moralement l'action de Scarpia d'exécuter Cavaradossi revient à déterminer si son action était requise, optionnelle ou interdite.

La notion de *bien* peut être rapprochée de la notion de valeur. En effet, considérer que quelque chose a une valeur positive intrinsèque équivaut à dire que le bien est propre à cette chose, elle est bonne en elle-même ou en tant que telle, elle est *intrinsèquement bonne*. Selon le même raisonnement, quelque chose a une valeur négative intrinsèque si elle est *intrinsèquement mauvaise*, le mal est propre à cette chose. Si cette chose n'est ni intrinsèquement mauvaise ni intrinsèquement bonne, elle est considérée comme *intrinsèquement neutre*. Par exemple, certains philosophes diraient que la vie de Cavaradossi ou l'art qu'il produit ont une valeur positive intrinsèque, mais que sa sympathie pour les républicains est intrinsèquement neutre. Finalement, une chose est *extrinsèquement mauvaise ou bonne* si elle est reliée à autre chose qui respectivement a une valeur intrinsèque négative ou positive. Cette chose sera considérée comme ayant une valeur extrinsèque négative ou positive respectivement. Par exemple, même si Scarpia ne remettait pas en question que la sympathie de Caravadossi pour les républicains est intrinsèquement neutre, il pourrait considérer cette sympathie comme extrinsèquement mauvaise car remettant en question la stabilité de la société dans laquelle ils vivent, stabilité qu'il considère intrinsèquement bonne.

La notion de *vertu* peut aussi être rapprochée de la notion de valeur. Toutefois, il s'agit ici d'une valeur un peu particulière puisqu'il s'agit de la valeur attribuée à un agent autonome au sens étymologique du terme, i.e. qui se régit par ses propres lois. Il est question d'évaluer la valeur de la personne et son caractère. Scarpia est souvent décrit comme l'archétype du personnage incarnant le mal, il pourrait donc être considéré comme un personnage non vertueux. À l'opposé, Tosca qui fait tout pour sauver son amant au point de perdre la vie pourrait représenter, dans le contexte social un peu désuet de cet opéra, un personnage vertueux.

À partir des trois notions venant d'être évoquées s'articulent les théories morales occidentales les plus importantes. Toutes ces théories partagent deux objectifs, un pratique, l'autre théorique. L'*objectif pratique* consiste à procurer une procédure fiable de décision permettant aux agents rationnels et bien informés de produire des verdicts moraux corrects. En définitive, l'objectif peut se résumer à permettre d'évaluer moralement les actions et donc à déterminer leur statut déontique. L'*objectif théorique* consiste à identifier quels aspects d'une action, d'un état du monde ou d'une personne font qu'ils peuvent être qualifiés respectivement de juste ou non juste, de bons ou mauvais, de vertueux ou non vertueux.

1.1.2 Trois piliers structurant la théorie morale

Dans la citation en début de chapitre, RAWLS [1974] énonce qu'un des éléments principaux qui différencie les théories morales entre elles est la façon dont sont structurées les

notions de juste, de bien et de vertu. En effet, chaque théorie morale propose sa propre organisation de ces notions en vue de satisfaire leur objectif pratique et théorique. Plus spécifiquement, ce qui varie selon chaque théorie est la hiérarchie entre les notions, la relation qu'elles ont entre elles, et bien évidemment la définition de ce qu'est le juste, le bien ou le vertueux.

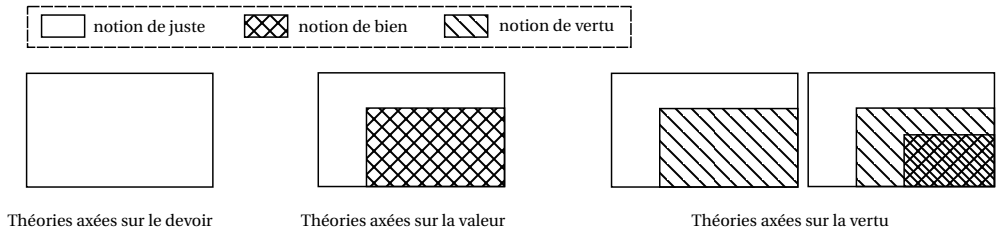


FIGURE 1.1 – Illustration de la façon dont sont structurées les notions de juste, bien et vertu selon les trois catégories de théories morales.

Malgré la spécificité de chaque organisation existante, il est possible de regrouper la plupart des théories morales dans trois catégories. La figure 1.1 est une illustration de ces trois familles de théories morales. Cette décomposition a de nombreuses similarités avec celle couramment utilisée qui parle de théories déontologiques, conséquentialistes et de la vertu. Toutefois, la décomposition de TIMMONS [2012] que nous présentons ci-dessous et que nous adoptons dans cette thèse a le mérite de subsumer la décomposition classique tout en clarifiant certaines zones d'ombre quant à la classification de certaines théories.
La première catégorie est composée des théories qui placent le juste au premier plan. Pour déterminer le statut déontique d'une action, les théories appartenant à cette catégorie définissent directement leur *théorie du juste*. Dit autrement, elles définissent ce qui est requis, optionnel ou interdit, sans que cela ne fasse intervenir les notions de bien ou de vertu. C'est le cas de la théorie du commandement divin, le relativisme moral ou la théorie morale de Kant. Ces théories sont qualifiées de *théories axées sur le devoir*.
La deuxième catégorie est composée des théories qui placent le bien au premier plan. Pour déterminer le statut déontique d'une action, et donc leur théorie du juste, les théories appartenant à cette catégorie définissent en premier leur *théorie du juste*. La première étape est donc de définir ce qui est considéré comme ayant de la valeur intrinsèquement. Puis, le fait que l'action contribue à amener ou non de la valeur est utilisé pour définir le statut déontique de l'action. C'est le cas de la théorie du droit naturel et de toutes les théories conséquentialistes comme l'utilitarisme de l'acte, l'utilitarisme espéré, l'utilitarisme de la règle ou l'égoïsme éthique. Ces théories sont qualifiées de *théories axées sur la valeur*.
La troisième catégorie est composée des théories qui utilisent ce qui est défini comme étant la vertu pour déterminer leur théorie du juste. Il serait erroné de dire que ces théories placent la vertu au premier plan. En effet, selon la théorie morale, il se peut que leur *théorie de la vertu* ne soit pas le premier élément défini, mais que celui-ci se base sur une théorie du bien. Toutefois, ces théories ont toutes en commun, qu'en plus de proposer une procédure pour déterminer le statut déontique de l'action, elles essaient d'indiquer quelle devrait être la manière d'être d'un agent avec leur théorie de la vertu. C'est le cas de l'éthique aristotélicienne et l'éthique de la sollicitude (*ethics of care* en anglais). Ces théories sont qualifiées de *théories axées sur la vertu*.

1.1.3 Méthodologie de présentation des théories morales

Chacune des sections 1.2, 1.3 et 1.4 est consacrée à la présentation de théories morales d'une des familles introduites dans la section 1.1.2. Le choix des théories morales présentées en détail a été fait de façon à donner l'aperçu le plus large possible des structures existantes.

Dans la mesure où cette thèse s'inscrit dans un domaine qui cherche à formaliser notre capacité en tant qu'êtres rationnels à évaluer moralement une action, la présentation des théories morales est davantage orientée vers l'objectif pratique que vers l'objectif théorique. Ainsi, nous présentons uniquement ce qui est nécessaire pour déterminer le statut déontique d'une action.
Un *principe moral* est une règle générale posant les conditions selon lesquelles une action est juste, un état du monde est bon ou une personne est vertueuse. Présenter une théorie morale dans le cadre de cette thèse se résumera à énoncer le principe moral posant les conditions selon lesquelles une action est juste, en ayant auparavant introduit les éléments nécessaires à sa compréhension. Selon le même raisonnement, nous n'allons pas nous attarder sur les critiques qui sont faites sur chaque théorie. Quelques unes de ces critiques seront abordées dans le chapitre 4 où nous proposerons un cadre commun pour la formalisation du raisonnement éthique dans un but de clarification de concepts pour l'éthique computationnelle.

Pour faciliter la compréhension, la même structure est adoptée pour la présentation de l'ensemble des théories morales. Toutes les théories comptent avec au moins deux parties : une première partie introduit l'idée générale, une autre introduit le principe moral permettant de déterminer le statut déontique des actions. Lorsqu'une théorie s'appuie sur la notion de valeur et/ou de vertu pour définir le juste, une partie sera consacrée à chaque notion supplémentaire. Les théories axées sur le devoir sont présentées en deux parties puisqu'elles définissent le statut déontique sans que cela ne fasse intervenir les notions de bien ou de vertu. Les théories axées sur la valeur sont présentées en trois parties puisque le statut déontique de l'action est définit par rapport à ce qui est considéré comme ayant de la valeur. Les théories axées sur la vertu sont quant à elles présentées en trois ou quatre parties en fonction de si la vertu est définie indépendamment du bien ou non.

Avant de finir cette présentation des concepts de base en théorie morale, précisons la différence qu'il existe entre les règles morales et les principes moraux. Une *règle morale* est une déclaration morale au sujet d'un type d'action, d'un état du monde ou d'un groupe d'agents. Une règle morale peut donc être considérée comme une déclaration morale moins générale qu'un principe moral qui est destiné à s'appliquer à toutes les actions, tous les états du monde ou tous les agents. En guise d'analogie nous pourrions dire que la règle morale est au principe moral ce qu'en logique le fait est à la règle. Un principe moral pour Scarpia pourrait être qu'une action est juste si celle-ci permet d'augmenter la stabilité dans la société. Étant donné son comportement, nous pouvons imaginer que pour lui il existe une règle morale déclarant que les idées avancées par les républicains nuisent à l'ordre public et une autre indiquant que le maintient de l'ordre publique est essentiel pour la stabilité dans la société. Ces deux faits permettent à Scarpia de conclure, d'après le principe moral, qu'il est juste d'exécuter Cavaradossi.

1.2 Théories axées sur le devoir

Cette section traite des théories morales où le concept de devoir est au moins aussi fondamental que les concepts de valeur. Pour déterminer le statut déontique d'une action, les théories appartenant à cette catégorie définissent directement ce qui est requis, optionnel ou interdit, sans que cela ne fasse intervenir les notions de bien ou de vertu. La section 1.2.1 introduit deux théories morales se basant sur des codes de conduite. La section 1.2.2 présente la théorie morale de Kant.

1.2.1 Des théories reposant sur des codes de conduite

Un *code de conduite* est un ensemble de règles morales qui dictent la façon juste d'agir. Dans la mesure où nous parlons ici de règles morales exclusivement liées aux actions et leur statut déontique, il s'agit d'un ensemble de règles morales de nature bien spécifique. Pour être plus exacts, nous parlerons plutôt d'un ensemble de *normes morales*, i.e. un ensemble de déclarations morales au sujet d'actions.

Au sein d'un code de conduite il est possible de trouver des normes morales qui dérivent d'autres. Il est donc possible d'en qualifier certaines de plus basiques que d'autres. Par exemple, un code de conduite peut contenir à la fois une norme morale indiquant qu'il est interdit d'ôter la vie à un être sentient et une autre indiquant qu'il est interdit d'empoisonner un être sentient. Nous pouvons considérer qu'ôter la vie est un type d'action dont l'empoisonnement mortel en est une spécification. C'est pourquoi la première norme est dite plus basique que la deuxième.

La section 1.2.1.1 introduit la théorie du commandement divin où le code de conduite est dicté par un dieu. La section 1.2.1.2 présente la théorie du relativisme moral où le code de conduite est dicté par la culture de l'individu.

1.2.1.1 Théorie du commandement divin

De l'idée qu'un dieu est source et créateur de tout, découle l'idée selon laquelle ce qui est juste, bien et vertueux ne dépend pas de ce que les êtres de la création en pensent, mais uniquement d'une « activité volitionnelle » de ce dieu, comme ce qu'il désire, ce qu'il croit ou ce qu'il commande. La théorie du commandement divin, et toute autre théorie adoptant cette idée, sont dites comme appartenant au volontarisme théologique.

Ce qui fait la spécificité de la théorie du commandement divin est que le juste, le bien et le vertueux dépendent exclusivement de ce que ce dieu commande. Une action est juste si ce dieu commande de la faire. Si dieu commande de tout faire pour sauver les êtres sentients qui nous sont chers, alors les actions de Tosca étaient requises, même faire croire à Scarpia qu'elle céderait à ses avances pour obtenir un sauf-conduit et le poignarder ensuite.
Rien n'est intrinsèquement bon ou mauvais, un état du monde ne peut avoir de la valeur qu'extrinsèquement car cela dépend de sa relation avec le commandement divin. L'état du monde où Scarpia est mort suite à sa blessure a une valeur extrinsèque positive dans ce cas de figure.
Un être est dit vertueux si ses actions suivent à la lettre les commandements divins. Lors d'une rencontre entre Tosca et Cavaradossi, la cantatrice laisse éclater sa jalousie en voyant que son amant a utilisé la marquise Attavanti comme modèle pour la Marie-Madeleine de son tableau. Si ce dieu commande qu'il ne faut pas douter des êtres sentients qui nous sont

chers, Tosca ne peut pas être considérée comme vertueuse. Au contraire, dans l'opéra Lohengrin de Wagner, il existe l'idée que Lohengrin et tous les chevaliers du Graal ont la plus pure vertu car ils vouent leur vie à accomplir le mandat divin de protéger le Graal.

Théorie du juste Le principe moral selon lequel une action est juste pour la théorie du commandement divin peut s'énoncer ainsi :

> TCD [théorie du commandement divin] : une action est requise si et seulement si un dieu source et créateur de tout, commande qu'elle soit réalisée. Une action est interdite si et seulement si ce dieu commande, de ne pas la réaliser. Une action qui n'est ni requise, ni interdite est optionnelle.

Il est possible de voir la TCD comme un code de conduite dont les normes morales qui le compose sont déterminées par le commandement divin.

Dans le film Breaking the Waves sorti en 1996 et réalisé par Lars von Trier, la protagoniste est l'archétype d'un être adhérant à la TCD. Bess McNeill, femme vivant dans un village écossais où la pratique de la religion peut être qualifiée de rigoriste, se marie avec Jan Nyman qui peu après son mariage se retrouve paralysé suite à un accident de travail. Convaincue de parler directement avec son dieu, elle va vivre des expériences dangereuses et condamnées par sa communauté étant donné que dieu lui ordonne d'agir ainsi pour sauver son mari. Bess désobéit aux normes morales suivies par sa communauté, qui pourtant sont censées être imposées par dieu, car elle pense recevoir les commandements de dieu directement, sans devoir passer par les écritures, moyen majoritairement admis comme celui par lequel dieu transmet ses commandements. En plus de donner un exemple illustrant ce à quoi ressemble l'adoption de cette théorie morale, ce film fait ressortir un des problèmes auxquels cette théorie est confrontée, déterminer le moyen par lequel le code de conduite est accessible.

1.2.1.2 Théorie du relativisme moral

Des observations anthropologiques selon lesquelles des cultures différentes adhèrent à des règles morales différentes sont nées les théories morales selon lesquelles ce qui est juste, bien et vertueux dépend des codes de conduite de chaque culture. Cesdits codes étant définis comme un ensemble de normes morales communément acceptées par les membres de la culture.
Ce qui fait la spécificité de la théorie du relativisme moral qui nous intéresse ici est que le juste, le bien et le vertueux pour un individu dépendent exclusivement de la culture à laquelle cet individu appartient.

Théorie du juste Le principe moral selon lequel une action est juste pour la théorie du relativisme moral peut s'énoncer ainsi :

> TRM [théorie du relativisme moral] : une action est requise, interdite ou optionnelle pour des membres d'une culture si et seulement si un code de conduite de cette culture le stipule ainsi.

Si deux individus appartiennent à des cultures adhérant à des codes de conduite différents, il est tout à fait possible que des conflits existent étant donné que ce qui est juste pour l'un ne l'est pas nécessairement pour l'autre. Il est important de distinguer le conflit

lié à l'aspect contextuel de l'éthique qui est possible quel que soit la théorie morale adoptée et dont l'existence est communément acceptée, du conflit dont nous parlons qui lui est possible dès lors qu'un individu adhère au relativisme moral. Le premier est qualifié de *désaccord moral non fondamental*; il s'agit d'un conflit qui prend ses racines exclusivement dans les faits non moraux le concernant. Un tel conflit se présenterait entre Tosca et Scarpia si tous deux partageaient la norme morale selon laquelle il est requis de préserver la stabilité de la société alors qu'ils ne sont pas d'accord sur le fait que les républicains perturbent cette stabilité. Ce type de conflit peut être résolu de façon rationnelle en faisant appel à la science étant donné qu'il dépend exclusivement de faits non moraux.
Le deuxième est qualifié de *désaccord moral fondamental*; il s'agit d'un conflit entre les normes morales les plus basiques, même l'unanimité sur l'ensemble des faits non moraux concernant le conflit ne permettrait pas de le résoudre. Un tel conflit se présenterait entre Tosca et Scarpia si l'un adhère à l'idée qu'il est requis d'abolir toute forme de société, alors que l'autre adhère à l'idée qu'il est requis de préserver la stabilité de la société. Cet exemple illustre un premier type de désaccord moral fondamental. Il en existe un deuxième qui repose sur les préférences entre normes morales. Imaginons une situation où Tosca et Scarpia partagent à la fois la norme morale selon laquelle il est requis de préserver la stabilité de la société et la norme morale selon laquelle il faut tout faire pour sauver les êtres sentients qui nous sont chers, mais que les actions possibles dans la situation ne permettent pas de satisfaire les deux normes. Le deuxième type de conflit se présenterait s'il y avait divergence sur le choix de la norme morale à respecter en priorité.

Il est intéressant de remarquer que l'existence d'un moment dans l'histoire où toutes les cultures partageraient exactement les mêmes règles morales n'impliquerait pas que ces théories sont fausses. Il serait toujours possible d'imaginer qu'une culture dont les règles morales diffèrent puisse se développer.

1.2.2 Théorie morale de Kant

De l'idée qu'adopte Kant selon laquelle une exigence morale est une exigence rationnelle, découle le fait que par sa théorie Kant cherche à déterminer ce qui est juste, bien et vertueux pour tous les être rationnels et non pas exclusivement pour les êtres humains. Le principe moral qu'il propose pour définir le juste, le bien et la vertu est connu comme l'Impératif Catégorique.

Les raisons pour lesquelles Kant parle d'impératif catégorique sont importantes pour comprendre sa théorie morale. En vue de cet objectif, nous allons en premier expliquer ce qu'est un impératif hypothétique pour le philosophe prussien. Dès lors qu'un agent rationnel a un objectif qui nécessite la réalisation de certaines actions pour être atteint, il existe un impératif hypothétique. Cela prend la forme suivante : si un agent rationnel a comme objectif O et reconnaît que faire A est nécessaire à l'accomplissement d'un tel objectif, alors il est impératif qu'il fasse A ou qu'il renonce à O. Aller à l'encontre d'un tel impératif serait tout simplement irrationnel. Si l'objectif de Tosca était de sauver Cavaradossi et que tuer Scarpia était une condition nécessaire à ce but, ne pas le faire aurait été irrationnel. Attention toutefois à ne pas déduire que parce que Tosca agit de facon rationnelle, alors elle agit moralement, l'objectif a un rôle important. En effet, l'impératif que nous avons là n'est valable que si l'agent fait sien l'objectif, il est par conséquent conditionnel et c'est pour cela que c'est un impératif hypothétique. Scarpia ne partageant pas l'objectif de Tosca, il n'est

en aucun cas tenu d'agir de cette façon. Pour Kant, une exigence morale est nécessairement inconditionnelle, un impératif hypothétique ne peut donc pas être un principe moral.

Ce qui différencie un impératif catégorique d'un impératif hypothétique est que le premier requiert qu'une action soit faite quels que soient les objectifs propres à l'agent, pour peu qu'il existe un objectif imposable à tous pour incarner cet impératif. Un tel objectif est dit être une fin en soi. Sous cette condition, un impératif catégorique possède l'inconditionnalité nécessaire pour être un principe moral. Kant propose l'autonomie pour incarner son Impératif Catégorique, i.e. la capacité inhérente à tout agent rationnel d'agir librement sur la base de la raison et indépendamment de ses désirs propres. Il appelle « humanité » cette autonomie propre aux agents rationnels.

Plusieurs formulations de l'Impératif Catégorique ont été proposées par le philosophe, formulations qu'il revendique équivalentes. Voici celle qui fait apparaître explicitement la notion d'humanité :

> HFS [humanité comme fin en soi] : agis toujours de manière à traiter l'humanité, aussi bien dans ta personne que dans la personne des autres, comme une fin et à ne t'en servir jamais comme d'un simple moyen. [KANT, 1785]

Deux parties composent ce principe moral, une première positive et une deuxième négative. La partie positive exige de traiter l'humanité comme une fin. Cela équivaut pour Kant à agir de sorte à promouvoir le perfectionnement de soi et le bonheur des autres. Le processus par lequel il arrive à cette conclusion appartenant exclusivement à l'objectif théorique de la théorie morale, nous ne nous attarderons pas dans les détails du raisonnement.
La partie négative interdit de se servir des autres comme de simples moyens. Bien que la signification exacte de ce que cela veut dire suscite de nombreuses discussions en philosophie, il semblerait qu'il existe un courant majoritaire qui associe cela à traiter les autres de façon à ce qu'ils ne puissent pas consentir rationnellement. Le meurtre de Scarpia rentre clairement dans cette catégorie, il ne consent pas à être poignardé. Mais ce n'est pas le seul moment où une action de Tosca va à l'encontre du HFS. En mentant sur ses intentions à Scarpia pour obtenir un sauf-conduit, Tosca l'empêche de prendre une décision éclairée et consentie.

Des deux parties qui composent le HFS, Kant déduit tout un ensemble de devoirs qui lui permettent de créer un code de conduite. Ce code est structuré du plus général, où se trouve l'Impératif Catégorique, au plus spécifique, où se trouvent des normes morales qui s'appliqueraient dans des situations concrètes. En définitive, cette théorie pourrait tout à fait avoir sa place dans la section 1.2.1. Cependant, du fait que Kant propose une autre version de son Impératif Catégorique qui suggère un mécanisme d'évaluation des actions intéressant par sa construction, ce serait une erreur de s'arrêter là. Dans la mesure où notre objectif est de donner l'aperçu le plus large possible des structures existantes, soit nous devons traiter cette deuxième formulation, soit nous devons abandonner notre objectif. Les lecteurs les plus attentifs identifieront que la phrase venant d'être énoncé est un impératif hypothétique.

Théorie du juste Parmi les trois formulations de l'Impératif Catégorique existantes, celle qui suit est, selon TIMMONS [2012], celle qui répond le plus directement à l'objectif pratique de la théorie morale. C'est pour cette raison que le HFS a été présenté dans la partie introductive de la théorie et non dans la partie propre à la théorie du juste.

> LU [loi universelle] : agis uniquement d'après la maxime qui fait que tu puisses vouloir en même temps qu'elle devienne une loi universelle. [KANT, 1785]

Une *maxime* est une déclaration de la forme : je vais faire A, si certaines conditions sur l'état du monde S sont réunies, afin d'accomplir O. C'est une modélisation d'un état psychologique d'un agent indiquant ce qu'il désire faire. Adopter une maxime revient à agir selon cette maxime. Il est donc possible d'imaginer que Tosca a adopté la maxime : je vais aller jusqu'à tuer une personne, si cette personne agit de façon injuste et que sa mort est nécessaire, afin de sauver une personne poursuivie pour ses idéaux politiques.

Le principe moral LU indique que pour savoir si une action est juste moralement, un agent doit d'abord identifier la maxime selon laquelle cette action est réalisée. Puis, il doit considérer ce que serait le monde si tous les êtres rationnels adoptaient cette maxime. Si dans cette situation hypothétique aucune incohérence n'apparaît, la maxime est vouée à être une loi universelle et l'action est requise. Dans le cas contraire, l'action est interdite. Reprenons la maxime ci-dessus et essayons de l'universaliser. Y aurait-il une incohérence dans le monde si tous les individus étaient prêts à tuer une personne si celle-ci agissait de façon injuste et que sa mort était nécessaire pour sauver une personne poursuivie pour ses idéaux politiques ? Il semblerait que oui. Il est possible de considérer que Scarpia est assassiné alors qu'il essaie juste de protéger la société d'un individu qu'il considère comme un danger car menaçant la stabilité de la société. Si nous adoptons le point de vue d'un sbire de Scarpia, il peut considérer que son chef est assassiné pour les idéaux politiques par lesquels il agit et que Tosca agit de façon injuste. De ce fait, l'universalisation de cette maxime impliqueraient qu'il serait juste pour lui de tuer Tosca. Le monde dans lequel la maxime est universalisée ne peut pas être souhaitable pour Tosca ou n'importe quel agent qui tient à sa vie.

Le principe moral selon lequel une action est juste pour la théorie morale de Kant peut s'énoncer ainsi :

> TMK [théorie morale de Kant] : une action est requise si et seulement si la maxime selon laquelle la réalisation de l'action est omise n'est pas universalisable. Une action est interdite si et seulement si la maxime selon laquelle l'action est réalisée n'est pas universalisable. Une action qui n'est ni requise, ni interdite, ou à la fois requise et interdite est optionnelle.

TIMMONS [2012], dont une partie des travaux a porté sur la théorie morale de Kant, suggère de combiner les principes moraux HFS et LU. Il propose d'utiliser le principe moral LU comme procédure de décision et le principe moral HFS comme une sorte d'heuristique permettant de formuler les maximes selon lesquelles les actions sont réalisées. La maxime que nous avons formulée qui correspondrait à l'action de Tosca pourrait s'exprimer d'une infinité de façons différentes. En l'occurrence, il serait possible de formuler la maxime ainsi : je vais aller jusqu'à tuer Scarpia dans le quartier général de la police de Rome la nuit, si cette personne agit de façon injuste envers Cavaradossi et que sa mort est nécessaire, afin de sauver une personne à laquelle je tiens sentimentalement poursuivie pour ses idéaux politiques. Dans ce cas là, l'universalisation de la maxime ne semble plus mener à l'incohérence constatée précédemment ; celle-ci est devenue si spécifique que son universalisation se restreint presque à une unique situation, celle du cas évalué. Nous nous retrouvons dans la situation problématique où le statut déontique d'une même action changerait selon la façon dont la maxime justifiant cette action est formulée. Ce que TIMMONS [2012] suggère

est que le principe HFS peut être utilisé comme une heuristique pour savoir quels détails de l'action, des circonstances et de l'objectif peuvent apparaître dans la maxime. Selon cette idée, tous les éléments rajoutés dans la dernière version de la maxime n'ont pas lieu d'être. Par contre, le fait que l'individu à sauver soit poursuivi pour ses idéaux politiques semble être pertinent car cela est une attaque même à l'autonomie de cet individu et donc à son « humanité ».

1.3 Théories axées sur la valeur

Cette section traite des théories morales où le concept de valeur est antérieur au concept de devoir. Elles partagent l'idée qu'il est possible, sans faire appel à la notion de juste, d'attribuer une valeur aux constituants des états du monde et donc comparer différents constituants ou même différents états du monde entre eux. Dans toutes ces théories morales, le concept de valeur est utilisé pour déterminer le statut déontique d'une action. La section 1.3.1 introduit quelques théories morales appartenant à la grande famille des théories conséquentialistes selon lesquelles toute évaluation morale se base sur les conséquences des actions. La section 1.3.2 présente la théorie du droit naturel de Thomas d'Aquin, théorie sur laquelle s'appuie la morale catholique romaine.

1.3.1 Conséquentialisme

Sur l'idée que la valeur des conséquences associées à une action est la seule chose à prendre en compte pour déterminer si une action est juste, se base l'ensemble des théories morales conséquentialistes. À partir de là, un grand nombre de variantes peut être construit en faisant varier, aussi bien ce qui est considéré comme ayant de la valeur intrinsèque, que la façon dont ce qui a de la valeur détermine ce qu'il est juste de faire. Dit plus simplement, les différentes théories morales diffèrent autant par la théorie du bien que par la théorie du juste qu'elles proposent.

Dans cette section nous présentons quelques unes des théories morales conséquentialistes existantes. Comme pour l'ensemble du chapitre 1, le choix des théories morales présentées en détail a été fait de façon à donner l'aperçu le plus large possible des structures existantes. Toutefois, pour donner une idée du grand nombre de théories morales qui peuvent être construites sur la simple idée de base du conséquentialisme, nous survolons les éléments qui sont le plus communément modifiés et combinés. Cela permet de mieux identifier les points de divergence entre les différentes théories morales que nous présentons.

— Ce qui peut varier dans la théorie du bien :

1. *ce qui est considéré comme étant bon intrinsèquement.* La posture la plus commune est connue sous le nom de *welfarisme*. C'est l'idée selon laquelle la seule chose qui a de la valeur intrinsèquement est le bien-être des individus. Au sein de cette vision, différents points de vue s'opposent. L'utilitarisme, famille de théorie conséquentialistes la plus répandue, adopte un point de vue *hédoniste*. Cela consiste à dire que le bien-être d'un individu peut se résumer au plaisir et à la souffrance qu'il ressent. De ce fait, les expériences de plaisir sont les seules à être intrinsèquement bonnes et les expériences de souffrance sont les seules à

être intrinsèquement mauvaises. Après avoir obtenu le sauf-conduit et tué Scarpia, Tosca se rend à l'exécution de Cavaradossi et le prévient qu'il devra simuler sa mort car les fusils seront chargés à blanc. Les deux amants se croient sauvés et bientôt libres de fuir Rome avec le sauf-conduit. Selon un hédoniste, si cette situation a de la valeur c'est uniquement par le fait qu'elle leur procure du plaisir.
D'un autre côté, d'autres théories adoptent plutôt l'idée que c'est dans l'assouvissement des désirs où se trouve le bien-être des individus et dans leur frustration où se trouve leur mal-être. C'est donc dans l'assouvissement des désirs où se trouve la valeur intrinsèque. Dans tous les cas, c'est le bien-être des individus qui est au centre de cette théorie du bien. Selon ce point de vue, la situation avant l'exécution n'a de la valeur intrinsèque que par le fait que cet état du monde assouvit le désir de Tosca de sauver le peintre.
Parmi les alternatives au welfarisme se trouve le perfectionnisme. Il s'agit de l'idée selon laquelle certains états du monde rendent meilleure la vie des individus sans nécessairement contribuer à leur bien-être ou leur bonheur. De tels états contribuent au perfectionnement de soi et donc ont de la valeur pour nous les humains. Une condition nécessaire pour se parfaire est d'être en vie, la situation avant l'exécution a de la valeur pour un perfectionniste par le fait que Cavaradossi est en vie. La section 1.3.2 présente une théorie s'appuyant sur une théorie du bien perfectionniste.

2. *l'importance donnée à chaque individu.* Une première approche consiste à dire que la valeur d'une conséquence qui affecte un agent sera la même quel que soit l'agent. Cette version est qualifiée de *conséquentialisme impartial*. En réalité, cette notion peut être décomposée en deux autres. La première consiste à dire que l'assignation de valeur à une conséquence est complètement indépendante de l'agent que cela affecte. Une théorie morale conséquentialiste adoptant cette perspective est qualifiée de *conséquentialisme agent-neutre*. La deuxième consiste à dire que l'appartenance d'un agent à un groupe n'a aucun impact sur la valeur qu'a une conséquence pour lui. Une théorie morale conséquentialiste adoptant cette perspective est qualifiée de *conséquentialisme non priorisant*.
Pour obtenir une théorie conséquentialiste non impartiale il suffit d'aller à l'encontre d'au moins une de ces deux notions composant l'impartialité. L'égoïsme éthique est un exemple de théorie conséquentialiste agent-relative. En effet, elle se base sur l'idée que seules les conséquences de l'action d'un agent qui l'affectent lui-même ont de la valeur, et donc que seules ces conséquences comptent pour évaluer l'action. Cette théorie incite les agents à agir dans leur propre intérêt uniquement. Pour évaluer si poignarder Scarpia est juste, Tosca devrait considérer les conséquences qui l'affectent elle et ignorer complètement les conséquences sur Scarpia ou tout autre individu.
Une théorie est dite priorisante si par exemple elle défend l'idée qu'une blessure pour un humain a une valeur négative plus importante que la même blessure pour un animal. Dans ce même ordre d'idées, la théorie morale suivie par Scarpia serait priorisante s'il était convaincu que la souffrance causée a un de ses sbires avait plus de valeur que la même souffrance causée à un républicain.

— Ce qui peut varier dans la théorie du juste :

1. *la façon dont l'action est évaluée.* L'approche la plus simple consiste à évaluer les conséquences spécifiques de la réalisation de l'action dans un contexte donné. Ce conséquentialisme est qualifié de *conséquentialisme direct*. Pour évaluer l'action de Tosca consistant à poignarder Scarpia, il est nécessaire d'évaluer l'action telle qu'elle a été faite, dans le contexte précis où elle a été faite. C'est à dire poignarder Scarpia dans le quartier général de la police de Rome la nuit. Nous en verrons un exemple dans la section 1.3.1.1.
   Les alternatives sont qualifiées d'indirectes. Une première approche consiste à s'intéresser aux motivations qui ont poussé l'agent à agir de cette façon. Ce ne sont pas les conséquences de l'action dont la valeur nous intéresse pour évaluer le juste, mais les conséquences généralement produites lorsqu'un agent agit selon ces motivations. Ici ce ne sont pas les conséquences de l'action poignarder Scarpia qui sont évaluées, mais les conséquences généralement produites lorsqu'un agent agit pour sauver un autre agent auquel il tient.
   Une deuxième approche possible repose sur un code de conduite. L'idée est de construire un code de conduite dit idéal en évaluant les conséquences de l'adoption de chaque règle par tous les individus. Puis, une action sera évaluée comme juste ou non en fonction de ce que disent les règles dans ce code idéal. Si agir pour sauver un agent auquel un individu tient est dans le code de conduite idéal, alors l'action de Tosca est juste. Cette approche de conséquentialisme indirect sera présentée plus en détail dans la section 1.3.1.2.
2. *les individus pris en compte dans le calcul.* La posture classique est qualifiée de *conséquentialisme universaliste*. Elle consiste à considérer que tous les individus impactés par les conséquences de l'action doivent être pris en compte dans son évaluation. Pour comprendre la spécificité de cette posture, il est intéressant de la rapprocher d'une approche agent-neutre avec laquelle elle peut être confondue.
   D'un côté, la théorie du bien d'une théorie conséquentialiste non universaliste mais agent-neutre va attribuer de la valeur à toutes les conséquences d'une action impactant un individu, mais la théorie du juste n'utilisera pas toutes les conséquences pour évaluer le statut déontique de l'action. Certes, la souffrance causée aux sbires de Scarpia par sa mort a une valeur, mais elle n'est pas prise en compte dans l'évaluation de l'action. De l'autre côté, la théorie du bien d'une théorie conséquentialiste universaliste mais agent-relative ne va attribuer de la valeur qu'aux conséquences impactant certains individus, mais la théorie du juste utilisera toutes ces conséquences pour évaluer le statut déontique de l'action. Pour Tosca, la souffrance causée aux sbires de Scarpia n'a aucune valeur, donc bien que tous les individus soient pris en compte, leur souffrance n'a aucune influence sur l'évaluation de l'action. Toutefois, si un des sbires était un ami d'enfance de Tosca, sa souffrance pourrait avoir de la valeur et donc influencer l'évaluation de l'action.
3. *le type de conséquences prises en compte dans le calcul.* Une première approche repose sur l'idée que ce sont les conséquences réelles qui doivent être prises en compte, celles qui ont effectivement eu lieu. Une théorie morale conséquentialiste adoptant cette perspective est qualifiée de *conséquentialisme effectif*. Se-

lon cette idée, si le meurtre de Scarpia a été à l'origine d'une intensification des poursuites aux républicains, toutes les conséquences qui en découlent doivent être prises en compte dans l'évaluation de l'action. De plus, il se trouve que Scarpia a menti à Tosca, les fusils ne sont pas chargés à blanc, il s'agit d'une vrai exécution. Pour le conséquentialisme effectif, sauver la vie de Cavaradossi ne peut pas être considéré comme une conséquence de l'action de Tosca, puisque celui-ci a finalement été exécuté.
Parmi les alternatives se trouvent des théories morales qui considèrent que ce sont plutôt les conséquences que l'agent espérait que l'action ait qui doivent compter pour évaluer son statut déontique. L'utilitarisme espéré considère que pour agréger la valeur des conséquences espérées des actions il faut, avant de sommer le tout, multiplier la valeur de chaque conséquence par la probabilité que l'agent attribue à leur production effective. Selon cette variante, l'intensification des poursuites aux républicains n'étant pas une conséquence que Tosca avait envisagée, la valeur de l'ensemble des conséquences qui en découlent sont multipliées par une probabilité nulle et donc n'ont aucun impact dans l'évaluation de l'action. Par le même raisonnement, sauver la vie de Cavaradossi étant une conséquence espérée et dans laquelle Tosca croyait jusqu'à la fin, celle-ci serait pris en compte dans l'évaluation de l'action.

4. *la condition pour déterminer si l'action est juste*. Une approche exigeante consiste à dire qu'une action est juste si elle produit la plus grande quantité totale de valeur parmi toutes les alternatives. Une théorie morale conséquentialiste adoptant cette perspective est qualifiée de *conséquentialisme maximisant*. Poignarder Scarpia n'est considéré comme juste que si aucune autre action n'avait permis de produire la même quantité totale de valeur.
Une façon de relâcher la contrainte est de regarder la quantité totale de valeur positive ou négative seulement, au lieu de prendre la quantité totale. C'est le cas de l'utilitarisme négatif qui donne une priorité à la minimisation du mal par rapport à la maximisation du bien.
Une autre alternative cherchant à être encore moins contraignante est le conséquentialisme satisfaisant. Cette version du conséquentialisme adopte l'idée qu'une action est juste si ses conséquences sont suffisamment bonnes. Cette approche sera présentée plus en détail dans la section 1.3.1.3. Selon cette variante, poignarder Scarpia est considéré comme juste du moment où la valeur globale est supérieure à un seuil, quelques soient les actions alternatives que Tosca pouvait faire.

La section 1.3.1.1 introduit la théorie conséquentialiste la plus répandue, l'utilitarisme de l'acte. La section 1.3.1.2 présente l'utilitarisme de la règle faisant varier la façon dont l'action est évaluée par rapport à l'approche directe de l'utilitarisme de l'acte. Finalement, la section 1.3.1.3 présente le conséquentialisme satisfaisant faisant varier la condition pour déterminer si l'action est juste par rapport à l'approche maximisante de l'utilitarisme de l'acte.

1.3.1.1 Utilitarisme de l'acte

L'utilitarisme de l'acte est sans aucun doute la théorie conséquentialiste la plus connue. Jeremy Bentham et John Stuart Mill sont considérés comme les plus grands contributeurs aux bases de ce conséquentialisme. Dans la mesure où la majorité des éléments pouvant varier dans les différentes théories conséquentialistes ont été introduits, il est possible de classifier précisément l'utilitarisme de l'acte dans cette grande famille. L'utilitarisme de l'acte peut être considéré comme un conséquentialisme hédoniste, agent-neutre, non priorisant, direct, effectif, universaliste et maximisant.

Théorie du bien Les théories utilitaristes parlent d'*utilité* pour faire référence à la valeur des conséquences d'une action. Elle s'obtient en sommant la valeur de toutes les conséquences négatives de l'action, puis en sommant la valeur de toutes les conséquences positives de l'action, pour enfin sommer ces deux résultats et ainsi obtenir la valeur totale de l'action, son utilité. L'utilitarisme de l'acte étant un conséquentialisme hédoniste, seul les expériences de plaisir et de souffrance peuvent avoir une valeur intrinsèque. En conséquence, l'utilité d'une action est la somme de la souffrance causée à laquelle s'ajoute la somme du plaisir causé.

En admettant qu'un individu sache attribuer une valeur aux expériences de plaisir et de souffrance, le calcul de l'utilité est clairement défini. Mais ce calcul est en réalité plus propre à la théorie du juste que celle du bien. Le rôle de cette dernière est justement de spécifier comment attribuer une valeur aux expériences de plaisir et de souffrance. Bentham et Mill ont chacun proposé leur propre théorie du bien pour l'utilitarisme de l'acte. Nous allons voir en quoi elles consistent et en quoi elles diffèrent.

Dans *An Introduction to the Principles of Morals and Legislation*, BENTHAM [1789] propose cinq critères principaux à prendre en compte pour attribuer une valeur aux expériences de plaisir et de souffrance : leur intensité, leur durée, leur fécondité (i.e. leur capacité à produire en chaîne une autre expérience de même valeur), leur pureté (i.e. le fait qu'une expérience d'une certaine valeur ne produise pas d'autres de la valeur opposée), leur étendue (i.e. le nombre de personnes que cela impacte). Il ajoute à cela deux critères supplémentaires qui ne sont pas uniquement focalisés sur l'expérience, mais sur la relation entre l'action et l'expérience : la certitude sur le fait que l'action amène l'expérience et la proximité temporelle entre l'action et l'expérience. Il est important de souligner que l'utilitarisme que propose Bentham suppose une commensurabilité de toutes les expériences de plaisir ou de souffrance, i.e. qu'il y a une échelle commune pour les expériences qu'elles soient physiques, intellectuelles ou esthétiques. Le plaisir que procure la capture de Cavaradossi à Scarpia peut être mis sur la même échelle que le plaisir esthétique de l'art que produit Cavaradossi. Si Scarpia pouvait retrouver ce même plaisir à chaque fois qu'il racontait la capture, après un certain nombre de fois, la valeur associée aux conséquences positives de la capture pourrait être plus élevée que la valeur des conséquences négatives. Et cela, même si Cavaradossi était sur le point de produire le premier tableau cubiste, bien avant Picasso et Braque, et que cela l'en empêchait.

Dans *Utilitarianism*, MILL [1863] critique la théorie de Bentham en la qualifiant de purement quantitative et donc non adaptée aux êtres humains, mais plutôt à des êtres moins rationnels comme les animaux. La version qu'il propose ajoute, à l'aspect purement quantitatif du calcul de la valeur des expériences, un aspect qualitatif en envisageant l'existence de types de plaisirs différents. Il défend l'existence de plaisirs plus désirables car ayant plus de

valeur que d'autres. En introduisant cette idée, certaines expériences de plaisir et de souffrance deviennent incommensurables. Il est possible de dire dans la version de Mill que, quel que soit le nombre de fois que Scarpia raconte la capture de Cavaradossi et la quantité de plaisir que cela lui procure, cela ne compensera jamais la perte artistique.

Théorie du juste Le principe moral selon lequel une action est juste pour l'utilitarisme de l'acte peut s'énoncer ainsi :

> UA [utilitarisme de l'acte] : une action est requise si et seulement si son utilité est supérieure à l'utilité de toutes les actions alternatives. Une action est interdite s'il existe une action alternative dont l'utilité est supérieure à la sienne. Une action qui n'est ni requise, ni interdite est optionnelle.

1.3.1.2 Utilitarisme de la règle

Considérons l'utilitarisme de la règle correspondant à un utilitarisme de l'acte où la seule chose qui varie est l'aspect direct. Ce conséquentialisme est indirect car il ne va pas évaluer les conséquences des actions individuelles faites dans un contexte donné pour déterminer si une action est juste ou non. Il va évaluer les conséquences qu'aurait l'adoption d'une norme morale par tous les individus. Au lieu d'évaluer si l'action spécifique faite par Tosca est juste, il faut en premier lieu regarder les conséquences de l'adoption par tous les individus de la norme morale : agir pour sauver un agent auquel un individu tient est juste. À partir de cette opération élémentaire, ce conséquentialisme propose de construire un code de conduite dit « idéal ». Un tel code de conduite correspond à un ensemble de normes morales dont les conséquences de son adoption sont supérieures ou égales à celles de tout autre ensemble de normes morales.

Il est important de préciser que l'utilitarisme de la règle et le principe de loi universelle de Kant ne sont pas la même chose. Alors que le premier s'intéresse à la valeur des conséquences de l'universalisation d'une norme morale pour savoir si une telle universalisation est souhaitable, le deuxième va plutôt regarder si l'universalisation entraîne une contradiction dans le monde obtenu.

Théorie du juste Le principe moral selon lequel une action est juste pour l'utilitarisme de la règles peut s'énoncer ainsi :

> UR [utilitarisme de la règle] une action est requise, interdite ou optionnelle si et seulement si le code de conduite idéal le stipule ainsi.

Cette version du conséquentialisme satisfait l'idée attrayante que certaines actions, comme tuer pour le plaisir, sont interdites par le type même d'action qu'elles sont, et donc en toutes circonstances. Toutefois, contrairement aux précédentes théories basées sur des codes de conduite, elle le justifie d'un point de vue conséquentialiste.

1.3.1.3 Conséquentialisme satisfaisant

Considérons le conséquentialisme satisfaisant correspondant à un conséquentialisme de l'acte où la seule chose qui varie est l'aspect maximisant. Cette version du conséquentialisme adopte l'idée qu'une action est juste si ses conséquences sont suffisamment bonnes.

Il s'agit d'une version moins exigeante. En effet, même s'il existe des alternatives meilleures, du moment où les conséquences de l'action sont suffisamment bonnes, alors l'action est juste. Poignarder Scarpia peut être juste, même s'il avait suffit de débattre avec lui durant quelques heures pour le convaincre que Cavaradossi n'est pas un danger pour la stabilité de la société.

Théorie du juste Le principe moral selon lequel une action est juste pour le conséquentialisme satisfaisant peut s'énoncer ainsi :

> CS [conséquentialisme satisfaisant] : une action est requise si c'est l'unique action dont l'utilité est supérieure ou égale au seuil d'utilité spécifié, ou si son utilité est supérieur à celle de toutes les alternatives et qu'aucune action n'a une utilité supérieure au seuil. Une action est interdite si son utilité est inférieure au seuil d'utilité spécifié et qu'il existe des actions dont l'utilité est supérieure. Une action qui n'est ni requise, ni interdite est optionnelle.

Selon cette variante, poignarder Scarpia est considéré comme juste du moment où la valeur globale est, par exemple, supérieure à zéro. Ce cas où le seuil est à zéro consiste à dire qu'une action est juste du moment où la valeur positive de ses conséquences est supérieure à leur valeur négative.

1.3.2 Théorie du droit naturel

De l'idée qu'il existe des principes moraux qui sont ancrés dans la nature humaine, découle l'idée selon laquelle ce qui est juste, bien et vertueux est valide pour tous les êtres humains. Il existerait donc un ensemble objectif de principes moraux dont l'autorité supérieure s'explique par leur source naturelle.

Thomas d'Aquin est considéré comme un des promoteurs les plus importants de ce courant de pensée. Il développe sa théorie morale dans la *Summa Theologiae* [D'AQUIN, 1266] et influence durablement la morale catholique romaine. Des différentes théories existantes adoptant cette pensée, c'est sa théorie que nous présentons.

Théorie du bien Thomas d'Aquin propose une théorie de la valeur perfectionniste comme base de sa théorie morale, d'où son appartenance au perfectionnisme moral. Il s'agit de l'idée selon laquelle ce qui a de la valeur pour nous les humains, sont les états du monde qui permettent le développement des capacités qui nous sont propres. Cela veut dire que plus un humain se rapproche d'un état de perfection, plus l'état du monde dans lequel cela se produit a de la valeur positive intrinsèque.

Pour comprendre ce qu'est l'état de perfection chez l'être humain, il faut remonter à la théologie grecque et faire le lien avec le droit naturel. D'après l'héritage d'Aristote, pour comprendre l'essence d'une chose il faut comprendre ses fins. Dès lors que les fins des êtres humains sont claires, l'état de perfection peut être atteint si ces fins sont pleinement développées. Thomas D'Aquin fait reposer sa théorie du bien perfectionniste sur des faits concernant la nature humaine. C'est cette nature commune dont nous ne pouvons pas nous détacher qui fait que certains états du monde ont une valeur intrinsèque pour nous.

L'étude des fins caractéristiques des êtres humains par Thomas D'Aquin lui font conclure qu'il existe quatre valeurs de base : la vie (humaine), la procréation, la connaissance et la

sociabilité. Le processus par lequel il arrive à cette conclusion appartenant exclusivement à l'objectif théorique de la théorie morale, nous ne nous attarderons pas dans les détails du raisonnement.

Théorie du juste Pour Thomas d'Aquin, il est du devoir des humains de comprendre et développer les capacités qui nous rapprochent de l'état de perfection étant donné que nous sommes des agents rationnels dotés de libre arbitre. Dans la mesure où la vie, la procréation, la connaissance et la sociabilité sont les valeurs de base, elles doivent être préservées et promues, aller à leur encontre doit être évité.

Le principe moral selon lequel une action est juste pour la théorie du droit naturel de Thomas d'Aquin peut s'énoncer ainsi :

> TDN [théorie du droit naturel] : une action est requise si et seulement si ne pas la réaliser résulte en une violation directe d'au moins une des valeurs de base. Une action est interdite si et seulement si la réaliser résulte en une violation directe d'au moins une des valeurs de base. Une action qui n'est ni requise, ni interdite est optionnelle.

En pratique, il est possible de considérer qu'une action est une violation directe d'une valeur de base si l'action ne peut pas être justifiée par la Doctrine du double effet. Composante essentielle à cette théorie morale, ce principe indique quand il est juste de faire une action qui promeut ce qui est mal alors que l'objectif est de promouvoir ce qui est bien. Plus généralement, ce principe permet de déterminer le statut déontique d'une action qui apporterait à la fois du bien et du mal dans le monde. C'est le cas des actions de Tosca, d'un côté elle pense sauver une vie, de l'autre elle trompe un individu et le tue. Cependant, comme mentionné ci-dessus, ce principe permet en fin de compte d'évaluer le statut déontique de toutes les actions. D'après ce principe, une action est juste si les conditions suivantes sont satisfaites :

1. *L'action ne doit pas être intrinsèquement interdite.* Cette première condition est une des composantes les plus distinctives de cette théorie, notamment par rapport aux autres théories axées sur la valeur. Il s'agit de l'absolutisme moral, idée selon laquelle certaines actions ne sont jamais justes quel que soit le contexte et les conséquences qui résulteraient de leur réalisation. Le statut déontique de ces actions est toujours le même, elles sont interdites.
   Le suicide ou l'homicide étant des actions consistant à aller à l'encontre de la vie de façon intentionnelle, ces actions sont interdites quel que soit le contexte.
   La procréation fait référence aussi bien à la conception qu'à l'éducation d'un enfant. D'après Thomas d'Aquin, l'adultère, les moyens artificiels de contraception et l'homosexualité vont à l'encontre de la procréation, ces actions sont interdites quel que soit le contexte.
   La connaissance fait référence à la connaissance de dieu, mais aussi du monde qui nous entoure et donc de sa création. Brûler des livres sacrés est donc interdit quel que soit le contexte.
   La sociabilité fait référence à la capacité des êtres humains à vivre en groupe et coopérer pour leur bénéfice mutuel. Des relations entre individus comme l'amitié ou le mariage sont des spécifications de cette valeur. Pour Thomas d'Aquin, mentir ou trahir est interdit quel que soit le contexte.

Aussi bien tromper Scarpia que le tuer ne satisfont pas cette première condition, ces deux actions sont intrinsèquement interdites. Selon cette théorie morale, les actions de Tosca sont une violation directe de la valeur vie et sociabilité, elle sont donc interdites.

2. *Le mal engendré par l'action ne doit pas être intentionnel.* C'est considéré comme étant le cas si : soit l'effet étant mauvais est une des fins de réaliser l'action, soit l'effet étant mauvais est un moyen pour arriver à une des fins de réaliser l'action. La fin de Tosca n'était pas la mort de Scarpia. Toutefois, sa mort peut être vue comme un moyen pour lui échapper et sauver Cavaradossi. Le mal qu'elle engendre est alors intentionnel aux yeux de la doctrine du double effet.
3. *Il y a une raison proportionnellement suffisante pour engendrer le mal.* C'est le cas s'il n'existe pas d'action alternative qui apporterait autant de bien sans apporter autant de mal et que le mal engendré n'est pas disproportionné par rapport au bien recherché. Si nous imaginons que toutes les autres alternatives n'avaient aucune chance d'aboutir au sauvetage de Cavaradossi, il est possible de dire qu'il y avait une raison proportionnellement suffisante pour engendrer le mal. En fin de compte, Tosca pense prendre une vie pour en sauver une autre.

1.4 Théories axées sur la vertu

Cette section traite des théories morales où le concept de vertu est antérieur au concept de devoir. Elles partagent l'idée qu'il est possible, sans faire appel à la notion de juste, d'attribuer une valeur aux agents. Dans toutes ces théories morales, le concept de vertu est utilisé pour déterminer le statut déontique d'une action. Dans cette section nous présentons une théorie axée sur la vertu très générale de sorte à introduire les fondements communs à ces théories.

De l'idée qu'une action est juste parce que c'est ce que ferait une personne vertueuse, découle l'idée selon laquelle pour déterminer ce qui est juste il est possible d'adopter une procédure de décision qui s'appuie sur ce que ferait et ne ferait pas une personne vertueuse. L'éthique de la vertu met en avant la notion de vertu aussi bien dans l'objectif pratique que théorique.

L'origine de cette famille de théories remonte à Platon et Aristote pour qui ce qui devait être recherché en éthique était le type de personne qu'il fallait être. La vertu occupait la place du juste dans l'objectif pratique. Pour mieux comprendre pourquoi la vertu occupait cette place, il faut comprendre les fondements de l'éthique d'Aristote. Le philosophe défend l'idée que toutes les activités humaines tendent vers un ensemble de fins et que pour comprendre l'essence d'une chose il faut comprendre ses fins. L'humain tendant naturellement vers ces fins, elles peuvent être considérées comme étant le bien. Aristote croit en l'existence d'un bien supérieur aux autres, le bonheur. Celui-ci est identifiable parmi les autres car toutes les autres fins sont poursuivies pour arriver à ce bien supérieur.

Théorie du bien Le philosophe grec utilise le bonheur comme valeur positive intrinsèque pour justifier les traits de caractères qui seront considérés comme vertus. De ce fait, dans la théorie d'Aristote la notion de valeur est plus élémentaire que la notion de vertu, ce qui n'empêche pas que ce soit une théorie axée sur la vertu.

Théorie de la vertu Une vertu est définie comme un trait de caractère, comptant dans l'évaluation positive d'un agent, qui implique qu'il a tendance à penser, sentir et agir d'une façon particulière dans des situations données. Un vice est défini comme un trait de caractère, comptant dans l'évaluation négative d'un agent, qui implique qu'il a tendance à penser, sentir et agir d'une façon particulière dans des situations données. Une vertu ou un vice ne se résume donc pas aux simples actions de l'agent, il y a également une composante intellectuelle, affective et motivationnelle.
Dans *In a Different Voice*, GILLIGAN [1982] pose les bases de l'éthique de la sollicitude (« ethics of care » en anglais), où sollicitude est en quelque sorte l'attention portée à autrui à laquelle s'ajoute la notion de soin. Prenons l'exemple de l'éthique de la sollicitude pour illustrer les différentes composantes qu'il faut posséder pour être considéré comme ayant cette vertu. La composante intellectuelle demande d'avoir la capacité de savoir quand est-ce qu'une personne est dans le besoin, mais également de savoir ce dont elle a besoin dans le contexte donné. La composante affective demande d'avoir la capacité de se réjouir des réussites des autres et de déplorer leurs échecs. Enfin, la composante motivationnelle demande d'avoir la capacité de désirer aider les autres de façon désintéressée.

Certains philosophes considèrent qu'il existe une hiérarchie entre les vertus, certaines seraient plus basiques que d'autres. De nombreux philosophes grecs considéraient qu'il y avait quatre vertus basiques desquelles découlaient les autres : la sagesse, le courage, la tempérance et la justice. Parmi les nombreuses vertus qui ont été proposées dans l'histoire de l'éthique, quelques vertus morales le plus souvent reprises sont : la bienfaisance, la conscienciosité, le courage, la générosité, la gratitude, la justice, l'honnêteté, la loyauté et la tempérance.

Les vertus sont considérées comme acquises par l'exposition à de nombreuses situations où l'agent a la possibilité d'incarner cette vertu. Il ne faut donc pas confondre vertus et caractéristiques propres à un agent par naissance. Les vertus s'acquièrent par la pratique, ce qui justifie d'autant plus leur valeur.

Théorie du juste L'éthique de la vertu est souvent qualifiée comme étant centrée sur l'agent au lieu d'être centrée sur les actions, contrairement aux autres familles de théories morales. Cela peut faire penser que ces théories ne permettent pas de déterminer le statut déontique, ce qui est faux.

Le principe moral selon lequel une action est juste pour la théorie de la vertu peut s'énoncer ainsi :

> EV [éthique de la vertu] : une action est requise si et seulement si c'est une action qu'une personne vertueuse n'omettrait pas de faire dans ces circonstances. Une action est interdite si et seulement si c'est une action qu'une personne vertueuse omettrait de faire dans ces circonstances. Une action qui n'est ni requise, ni interdite est optionnelle.

Dans le principe EV il est supposé que la personne vertueuse dont il est question connaît tous les éléments pertinents pour agir et que dans tous les cas elle agira de façon vertueuse.

1.5 Un bref aperçu de l'éthique computationnelle

Cette section présente la structure générale du domaine qu'est l'éthique computationnelle et établit le périmètre de cette thèse en précisant que nous nous intéressons exclusivement aux approches qui cherchent à formaliser les théories morales provenant de l'éthique normative.

L'*éthique computationnelle* est un sous-domaine de l'intelligence artificielle. Il se préoccupe d'assurer que des systèmes informatiques qui automatisent au moins une partie d'un processus décisionnel prennent des décisions qui puissent être vues comme éthiques par les humains. D'autres noms, comme « moralité artificielle » ou « éthique des machines » [TOLMEIJER et collab., 2021], sont parfois utilisés pour y faire référence. L'émergence et développement de ce domaine est principalement dû à la prolifération d'outils du quotidien utilisant des résultats obtenus dans la discipline scientifique qu'est l'intelligence artificielle. Cette diffusion, couplée à des discours avertissant de certains dangers associés à leur utilisation, a suscité une demande croissante pour des outils « dignes de confiance », i.e. offrant des garanties selon lesquelles les décisions prises sont transparentes, explicables et éthiques. L'éthique computationnelle a pour objectif de répondre à la dernière garantie : que les décisions puissent être vues comme éthiques. Comme nous l'avons vu tout au long de ce chapitre, déterminer si une décision prise par un humain est éthique n'est pas trivial. Cette question a fait, et fait toujours, l'objet de nombreux débats en éthique normative, cette composante de la discipline philosophique qu'est l'éthique.
Les solutions proposées en éthique computationnelle pour répondre à ce défis sont usuellement classifiées dans trois grandes familles [ALLEN et collab., 2005] : les approches descendantes ou prescriptives (« top-down » en anglais), les approches ascendantes ou descriptives (« bottom-up » en anglais) et les approches hybrides. Les approches descendantes consistent à formaliser un raisonnement pour qu'il puisse être réalisé par des systèmes informatiques. Derrière ces approches existe l'hypothèse que le raisonnement à formaliser est suffisamment compris pour pouvoir être implémenté dans une machine. Ces approches reposent sur des méthodes qui se prêtent au raisonnement déductif. Les approches ascendantes consistent à apprendre comment reproduire le raisonnement à partir d'observations pour qu'il puisse ensuite être réalisé par des systèmes informatiques. Derrière ces approches existe l'hypothèse que le raisonnement à formaliser peut être entièrement déduit de faits observables. Ces approches reposent donc sur des méthodes qui se prêtent au raisonnement inductif. Finalement, les approches hybrides sont le produit d'une combinaison entre les approches descendantes et ascendantes. Les hypothèses derrière ces approches comme les méthodes utilisées sont une combinaison propre à chacune. TOLMEIJER et collab. [2021] explique que l'éthique computationnelle est un domaine archipélisé, les familles se développent individuellement et communiquent peu entre elles.

Dans cette thèse nous n'allons pas déroger à cette tendance et allons exclusivement nous intéresser aux approches descendantes. Trois raisons à cela : l'éthique sied plus aux approches descendantes ; les approches ascendantes doivent surmonter plus d'obstacles pour satisfaire toutes les garanties demandées pour avoir un outil digne de confiance ; l'approche descendante permet de mieux comprendre la théorie morale formalisée. Nous allons détailler chaque raison une par une.
Pour expliquer pourquoi les approches descendantes sont plus aptes à remplir l'objectif de l'éthique computationelle, rappelons l'objectif recherché par le domaine : assurer que des

systèmes informatiques qui automatisent au moins une partie d'un processus décisionnel prennent des décisions qui puissent être vues comme éthiques par les humains. Étudions ce dernier aspect. Nous défendons que les machines ne peuvent pas être des agents pleinement éthiques. Pour cela il faudrait qu'elles puissent être considérées comme des agents dans le sens de la philosophie de l'action où le principe d'intentionnalité est primordial. Comme le montrent CHANGEUX et CONNES [2008] elles en sont complètement dépourvues. Ne pouvant pas définir leur propre loi, elles ne peuvent pas être à l'initiative de l'action et donc ne sont pas des agents dans ce sens là. Elles ne sont donc pas non plus autonomes dans le sens étymologique du terme qui est « le fait de se donner à soi-même sa loi », mais automatiques ou d'une « autonomie technique » tout au plus. De ce fait, elles peuvent au plus prendre des décisions vues comme éthiques par les humains. Pour remplir l'objectif de l'éthique computationnelle il faut donc s'intéresser à l'éthique pour les êtres humains et donc à l'éthique normative. Comme nous l'avons vu tout au long de ce chapitre, l'éthique occupe les philosophes depuis au moins vingt-quatre siècles si nous prenons Aristote comme point de départ. Il n'est pas insensé de considérer que nous disposons de suffisamment d'éléments pour pouvoir implémenter le raisonnement éthique dans une machine et donc que l'hypothèse derrière les approches descendantes est satisfaite. Qu'en est-il de l'hypothèse des approches ascendantes ? Pour rappel, ces approches assument que le raisonnement à reproduire peut être entièrement déduit de faits observables. Cette assomption est dangereuse lorsqu'il est question d'éthique. En effet, accepter cette hypothèse est une pente glissante menant vers la guillotine de HUME [1748], aussi connue comme le « is-ought problem » [KIM et collab., 2021; SINGER, 2015]. Celui-ci consiste à déduire ce qui doit être uniquement de ce qui est, donc ici des obligations morales de simples faits non moraux : « No normative truth is determined by any non-normative truths » ou « [...] the is-ought gap says that we cannot get from non-normative premises to conclusions about how things ought to be in cases where the premises are true » [SINGER, 2015]. L'hypothèse des approches ascendantes est donc plus difficilement acceptable.
Ensuite, les approches ascendantes doivent surmonter plus d'obstacles pour satisfaire toutes les garanties demandées pour avoir un outil digne de confiance. Pour rappel, cela demande que les décisions prises par les outils soient transparents et explicables en plus d'être éthiques. Les approches descendantes sont par nature propices à être explicables et transparentes. Au contraire, les méthodes d'apprentissage machine et d'apprentissage profond sont par nature des boîtes noires qui ne peuvent devenir transparentes et explicables que par l'ajout de mécanismes. Ces méthodes aujourd'hui plébiscitées dans cette famille d'approches ne remplissent pas ces conditions a priori.

> A lack of engagement with philosophical literature also makes automated ethics less explainable, as seen in the example of Delphi, which uses deep learning to make moral judgements based on a training dataset of human decisions [JIANG et collab., 2021]. Early versions of Delphi gave unexpected results, such as declaring that the user should commit genocide if it makes everyone happy [VINCENT, 2021]. Because no explicit ethical theory underpins Delphi's judgements, we cannot determine why Delphi thinks genocide is obligatory. Machine learning approaches like Delphi often cannot explain their decisions. This reduces human trust in a machine's controversial ethical judgements. The high stakes of automated ethics require explainability to build trust and catch mistakes. [SINGH, 2022]

Finalement, les approches descendantes ont l'avantage de permettre de mieux comprendre la théorie morale formalisée. En effet, le processus de formalisation passe nécessairement par une modélisation des concepts qui composent le raisonnement et du raisonnement lui même. La rigueur mathématique de la modélisation peut faire apparaître des voies non explorées, des nuances ou des incohérences. Tout au long de cette thèse nous en verrons quelques exemples, aussi bien dans la modélisation du raisonnement éthique que du raisonnement causal. L'intelligence artificielle n'est pas le seul domaine à profiter du point de vue apporté par la philosophie dans l'éthique computationnelle, la philosophie peut aussi se voir bénéficié. C'est pourquoi nous allons principalement nous intéresser dans cette thèse aux approches qui formalisent des théories morales. Nous parlons alors d'un sous-domaine de l'éthique computationnelle que nous appelons *éthique computationnelle normative*.

Cette première partie consacrée à l'état de l'art ne contient pas de chapitre plus détaillé concernant l'éthique computationnelle. La première raison à cela est que TOLMEIJER et collab. [2021] ont récemment proposé un large aperçu de tout le domaine de l'éthique computationnelle. Nous invitons les lecteurs qui voudraient approfondir cet aperçu global du domaine à s'y référer. La deuxième raison est que dans le chapitre 4 nous proposons un cadre commun pour la formalisation du raisonnement éthique qui se prête à la réalisation d'une étude comparative des travaux en éthique computationnelle normative. Nous avons donc opté pour omettre de faire un état de l'art très général déjà existant pour plutôt proposer une étude comparative plus approfondie sur une partie du domaine en particulier. Cette étude comparative est faite dans la section 4.3.

1.6 Conclusion

Dans ce chapitre nous avons introduit les concepts de base nécessaires à la compréhension et à la structuration de la théorie morale. Puis, nous avons présenté quelques unes des théories morales occidentales les plus importantes. En premier ont été présentées les théories axées sur le devoir, en deuxième celles axées sur la valeur, pour finir sur celles axées sur la vertu. La présentation de toutes ces théories morales a permis de donner un large aperçu des structures existantes. Cet aperçu est d'une grande utilité pour cette thèse puisqu'elle s'inscrit dans le domaine de l'éthique computationnelle qui se préoccupe d'assurer que des systèmes informatiques qui automatisent au moins une partie d'un processus décisionnel prennent des décisions qui puissent être vues comme éthiques par les humains. Comme nous l'avons vu tout au long de ce chapitre, déterminer si une décision prise par un humain est éthique n'est pas trivial. Cette question a fait, et fait toujours, l'objet de nombreux débats en éthique normative. Dans la dernière partie de ce chapitre nous avons donné un aperçu du domaine qu'est l'éthique computationnelle et avons vu trois types d'approches utilisées pour essayer d'atteindre cet objectif. Nous avons exposé les trois raisons principales qui nous poussent à nous intéresser exclusivement aux approches descendantes dans le cadre de cette thèse et plus précisément, aux approches qui formalisent des théories morales. Nous avons alors délimité le sous-domaine de l'éthique computationnelle qui nous intéresse, l'éthique computationnelle normative.
Deux remarques importantes pour la suite doivent être faites avant de conclure ce chapitre :

— L'éthique est sensible au contexte. Cela veut dire que le statut déontique d'une action dépend en partie de faits non moraux reliés au contexte. Ces faits non moraux peuvent

aussi bien être reliés aux agents qu'aux circonstances. Cela s'applique non seulement au statut déontique mais aux principes moraux en général. Par conséquent, à l'heure d'être appliquées, toutes les théories morales requièrent que l'agent dispose d'une bonne capacité à identifier les détails propres à chaque cas, les subtilités du problème éthique.
C'est le cas pour l'application des codes de conduite dans la théorie du commandement divin ou le relativisme moral, pour la construction de maximes et leur universalisation dans la théorie morale de Kant, pour l'attribution de valeurs aux conséquences des actions dans le conséquentialisme, pour appliquer la doctrine du double effet dans la théorie du droit naturel et pour développer toutes les composantes permettant d'avoir une vertu dans l'éthique de la vertu.

— Dans la suite de cette thèse nous ne traiterons pas de théories morales axées sur la vertu. Malgré le fait que l'éthique de la vertu propose une procédure pour déterminer le statut déontique d'une action, elle reste particulièrement centrée sur l'agent contrairement aux autres familles de théories morales. Les différentes composantes nécessaires pour considérer qu'un agent est vertueux en témoignent.

Chapitre 2

État de l'art : modélisation et représentation de l'action et du changement

« *The Toronto group's approach [Reiter, R., Pinto, J., Lin, F., Levesque, H., Lesperance, Y., and Scherl, R.] and the approach based on $\mathscr{A}$ share the same view of a dynamic world [...]. The difference is in the type of language used to describe actions and in the type of logic associated with this language. The former uses general purpose classical logic while the latter prefers a special purpose, high level language with nonmonotonic, specialized semantics.* »

BARAL et GELFOND [1997]

Sommaire

Ce chapitre se veut un aperçu de ce qu'est la modélisation et la représentation de l'action et du changement. Décrire des changements causés par l'exécution d'actions est un des premiers problèmes que les chercheurs en intelligence artificielle ont essayé de résoudre. L'étude de ce problème est indispensable pour cette thèse dans le domaine de l'éthique computationnelle qui cherche à formaliser notre capacité en tant qu'êtres rationnels à évaluer moralement une action. En effet, comme nous l'avons vu dans le chapitre 1, de nombreuses théories demandent un raisonnement sur l'état du monde, les actions des agents et leurs conséquences. Plus généralement, tout raisonnement éthique demande une certaine représentation du monde. C'est ce à quoi nous nous intéressons dans ce chapitre.

Faisons un bref rappel de ce qu'est formaliser. Le processus de *formalisation* consiste en trois étapes [SAINT-CYR et collab., 2014] : *modéliser*, *représenter* et *automatiser*. L'étape de modélisation consiste à définir rigoureusement les concepts et les processus d'un point du vu mathématique. L'étape de représentation consiste à indiquer comment les concepts doivent être codés pour être traités par un ordinateur. Finalement, l'étape d'automatisation consiste à indiquer comment les processus sont reproduits par des algorithmes. Dans ce chapitre il est question de modélisation et de représentation.
Il existe différents modèles d'action et de changement. Dans ce chapitre et tout au long de cette thèse nous nous intéressons uniquement à des modèles discrets. Dans ce cadre, le modèle le plus courant est le système de transition d'états. Il existe également différentes représentations de l'action et du changement. Dans ce chapitre nous présentons différents langages qui permettent de coder un tel modèle. Plus exactement, nous présentons le Planning Domain Description Language (PDDL), le langage sur lequel repose The Stanford Research Institute Problem Solver (STRIPS), le langage de description d'action $\mathscr{A}$, $\mathscr{B}$ et $\mathscr{C}$, le Calcul des Situations et le Calcul des Évènements. Nous avons choisi de parler de ces langages spécifiquement car ils nous permettent de situer notre contribution dans la représentation de l'action et du changement, mais aussi dans la représentation de la causalité.

Un modèle complet de raisonnement sur l'action et le changement est constitué de deux parties : un modèle sur le système avec ses aspects physiques et un modèle de l'agent avec ses croyances. Bien que dans la chapitre 1 nous ayons vu que pour certaines théories morales il est nécessaire d'avoir le modèle complet, représenter le modèle de l'agent de façon satisfaisante représente un vrai défi auquel des domaines entiers de recherche sont consacrés. Cela dépasse le cadre de cette thèse, mais reste une voie importante à explorer dans de travaux futurs. Dans la suite de ce chapitre nous parlerons donc exclusivement du modèle sur le système. Dans le chapitre 4 où il sera question de modélisation de différentes théories éthiques, nous parlerons de concepts propres au modèle de l'agent. Toutefois, dans cette thèse ces concepts ne passeront jamais à l'étape de représentation et encore moins d'automatisation.

Ce chapitre est divisé en deux sections. La section 2.1 décrit le modèle le plus courant, le système de transition d'états. La section 2.2 présente différents langages qui permettent de représenter un tel modèle. La proposition de GELFOND et LIFSCHITZ [1998] peut être vue comme une structure permettant de présenter de façon concise la syntaxe et la sémantique d'un langage de description d'action. Cette section s'inspire de ce travail. Pour commencer nous présentons le PDDL qui nous permet d'avoir un aperçu global. Puis, nous reprenons et adaptons la proposition de GELFOND et LIFSCHITZ [1998] pour STRIPS, $\mathscr{A}$, $\mathscr{B}$ et $\mathscr{C}$, et nous proposons une présentation selon la même structure du Calcul des Situations et du Calcul des Évènements.

2.1 Modèle : système de transition d'états étiqueté

Dans cette section nous présentons le modèle d'action et de changement le plus courant, le Système de Transition d'États Étiqueté (STEE). Le choix de ce modèle se justifie par l'information que nous avons à notre disposition et ce que nous souhaitons en faire. En effet, l'éthique computationnelle cherchant à formaliser notre raisonnement éthique, il semble pertinent d'avoir un modèle de l'évolution du monde qui se rapproche le plus de notre vision intuitive de ce phénomène lorsqu'il est mobilisé pour un tel raisonnement.

La structure classique d'un système de transition d'états étiqueté est la suivante. D'un côté, l'état du monde peut être défini comme une collection de variables le décrivant. De l'autre, les transitions entre états du monde est le résultat de l'occurrence d'un ensemble d'actions. L'évolution du monde peut être vue comme une suite d'états qui s'enchaînent les uns après les autres au fur et à mesure que des actions se produisent.
Nous notons $\mathbb{F}$ l'ensemble de variables décrivant l'état du monde. Ces variables représentent classiquement les propriétés du monde pouvant varier dans le temps, elles sont appelées des *fluents*. Un *État* dans un STEE est généralement noté S de l'anglais « state », et est défini comme un ensemble de fluents. Selon le STEE, des conditions différentes doivent être satisfaites pour que l'ensemble puisse être un état.
Puis, nous notons $\mathbb{A}$ l'ensemble de variables décrivant les transitions. Ces variables sont appelées des *actions*. Comme nous n'avons ici pas de modèle d'agent, les effets des actions sont uniquement des *effets ontiques*, i.e. des effets qui portent sur le monde physique et non sur l'état mental des agents ; nous ne représenterons pas les possibles *effets épistémiques*.
Si une action se produit dans un état du monde elle mène à un autre état. L'état d'arrivée dépend bien sûr de l'action, mais aussi de l'état de départ. L'hypothèse de Markov consiste à considérer que l'état de départ est le seul état nécessaire pour déterminer celui d'arrivée, « hypothèse qui est sans perte de généralité car tout système peut être modélisé de façon markovienne [...] » [Saint-Cyr et collab., 2014]. Tout système de transition est un système markovien. Si en plus de markovien le système est déterministe, l'état d'arrivée est unique.

Définition 2.1 [*Signature d'action*]**.** *Une signature d'action est un couple* $\langle \mathbb{F}, \mathbb{A} \rangle$ *où* $\mathbb{F}$ *est un ensemble de fluents et* $\mathbb{A}$ *est un ensemble d'actions.*

Parfois, la signature d'action est décrite comme un triplet contenant en plus un ensemble de valeurs $\mathbb{V}$ correspondant aux valeurs que peuvent prendre les fluents. Dans cette thèse nous ne parlerons que de langages où les fluents prennent les valeurs de vérité classiques en logique, $\mathbb{V} = \{false, true\}$. Toutes les signatures d'action que nous verrons sont propositionnelles, d'où notre choix de faire abstraction de cet ensemble.

Dans la suite, nous parlons de fluents et de leur négation. Nous introduisons les littéraux de fluents pour simplifier certaines notations. Un *littéral de fluent* est soit un fluent $f \in \mathbb{F}$, ou sa négation $\neg f$. L'ensemble des littéraux de fluents dans $\mathbb{F}$ est noté $Lit_{\mathbb{F}}$, il est défini par $Lit_{\mathbb{F}} = \mathbb{F} \cup \{\neg f \mid f \in \mathbb{F}\}$. Le complément d'un littéral de fluent l, noté $\overline{l}$, est défini comme $\overline{l} = \neg f$ si $l = f$, ou $\overline{l} = f$ si $l = \neg f$. Par extension, le complément d'un ensemble de littéraux $\mathrm{L} \subseteq Lit_{\mathbb{F}}$, noté $\overline{\mathrm{L}}$, est défini comme $\overline{\mathrm{L}} = \left\{\overline{l} | l \in \mathrm{L}\right\}$. Finalement, un ensemble de littéraux de fluents L est dit cohérent si $\forall l \in \mathrm{L}, \overline{l} \notin \mathrm{L}$

Définition 2.2 [*Système de transition d'états étiqueté* $\mathscr{S}$]**.** *Un système de transition d'états étiqueté pour une signature d'action est un triplet* $\langle \mathbb{S}, V, \tau \rangle$ *composé de :*

1. *un ensemble d'états* $\mathbb{S}$ *;*
2. *une fonction* V*:* $Lit_{\mathbb{F}} \times \mathbb{S} \rightarrow \{false, true\}$
3. *un ensemble de relations étiquetées de transition entre états* $\tau \subseteq \mathbb{S} \times 2^{\mathbb{A}} \times \mathbb{S}$.

Pour faciliter l'homogénéisation et les notations dans ce chapitre, nous choisissons de décrire les états comme un ensemble de littéraux de fluents. La notation classique utilisant plutôt un ensemble de fluents a l'avantage d'être plus concise à premier abord, mais a le désavantage d'alourdir considérablement la définition des conditions que les τ doivent respecter pour les différents langages que nous présentons. Nous considérons des états $S \in \mathbb{S}$ *cohérents* ($\forall l \in S, \bar{l} \notin S$) et *complets* ($\forall f \in \mathbb{F}, f \in S$ ou $\neg f \in S$). $V(l, S)$ donne la valeur de vérité du littéral de fluent l dans l'état S. Les états S' tel que $(S, A, S') \in \tau$ sont les résultats possibles lorsque l'ensemble d'actions A se produit dans S. Il est couramment dit que A est exécutable dans S s'il existe au moins un état d'arrivée S'. Comme nous l'avons mentionné précédemment, si le cadre est déterministe, S' est unique pour un A et un S donné.

Un STEE peut être vu comme un graphe orienté étiqueté où les états sont les nœuds et les arêtes sont les actions. Dans ce graphe, un triplet de deux nœuds reliés par une arête est une relation étiquetée de transition entre états appartenant à l'ensemble τ.

2.2 Représentation : langages d'action et du changement

Dans cette section nous présentons différentes représentations de l'action et du changement. Si le choix du modèle doit reposer sur l'information à disposition et ce que l'on souhaite en faire, le choix de la représentation à utiliser repose sur différents critères concernant l'efficacité du langage, comme son expressivité ou sa facilité d'utilisation. Le choix des langages que nous présentons dans cette section se justifie par le fait qu'ils nous permettent de situer notre contribution dans la représentation de l'action et du changement, mais aussi dans la représentation de la causalité.

Avant de décrire certains langages permettant de représenter un STEE, faisons un point sur les principales difficultés auxquelles ils doivent faire face. Nous nous intéresserons à deux d'entre elles : le *problème du décor* (« frame problem » en anglais) et le *problème de ramification* (« ramification problem » en anglais).
Le problème du décor concerne les effets des actions, et plus exactement les effets qu'une action n'a pas. Dans un système de transition, l'état d'arrivée est uniquement déterminé par l'état de départ et les actions réalisées. Lorsqu'une représentation des effets d'une action est faite, il faudrait pour être exhaustif spécifier ce qui change et ce qui ne change pas. Prenons l'exemple d'un agent qui veut allumer la lumière dans une pièce avec des murs bleus en appuyant sur un interrupteur. Pour pouvoir représenter l'évolution du monde, représenter l'action consistant à appuyer sur l'interrupteur demande que dans les effets il y ait le fait de changer l'état de la lumière, mais aussi de spécifier que suite à l'action la couleur des murs ne change pas. Plus le nombre de fluents pour décrire l'état du monde est grand, plus cette contrainte devient problématique. Une des solutions possibles à ce problème, celle adoptée par tous les langages que nous présentons, est de considérer que les fluents sont soumis à l'inertie, i.e. si leur valeur n'est pas affectée par l'effet des actions de la transition, alors dans

l'état d'arrivée leur valeur est la même que dans l'état de départ. De cette façon, il est uniquement nécessaire de spécifier ce que l'action change dans le monde.
Le problème de ramification concerne également les effets des actions. Lorsqu'une représentation des effets d'une action est faite, il faudrait pour être exhaustif spécifier tout ce qui change, directement et indirectement, suite à la réalisation d'une action. Prenons l'exemple d'un agent qui veut allumer la lumière dans une pièce en appuyant sur un interrupteur. Il est possible que l'agent cause un court-circuit et déclenche un incendie car l'installation électrique était défectueuse. Lorsque l'action consistant à appuyer sur l'interrupteur est représentée, il n'est pas intuitif de prendre en compte le court-circuit puis l'incendie comme faisant partie de ses effets intrinsèques. Différentes solutions ont été proposées pour résoudre ce problème également, nous en verrons deux tout au long de ce chapitre. Cependant, les deux consistent en bref à ajouter des éléments au langage pour faire en sorte que les changements indirects ne doivent plus être spécifiés comme des effets intrinsèques de l'action.

Dans la section 2.2.1 nous présentons le PDDL qui par sa volonté d'être une représentation commune nous permet d'avoir un aperçu général des différents concepts qui seront abordés dans les autres langages. Dans la section 2.2.2 nous présentons le STRIPS, dans les sections 2.2.3, 2.2.4 et 2.2.5 le langage de description d'action $\mathscr{A}$, $\mathscr{B}$ et $\mathscr{C}$ respectivement, dans la section 2.2.6 le Calcul des Situations et dans la section 2.2.7 le Calcul des Évènements.

2.2.1 Planning Domain Description Language

Le PDDL a été introduit par GHALLAB et collab. [1998]. Il s'agit d'un langage formel de représentation des connaissances qui a été développé progressivement par la communauté de la planification et qui est utilisé dans les différentes éditions de la « International Planning Competition ». C'est le résultat d'un travail commun pour faciliter l'interchangeabilité dans le domaine. Il a été développé comme un langage commun pour représenter un problème de planification. L'objectif est que, même si chaque approche de planification utilise un langage différent, le fait de partir du même problème facilite leur comparaison. Pour cette raison, le langage a été structuré autour des différents concepts qu'il permet de représenter appelés « requirements ». Chaque approche de planification peut alors identifier clairement ce qu'elle peut ou ne peut pas traiter. L'état de maturité avancé de PDDL, sa vocation à faciliter l'interchangeabilité, et son utilisation par une large communauté, sont autant d'arguments significatifs en faveur de ce formalisme graduellement étendu par différents fragments [GHALLAB et collab., 1998; HASLUM et collab., 2019].

Une présentation détaillée de ce langage et de tous les concepts qu'il permet de représenter est en dehors du cadre de cette thèse. Pour les lecteurs intéressés, HASLUM et collab. [2019] proposent une présentation très complète et claire de celui-ci. Toutefois, nous introduisons certains des concepts et leur syntaxe au travers de l'exemple 2.1.

Exemple 2.1 [voiture hétéronome]**.** *Afin d'illustrer au mieux ce que ces différents concepts permettent de représenter, nous allons prendre un exemple illustratif. Il s'agit d'une simplification d'un cas de voiture à conduite automatisée, ou voiture « hétéronome » si nous voulons insister sur leur non autonomie, où les actions qui peuvent être réalisées sont freiner, accélérer, tourner à droite, tourner à gauche ou ne rien faire. De plus, le monde dans lequel se situe le véhicule est discrétisé et simplifié. En effet, nous supposons que le véhicule est initialement*

sur une case main_0, qu'il existe deux cases qui se succèdent en face de lui, main_1 et main_2, et qu'il a une case à sa droite et à sa gauche, rightSide_1 et leftSide_1. À cela nous ajoutons la possibilité qu'il y ait des personnes, des vélos, des voitures, des bus et autres éléments sur ces différentes positions, chacun avec un nombre de passagers précisé. L'objectif d'une formalisation du raisonnement éthique serait de pouvoir évaluer chacune des actions dans le contexte spécifié et de déterminer leur statut déontique selon diverses théories morales. Pour cela, il est nécessaire que nous connaissions les évènements qui arrivent, à quel moment ils arrivent et quelles sont leurs conséquences. Pour commencer, décrivons en PDDL une des possibles actions que peut réaliser le véhicule. Pour faciliter la compréhension, nous représentons ce qui identifie les préconditions, les effets et le type d'évènement en bleu clair. Nous représentons les opérateurs en bleu foncé.

```
(:action accelerate
    :precondition (on vha main_0)
    :effect (and (not (on vha main_0))
                 (on vha main_2)
                 (when (not (moving)) (moving))
            )
)
```

Il s'agit ici de l'action `accelerate`*. De par la façon dont nous avons définit le problème, cette action ne peut être réalisée que si le véhicule est sur* `main_0`*, élément que nous retrouvons dans les préconditions. Cette action à plusieurs effets, effets qui sont conjonctifs* `(and <conjunct1> ... <conjunctN>)`*. Lorsque le véhicule accélère, nous allons supposer qu'au temps d'après il sera à la position* `main_2` *et par conséquent ne sera plus à sa position initiale. De plus, cette action a un effet conditionnel,* `(when <condition> <effect>)`*, spécifiant que si le véhicule était à l'arrêt et qu'il accélère, nous considérerons par la suite qu'il est en mouvement. Les autres actions ne présentant pas de grandes différences avec celle présentée ci-dessus, ni un intérêt particulier pour l'illustration des capacités du PDDL, nous n'allons pas les détailler ici.*

Intéressons nous plutôt aux actions de type `:event`*. Elles ont été introduites par* Fox *et* Long *[2006] dans le PDDL+ et elles font partie du requirement* `:time` *tout comme les actions de type* `:process`*. Contrairement aux actions, les* `:event` *et les* `:process` *peuvent être vus comme des actions qui n'ont pas besoin de la volition d'un agent, en d'autres termes il s'agit d'évènements naturels. Ce sont des évènements qui seront déclenchés dès que leurs préconditions seront remplies. La différence entre les* `:event` *et les* `:process` *est que le premier est un évènement instantané alors que le deuxième est un évènement qui va durer un intervalle de temps. Comme nous ne travaillons pas en temps continu, nous n'allons pas nous intéresser à cette partie du requirement* `:time`*. Ces évènements naturels sont utilisés dans notre exemple pour représenter des évènements tels qu'une collision ou l'arrêt involontaire du véhicule.*

```
(:event collision
    :parameters (?X - targets ?P - position ?L - occupants)
    :precondition (and (instance ?X ?P ?L ?D)
                       (on ?X ?P)
                       (not (= ?X vha))
                       (moving)
                       (or (on vha ?P)
                           (through ?P)
                       )
                  )
    :effect (and (not (on ?X ?P))
```

```
                (collide_with ?X)
                (when (alive ?L ?X) (not (alive ?L ?X)))
                (when (alive ?L vha) (not (alive ?L vha)))
            )
)
```

L'évènement naturel `collision` *est déclenché lorsque le véhicule est en mouvement et qu'un élément extérieur au véhicule, défini comme une instance avec certaines propriétés, est dans la même case que le véhicule, ou que celui-ci est passé par cette case. Les effets de cet évènement sont que les personnes impliquées dans la collision qui sont vivantes à ce moment là meurent et que l'élément avec lequel le véhicule a percuté n'est plus considéré comme étant sur cette case. Ce dernier effet est important pour pouvoir simuler des collisions qui pourraient s'enchaîner les unes après les autres dans certains cas particuliers. Cet exemple nous permet de voir l'utilisation des préconditions disjonctives* `(or <condition1> ... <conditionN>)`.

Revenons maintenant à cette question de multiples collisions. Nous voulons être capables d'exprimer le fait que suite à une collision, les conséquences ne seront pas les mêmes selon le contexte. D'un côté, nous voudrions que lorsque le véhicule percute un objet avec une masse importante il soit immobilisé par le choc. D'un autre côté, nous voudrions que lorsque l'élément percuté n'est pas aussi important, le mouvement du véhicule ne soit pas interrompu. L'évènement naturel en question est appelé `stop`*. Pour que celui-ci ait lieu, il faut que le véhicule soit en mouvement et, soit qu'il percute un élément externe considéré comme imposant, soit qu'il ait percuté tous les éléments sur une case. Ce dernier cas est ajouté pour permettre à la simulation de s'arrêter lorsque plus aucun évènement ne peut avoir une influence sur l'évaluation éthique postérieure. Comme le nom de l'évènement naturel l'indique, l'effet de celui-ci est d'interrompre le mouvement du véhicule. Cet évènement illustre la possibilité d'avoir des préconditions disjonctives et conjonctives imbriquées et l'utilisation de préconditions existentiellement quantifiées* `(exists <parameter> <conditions>)`*. Voici sa description :*

```
(:event stop
    :precondition (and (moving)
                       (or (alone)
                           (exists (?X - targets)
                                   (and (collide_with ?X)
                                        (heavy ?X)
                                   )
                           )
                       )
                  )
    :effect (not (moving))
)
```

Pour finir cette présentation d'exemple, intéressons nous à un dernier type d'élément différent de `:action` *et* `:event`*, il s'agit des axiomes correspondant à* `:derived`*. Ces éléments sont une alternative aux* `:event`*, ils se déclenchent également lorsque toutes leurs préconditions sont satisfaites. Il est important d'introduire deux nuances. Premièrement, les évènements naturels peuvent être considérés comme une transition d'un état à un autre dans un système de transition. Comme les actions, les évènements naturels peuvent donc être une arête dans le graphe orienté correspondant au STEE, les axiomes ne le sont pas. Deuxièmement, alors que les effets des actions et des évènements naturels impactent des fluents dits primitifs, les effets des évènements de type* `:derived` *ne peuvent impacter que des fluents dits dérivés [*MUELLER*, 2014], d'où le nom de cette extension : derived predicates. Cette spécificité est en lien étroit avec*

le premier point.
Nous voulons considérer que le véhicule est seul sur sa case uniquement lorsque celui-ci a percuté tous les éléments présents dans celle-ci. Cela est fait grâce au fluent dérivé `alone` *utilisé comme préconditions dans* `stop`*. Ci-dessous la description du* `:derived` *permettant de gérer ce fluent. Sa description est en plus une occasion de montrer l'utilisation de préconditions universellement quantifiées (*`forall <parameter> <conditions>`*).*

```
(:derived (alone)
    (and (not (= ?P main_0))
         (on vha ?P)
         (forall (?X - targets)
                 (or (= ?X vha)
                     (not (on ?X ?P))
                     (collide_with ?X)
                 )
         )
    )
)
```

Le choix de l'ensemble de concepts auquel nous allons nous intéresser dans cette thèse a été orienté par le type de problèmes que nous pouvons vouloir traiter en éthique computationnelle. Les différents *requirements* que nous avons gardés font partie du PDDL [Ghallab et collab., 1998], PDDL 2.1 [Fox et Long, 2003], PDDL 2.2 [Hoffmann et Edelkamp, 2005], et PDDL+ [Fox et Long, 2006]. Ci-dessous la liste de ces éléments avec leur hiérarchie et la version PDDL dans laquelle ils ont été présentés. Parmi ces éléments, ceux qui vont nous intéresser particulièrement sont : les préconditions disjonctives (`:disjunctive-preconditions`), les évènements naturels (`:event`) et les effets conditionnels (`:conditional-effects`).

```
PDDL:      :ucpop
              :adl
                :strips
                :typing
                :equality
                :conditional-effects
                :universal-effects
                :disjunctive-preconditions
                :quantified-preconditions
                   :existential-preconditions
                   :universal-preconditions
              :domain-axioms
PDDL2.1:   :negative-preconditions
PDDL2.2:   :derived-predicates
PDDL+:     :time (only :event)
```

Maintenant que nous avons vu différents concepts utilisés dans la représentation de l'action et du changement grâce au PDDL, étudions d'autres langages ayant vocation à être implémentés.

2.2.2 STRIPS comme un langage de description d'action

STRIPS a été introduit par Fikes et Nilsson [1971]. Nous décrivons ici le fragment STRIPS comme décrit en PDDL. Il s'agit d'un des fragments de base de ce langage, il inclut très peu

des concepts vus dans la section 2.2.1. En effet, ce langage n'inclut que `:strips` parmi la liste donnée précédemment.
Représenter le monde dans STRIPS demande de décrire les actions en leur attribuant des préconditions et des effets. Dans STRIPS un opérateur est un couple $\langle F, L\rangle$ où F est un ensemble de fluents et L est un ensemble de littéraux de fluents cohérent. Décrire une action revient à lui attribuer un opérateur. L'ensemble F de l'opérateur correspond alors aux préconditions et l'ensemble L aux effets.

Définition 2.3 [Description d'action D dans STRIPS]**.** *Soit la signature d'action décrite dans la définition 2.1 et l'opérateur* $\langle F, L\rangle$*, où* $F \subseteq \mathbb{F}$ *et* $L \subseteq Lit_{\mathbb{F}}$*. Une description d'action* D *dans STRIPS est une fonction qui associe à des actions des préconditions et des effets. Formellement :*

$$D : \mathbb{A} \rightarrow \langle F, L\rangle .$$

Le système de transition d'états étiqueté de STRIPS est simple. $\mathbb{S}$ est l'ensemble de toutes les interprétations de $Lit_{\mathbb{F}}$ et τ ne permet que des actions individuelles dans les triplets.

Définition 2.4 [Système de transition d'états étiqueté $\mathscr{S}_{\text{STRIPS}}$]**.** *Soit* D *une description d'action en STRIPS. Le système de transition d'états étiqueté* $\mathscr{S}_{\text{STRIPS}}$ *décrit par* D *est le triplet* $\langle \mathbb{S}, V, \tau\rangle$ *où :*

1. $\mathbb{S} = 2^{Lit_{\mathbb{F}}}$;
2. $V(l, S) = true$ *si* $l \in S$;
3. τ *est l'ensemble de tous les triplets* $(S, \{a\}, S')$ *tels que :*
   (a) S *satisfait les préconditions de* $D(\{a\})$;
   (b) $\begin{cases} l \in S' & \text{si } l \text{ un effet de } D(\{a\}) ; \\ l \in S' \Leftrightarrow l \in S & \text{sinon.} \end{cases}$

L'équivalence dans la condition 3b introduit l'inertie permettant à STRIPS de gérer le problème du décor.

2.2.3 Langage de description d'action $\mathscr{A}$

$\mathscr{A}$ a été introduit par Gelfond et Lifschitz [1993]. C'est le premier langage d'une grande famille de langages conçus spécifiquement pour représenter l'action et le changement. À part la possibilité d'avoir des préconditions négatives qu'offre $\mathscr{A}$, nous pourrions dire que ce langage est le fragment propositionnel de STRIPS. Ce langage inclut donc `:strips` et `:negative-preconditions` parmi la liste de concepts donnée précédemment. Représenter le monde dans $\mathscr{A}$ se fait à l'aide de propositions. Étant donné une action $a \in \mathbb{A}$, un littéral de fluent $l \in Lit_{\mathbb{F}}$ et une conjonction de littéraux de fluents $L \subseteq Lit_{\mathbb{F}}$, une proposition p dans $\mathscr{A}$ est une expression de la forme :

$$a \text{ causes } l \text{ if } L.$$

Comme mentionné précédemment, $\mathscr{A}$ est souvent décrit comme le fragment propositionnel de STRIPS. Toutefois, la représentation du monde par la déclaration de propositions permet à ce langage de gérer les effets conditionnels nativement. Ce langage inclut donc également `:conditional-effects` parmi la liste de concepts donnée précédemment.

Définition 2.5 [Description d'action D dans $\mathscr{A}$]**.** *Soit la signature d'action décrite dans la définition 2.1. Une description d'action* D *dans* $\mathscr{A}$ *est un ensemble de propositions* p.

Le système de transition d'états étiqueté de $\mathscr{A}$ ne diffère de celui de STRIPS que par la condition sur les triplets dans τ. $\mathbb{S}$ reste l'ensemble de toutes les interprétations de $Lit_{\mathbb{F}}$ et τ ne permet que des actions individuelles dans les triplets.

Définition 2.6 [Système de transition d'états étiqueté $\mathscr{S}_{\mathscr{A}}$]**.** *Soit* D *une description d'action en* $\mathscr{A}$. *Le système de transition d'états étiqueté* $\mathscr{S}_{\mathscr{A}}$ *décrit par* D *est le triplet* $\langle \mathbb{S}, V, \tau \rangle$ *où :*

1. $\mathbb{S} = 2^{Lit_{\mathbb{F}}}$;
2. $V(l, S) = true$ *si* $l \in S$;
3. τ *est l'ensemble de tous les triplets* $(S, \{a\}, S')$ *tels que :*

$$S' = \underbrace{\{l \in Lit_{\mathbb{F}} | \exists p \in D, (p = (a \text{ causes } l \text{ if } L)) \wedge (S \vDash L)\}}_{\epsilon} \cup (S \setminus \bar{\epsilon}).$$

Le premier terme de l'union dans la condition 3 demande que tous les effets de a qui le doivent soient bien dans S'. Le deuxième terme introduit l'inertie permettant à $\mathscr{A}$ de gérer le problème du décor. Notez que, bien que la proposition « a causes l if L » tolère des ensembles inconsistants pour L, les conditions sur les τ de la définition 2.6 excluent l'utilisation de tels ensembles dans la construction du STEE.

D'autres expressions sont possibles en $\mathscr{A}$ qui rendent son utilisation plus concise. Toutefois, toutes ces expressions peuvent s'exprimer comme une proposition p. Étant donné un littéral de fluent $l \in Lit_{\mathbb{F}}$, la proposition permettant de décrire la situation initiale :

$$true \text{ causes } l \text{ if } true$$

où $true$ est la notation utilisée pour $\varnothing$, peut être simplifiée dans $\mathscr{A}$ sous la forme :

$$\text{initially } l.$$

Quelques années après l'apparition de $\mathscr{A}$, BARAL et GELFOND [1997] proposent $\mathscr{A}_c$, une extension de $\mathscr{A}$ qui permet de traiter la cooccurrence d'actions. Ces deux langages partagent une sémantique et une syntaxe très similaire. Leur principale différence est la possibilité dans $\mathscr{A}_c$ d'avoir des triplets dans τ de la forme (S, A, S'), où $A \subseteq \mathbb{A}$ est un ensemble d'actions.

2.2.4 Langage de description d'action $\mathscr{B}$

$\mathscr{B}$ peut être vu comme une sous partie du langage proposé par TURNER [1997]. Il s'agit d'une extension de $\mathscr{A}$ qui propose une solution pour gérer les effets indirects des actions et donc le problème de ramification. Pour cela un nouveau type de proposition est utilisé, des lois statiques qui permettent d'inférer des littéraux de fluents. Ce langage inclut donc `:strips`, `:negative-preconditions` et `:conditional-effects` comme $\mathscr{A}$, mais aussi `:domain-axioms` et `:derived-predicates` parmi la liste de concepts donnée précédemment.
Comme dans $\mathscr{A}$, représenter le monde dans $\mathscr{B}$ se fait à l'aide de propositions. Cependant, celles-ci sont appelées des lois dans $\mathscr{B}$. Étant donné une action $a \in \mathbb{A}$, un littéral de fluent

$l \in Lit_{\mathbb{F}}$ et une conjonction de littéraux de fluents $L \subseteq Lit_{\mathbb{F}}$, une loi statique p_s dans $\mathscr{B}$ est une expression de la forme :

$$l \text{ if } L$$

et une loi dynamique p_d dans $\mathscr{B}$ est une expression de la forme :

$$a \text{ causes } l \text{ if } L.$$

Définition 2.7 [Description d'action D dans $\mathscr{B}$]**.** *Soit la signature d'action décrite dans la définition 2.1. Une description d'action* D *dans* $\mathscr{B}$ *est un ensemble de lois statiques* p_s *et dynamiques* p_d*.*

Le système de transition d'états étiqueté de $\mathscr{B}$ est plus complexe que celui de $\mathscr{A}$. La complexité supplémentaire vient du fait de devoir gérer les lois statiques. Elle se retrouve dans les conditions sur $\mathbb{S}$ et sur les triplets dans τ. L'ensemble $\mathbb{S}$ n'est plus simplement l'ensemble de toutes les interprétations de $Lit_{\mathbb{F}}$, mais l'ensemble de toutes les interprétations de $Lit_{\mathbb{F}}$ fermées sous l'ensemble des lois statiques P_s dans D. Soit P_s un ensemble de lois statiques dans D. Un ensemble $L' \subseteq Lit_{\mathbb{F}}$ est fermé sous un ensemble P_s si :

$$L' \supseteq \{l \in Lit_{\mathbb{F}} | \exists p_s \in P_s, (p_s = (l \text{ if } L)) \wedge (L' \vDash L)\}.$$

L'ensemble $Cn_{P_s}(L')$ de conséquences de L' sous P_s est le plus petit ensemble fermé sous P_s tel que $Cn_{P_s}(L') \subseteq L'$. Dans $\mathscr{B}$, τ ne permet pas non plus d'avoir autre chose que des actions individuelles dans les triplets.

Définition 2.8 [Système de transition d'états étiqueté $\mathscr{S}_{\mathscr{B}}$]**.** *Soit* D *une description d'action en* $\mathscr{B}$*. Le système de transition d'états étiqueté* $\mathscr{S}_{\mathscr{B}}$ *décrit par* D *est le triplet* $\langle \mathbb{S}, V, \tau \rangle$ *où :*

1. $\mathbb{S} = \left\{ I \in 2^{Lit_{\mathbb{F}}} | Cn_{P_s}(I) = I \right\}$;
2. $V(l, S) = true$ *si* $l \in S$;
3. τ *est l'ensemble de tous les triplets* $(S, \{a\}, S')$ *tels que :*

$$S' = Cn_{P_s}\left(\{l \in Lit_{\mathbb{F}} | \exists p_d \in D, (p_d = (a \text{ causes } l \text{ if } L)) \wedge (S \vDash L)\} \cup (S \cap S')\right).$$

Le premier terme de l'union dans la condition 3 demande que tous les effets de a qui le doivent soient bien dans S'. Le deuxième terme introduit l'inertie permettant à $\mathscr{B}$ de gérer le problème du décor. L'application de Cn_{P_s} à l'union de ces deux termes s'assure de la prise en compte des effets indirects. Si $P_s = \varnothing$, alors $\mathscr{S}_{\mathscr{B}} = \mathscr{S}_{\mathscr{A}}$. $\mathscr{B}$ peut donc être vu comme une généralisation de $\mathscr{A}$.

2.2.5 Langage de description d'action $\mathscr{C}$

$\mathscr{C}$ a été introduit par GIUNCHIGLIA et LIFSCHITZ [1998]. À différence de $\mathscr{A}$ et $\mathscr{B}$, celui-ci est fondé sur la théorie de l'explication causale proposée par MCCAIN et TURNER [1997] selon laquelle il n'est pas suffisant de trouver qu'un fluent est vrai, il faut pouvoir établir qu'il existe une cause du fait qu'il soit vrai.
Comme dans $\mathscr{B}$, dans $\mathscr{C}$ nous retrouvons aussi des lois statiques et dynamiques. Toutefois, $\mathscr{C}$ est plus expressif que $\mathscr{A}$, $\mathscr{A}_c$ et $\mathscr{B}$ car il permet de représenter des actions non déterministes, la cooccurrence d'actions, des préconditions disjonctives et l'inertie n'est pas

imposée à tous les fluents [LIFSCHITZ, 1997]. Ce langage inclut donc `:strips`, `:negative-preconditions`, `:conditional-effects`, `:domain-axioms` et `:derived-predicates`, mais aussi `:disjunctive-preconditions` parmi la liste de concepts donnée précédemment.
Comme dans $\mathscr{B}$, représenter le monde dans $\mathscr{C}$ se fait à l'aide de propositions appelées des lois. Soit $\mathscr{F}$ les formules propositionnelles de fluents dans $\mathscr{C}$ couramment appelées *formules d'état* et $\mathscr{G}$ les formules propositionnelles de fluents et d'actions. Étant donné deux formules $\psi, \psi' \in \mathscr{F}^2$, une loi statique p_s dans $\mathscr{C}$ est une expression de la forme :

$$\text{caused } \psi' \text{ if } \psi.$$

Puis, étant donné deux formules $\psi, \psi' \in \mathscr{F}^2$ et une formule $\varphi \in \mathscr{G}$, une loi dynamique p_d dans $\mathscr{B}$ est une expression de la forme :

$$\text{caused } \psi' \text{ if } \psi \text{ after } \varphi.$$

Définition 2.9 [Description d'action D dans $\mathscr{C}$]**.** *Soit la signature d'action décrite dans la définition 2.1. Une description d'action* D *dans* $\mathscr{C}$ *est un ensemble de lois statiques* p_s *et dynamiques* p_d*.*

Le système de transition d'états étiqueté de $\mathscr{C}$ est plus complexe que celui de $\mathscr{B}$. La complexité supplémentaire vient de fait de devoir gérer le non déterminisme, la cooccurrence d'actions et les préconditions disjonctives. Elle se retrouve dans la condition sur les triplets dans τ. Comme pour $\mathscr{B}$, il est nécessaire de gérer les lois statiques. L'ensemble $\mathbb{S}$ est donc également l'ensemble de toutes les interprétations de $Lit_{\mathbb{F}}$ fermées sous l'ensemble des lois statiques P_s dans D. Par contre dans $\mathscr{C}$, τ permet d'avoir des ensembles avec plusieurs actions dans les triplets.

Définition 2.10 [Système de transition d'états étiqueté $\mathscr{S}_{\mathscr{C}}$]**.** *Soit* D *une description d'action en* $\mathscr{C}$*. Le système de transition d'états étiqueté* $\mathscr{S}_{\mathscr{C}}$ *décrit par* D *est le triplet* $\langle \mathbb{S}, \mathrm{V}, \tau \rangle$ *où :*

1. $\mathbb{S} = \left\{ \mathrm{I} \in 2^{Lit_{\mathbb{F}}} | \mathrm{C}n_{\mathrm{P}_s}(\mathrm{I}) = \mathrm{I} \right\}$;
2. $\mathrm{V}(l, \mathrm{S}) = true$ *si* $l \in \mathrm{S}$;
3. τ *est l'ensemble de tous les triplets* $(\mathrm{S}, \mathrm{A}, \mathrm{S}')$ *tels que :*

$$\forall p_d \in \mathrm{D}, \left(p_d = \left(caused\ \psi'\ if\ \psi\ after\ \varphi\right)\right) \wedge \left(\mathrm{S} \cup \mathrm{A} \vDash \varphi\right) \wedge \left(\mathrm{S}' \vDash \psi\right) \Longrightarrow \mathrm{S}' \vDash \psi'.$$

D'autres expressions sont possibles en $\mathscr{C}$ qui rendent son utilisation plus concise. Toutefois, toutes ces expressions peuvent s'exprimer comme des lois statiques ou dynamiques. Le tableau 2.1 résume les simplifications possibles dans la syntaxe de $\mathscr{C}$.

Comme mentionné au début de cette section, $\mathscr{C}$ a été développé comme intégrant une notion causale. Toutefois, cette notion de causalité n'est que la brique interne au langage, elle ne contient pas le raisonnement plus complexe que nous traitons en détail dans cette thèse. Il est intéressant de souligner que la nécessité d'un tel raisonnement apparaissait déjà dans les réflexions des créateurs de ces langages : « Two ontological ideas that I do not know how to express in $\mathscr{C}$ are continuous time, and causal relations between actions » [LIFSCHITZ, 1997].

Hypothèses	Loi statique ou dynamique	Simplification
$\psi \in \mathscr{F}$	caused $false$ if $\neg\psi$	always ψ
$\psi, \psi' \in \mathscr{F}^2$	caused ψ' if $\psi' \wedge \psi$	default ψ' if ψ
$\psi' \in \mathscr{F}, \varphi \in \mathscr{G}$	caused ψ' if $true$ after φ	caused ψ' after φ
$\psi' \in \mathscr{F}$	caused ψ' if ψ' after ψ'	inertial ψ'
$\psi \in \mathscr{F}, \varphi \subseteq \mathbb{A}$	caused $false$ after $\psi \wedge \varphi$	non executable φ if ψ
$\psi, \psi' \in \mathscr{F}^2, \varphi \subseteq \mathbb{A}$	caused ψ' if ψ' after $\psi \wedge \varphi$	φ may cause ψ' if ψ

TABLEAU 2.1 – Simplifications possibles dans la syntaxe de $\mathscr{C}$.

2.2.6 Calcul des Situations

Introduit par MCCARTHY et HAYES [1969], le Calcul des Situations est le premier langage dédié au raisonnement sur les actions. Il s'agit d'un langage typé de la logique du premier ordre. Les types présents sont les fluents, les situations, les actions et les objets. Plusieurs versions ont été proposées au fil des années, notamment pour faire face au problème du décor. Nous présentons ici la version de REITER [1991, 2001]. Ce langage inclut tous les concepts parmi la liste de donnée précédemment.
Jusqu'à présent nous avons utilisé des variables qui semblent propositionnelles. Cela n'a pas été précisé mais celles-ci peuvent être instantiées. Le Calcul des Situations étant ancré par tradition dans la logique du premier ordre, nous utiliserons dans cette section une notation légèrement différente. Les actions comme les fluents seront des fonctions plutôt que des constantes. Pour représenter les arguments de ces fonctions, nous utiliserons $\vec{x}$ qui représente une séquence de termes. Ainsi, une action sera de la forme $\mathrm{A}(\vec{x})$ et un fluent $f(\vec{x})$.

Représenter le monde en Calcul des Situations revient à décrire une structure arborescente. La racine est une situation initiale σ_0 et les différents nœuds correspondent à d'autres situations σ. Le passage d'un nœud à un autre, donc d'une situation à une autre est fait grâce aux actions $\mathrm{A}(\vec{x})$. Le prédicat $\mathrm{P}oss(\mathrm{A}(\vec{x}), \sigma)$ indique que l'action $\mathrm{A}(\vec{x})$ est possible dans la situation σ, la fonction $do(a, \sigma)$ donne la situation qui résulte de la réalisation de a dans σ et, étant donné une séquence de termes $\vec{x}$, le prédicat $f(\vec{x}, \sigma)$ indique que le fluent $f(\vec{x})$ est vrai dans la situation σ.

Pour définir une description d'action D en Calcul des Situations, il est nécessaire d'introduire les axiomes de préconditions et les axiomes d'états successeurs. Une formule est dite uniforme en σ, si elle ne contient pas de prédicat $\mathrm{P}oss$, de relation d'ordre partielle entre situations $\sqsubseteq$, d'autres situations que σ ou de quantification sur σ. Étant donné $\Pi(\vec{x}, \sigma)$ une formule uniforme en σ, un axiome de préconditions est une formule de la forme :

$$\forall \vec{x}, \sigma, (\mathrm{P}oss(\mathrm{A}(\vec{x}), \sigma) \Leftrightarrow \Pi(\vec{x}, \sigma)).$$

Étant donné $\Psi^+(\vec{x}, a, \sigma)$ une formule uniforme en σ indiquant que la réalisation de a dans σ rend le fluent f vrai et $\Psi^-(\vec{x}, a, \sigma)$ une formule uniforme en σ indiquant que la réalisation de a dans σ rend le fluent f faux, un axiome d'état successeur (« successor state axiom » en anglais) est une formule de la forme :

$$\forall \vec{x}, \sigma, \big(f(\vec{x}, do(a, \sigma)) \Leftrightarrow \Psi^+(\vec{x}, a, \sigma) \vee \big(f(\vec{x}, \sigma) \wedge \neg\Psi^-(\vec{x}, a, \sigma)\big)\big).$$

Définition 2.11 [Description d'action D en Calcul des Situations]**.** *Soit la signature d'action décrite dans la définition 2.1. Une description d'action* D *en Calcul des Situations est un ensemble d'axiomes de préconditions, d'axiomes d'états successeurs et une situation initiale* σ_0.

Une situation σ est dite exécutable si chaque action réalisée pour arriver à cette situation était possible dans la situation où elle a été réalisée. Formellement :

$$executable(\sigma) \stackrel{\text{def}}{=} \forall a, \sigma', \left(do(a, \sigma') \sqsubseteq \sigma \implies Poss(a, \sigma')\right).$$

Définition 2.12 [Système de transition d'états étiqueté $\mathscr{S}_{\text{SC}}$]**.** *Soit* D *une description d'action en Calcul des Situations. Le système de transition d'états étiqueté* $\mathscr{S}_{\text{SC}}$ *décrit par* D *est le triplet* $\langle \mathbb{S}, \text{V}, \tau \rangle$ *où :*

1. $\mathbb{S}$ *est l'ensemble de toutes les situations ;*
2. $\text{V}(f, \sigma) = true$ *si* $\text{D} \models f(\sigma)$ *ou* $\text{V}(f, \sigma) = false$ *si* $\text{D} \not\models f(\sigma)$ *;*
3. τ *est l'ensemble de tous les triplets* $(\sigma, a, do(a, \sigma))$ *tel que :*

$$\text{D} \models executable(do(a, \sigma)).$$

Utilisant la logique du premier ordre, le Calcul des Situations a l'avantage d'être très expressif, il inclut tous les concepts parmi la liste donnée précédemment pour le PDDL. Toutefois, cette expressivité se paye par la difficulté à être utilisé et implémenté. C'est la raison pour laquelle des approches se basant sur la logique propositionnelle, comme celles présentées précédemment, ont été développées. Comme l'explique la citation de début de chapitre, le défi des créateurs du Calcul des Situations a été de trouver les bons axiomes, alors que celui de créateurs de langages comme $\mathscr{A}$, $\mathscr{B}$ et $\mathscr{C}$ a été de développer la sémantique appropriée [Baral et Gelfond, 1997].

2.2.7 Calcul des Évènements

Le Calcul des Évènements est un autre langage typé de la logique du premier ordre. Il a été conçu pour décrire les effets des occurrences d'actions dans une narration donnée, i.e. des actions qui se produisent à des temps donnés et des fluents vrais à certains temps. Une différence est donc faite entre une action a et ses occurrences en un ou plusieurs instants. Ce langage montre que pour raisonner sur des narratives il est pertinent d'utiliser une ontologie où le déroulé temporel est indépendant des occurrences d'actions qui s'y produisent. Le temps passe d'être implicitement représenté par les différents états, comme en Calcul des Situations, à être explicitement représenté. L'ontologie correspondante est donc différente à celle de tous les langages vus jusqu'ici ; en plus des actions et des fluents nous avons des points temporels.

Définition 2.13 [*signature d'action*]**.** *Une signature d'action en Calcul des Évènements est un quadruplet* $\langle \mathbb{F}, \mathbb{A}, \mathbb{T}, \leq \rangle$ *où* $\mathbb{F}$ *est un ensemble de fluents,* $\mathbb{A}$ *est un ensemble d'actions,* $\mathbb{T}$ *est un ensemble de points temporels et* $\leq$ *est un ordre partiel sur l'ensemble* $\mathbb{T}$*.*

Différentes versions ont été présentées au fil des années. Le « Original Event Calculus » a été présenté par Kowalski et Sergot [1986]. Puis, une version simplifiée, le « Simplified Event Calculus », a ensuite été proposée par Kowalski [1992]. Cette dernière a été étendue par Shanahan [1997] en permettant que l'inertie ne soit pas imposée à tous les fluents et en permettant de représenter les changements continus. Cette version est connue sous le nom de « Basic Event Calculus ». Ensuite, Miller et Shanahan [2002] ont proposé de

nombreuses formulations possibles du « Basic Event Calculus » que MUELLER [2004] a combiné pour créer le Calcul des Évènements. Cette version est la plus expressive de toutes. Par rapport au Calcul des Situations, le Calcul des Évènements traite plus facilement la cooccurrence d'évènements, le temps continu, les changements continus, les actions ayant une durée, les effets non déterministes, les évènements partiellement ordonnés et les évènements naturels [GELFOND et collab., 1991; MUELLER, 2008]. Dans le but d'avoir une version plus facilement implémentable, MUELLER [2014] propose une version discrète où le temps est représenté par des entiers ce qui lui permet de simplifier l'axiomatisation ; cela donne le « Discrete Event Calculus ». Pour une histoire détaillée de ce processus avec les axiomes de chaque version, voir [MUELLER, 2008].
Mais l'histoire ne s'arrête pas là, KAKAS et MILLER [1997] proposent le langage de description d'action $\mathscr{E}$. Il s'agit d'une version du Calcul des Évènements qui au lieu d'être représentée en logique du premier ordre, utilise la sémantique des langages de description d'action comme $\mathscr{A}$, $\mathscr{B}$ ou $\mathscr{C}$. MILLER et SHANAHAN [2002] ont prouvé l'équivalence entre $\mathscr{E}$ et le Calcul des Évènements auquel certains prédicats sont retirés. $\mathscr{E}$ est décrit par KAKAS et MILLER [1997] de façon à ce qu'aussi bien une représentation discrète que continue du temps puisse être utilisée. Nous décrivons ici la version discrète. Ce langage inclut tous les concepts parmi la liste de donnée précédemment.

Représenter le monde dans $\mathscr{E}$ se fait à l'aide de propositions comme dans $\mathscr{A}$. Étant donné une action $a \in \mathbb{A}$, un fluent $f \in \mathbb{F}$ et une conjonction de littéraux de fluents $\mathrm{L} \subseteq Lit_{\mathbb{F}}$, une c-proposition p_c dans $\mathscr{E}$ est une expression soit de la forme :

$$a \text{ initiates } f \text{ when L},$$

soit de la forme :

$$a \text{ terminates } f \text{ when L}.$$

Définition 2.14 [Description d'action D dans $\mathscr{E}$]**.** *Soit la signature d'action décrite dans la définition 2.13. Une description d'action* D *dans* $\mathscr{E}$ *est un ensemble de c-propositions.*

Le système de transition d'états étiqueté de $\mathscr{E}$ ne diffère de celui de $\mathscr{A}$ que par la condition sur les triplets dans τ. $\mathbb{S}$ reste l'ensemble de toutes les interprétations de $Lit_{\mathbb{F}}$ et τ ne permet que des actions individuelles dans les triplets.

Définition 2.15 [Système de transition d'états étiqueté $\mathscr{S}_{\mathscr{E}}$]**.** *Soit* D *une description d'action en* $\mathscr{E}$*. Le système de transition d'états étiqueté* $\mathscr{S}_{\mathscr{E}}$ *décrit par* D *est le triplet* $\langle \mathbb{S}, \mathrm{V}, \tau \rangle$ *où :*

1. $\mathbb{S} = 2^{Lit_{\mathbb{F}}}$;
2. $\mathrm{V}(l, \mathrm{S}) = true$ *si* $l \in \mathrm{S}$;
3. τ *est l'ensemble de tous les triplets* $(\mathrm{S}, \{a\}, \mathrm{S}')$ *tels que :*

$$\mathrm{S}' = \overbrace{\left\{f \in \mathbb{F} | \exists p_c \in \mathrm{D}, \left(p_c = \left(a \textit{ initiates } f \textit{ when } \mathrm{L}\right)\right) \wedge (\mathrm{S} \vDash \mathrm{L})\right\}}^{\epsilon_+} \cup$$
$$\underbrace{\left\{\neg f \in \mathbb{F} | \exists p_c \in \mathrm{D}, \left(p_c = \left(a \textit{ terminates } f \textit{ when } \mathrm{L}\right)\right) \wedge (\mathrm{S} \vDash \mathrm{L})\right\}}_{\epsilon_-} \cup \left(\mathrm{S} \setminus \left(\overline{\epsilon_+} \cup \overline{\epsilon_-}\right)\right).$$

Le premier terme de l'union ϵ_+ dans la condition 3 demande que tous les effets positifs soient bien dans S'. Le deuxième terme de l'union ϵ_- dans la condition 3 demande que tous les effets négatifs soient bien dans S'. Le troisième terme demande que la négation de tous

les effets positifs comme négatifs soient retirés de S' et ajoute l'inertie permettant à $\mathscr{E}$ de gérer le problème du décor.

Du fait que dans le Calcul des Évènements décrit les effets des occurrences d'actions dans une narration donnée, il est nécessaire d'introduire des éléments supplémentaires qui prennent en compte le temps. Étant donné une une action $a \in \mathbb{A}$, un littéral de fluent $l \in Lit_{\mathbb{F}}$ et un point temporel $t \in \mathbb{T}$. Une h-proposition p_h dans $\mathscr{E}$ est une expression de la forme :

$$a \text{ happens-at } t.$$

Une t-proposition p_t dans $\mathscr{E}$ est une expression de la forme :

$$l \text{ holds-at } t.$$

Définition 2.16 [Domaine de langage $\mathscr{E}$]**.** *Soit la signature d'action décrite dans la définition 2.13. Un domaine de langage* D_l *dans* $\mathscr{E}$ *est un ensemble de h-propositions et de t-propositions.*

Une interprétation de $\mathscr{E}$ est une cartographie :

$$H : \mathbb{F} \times \mathbb{T} \rightarrow \{false, true\}.$$

Définition 2.17 [Modèle]**.** *Une interprétation* H *de* $\mathscr{E}$ *est un modèle de* $D \cup D_l$ *si* $\forall t \in \mathbb{T}$, $\exists S \in \mathbb{S}$ *tel que :*

1. $\forall f \in \mathbb{F}, \big(H(f,t) = true \implies f \in S\big) \wedge \big(H(f,t) = false \implies \neg f \in S\big)$;
2. $\exists (a, S') \in \mathbb{A} \times \mathbb{S}$ *tel que :*
   (a) $(a \textit{ happens-at } t) \in D_l$;
   (b) $(S, \{a\}, S') \in \tau$;
   (c) $\forall f \in \mathbb{F}, \big(H(f,t+1) = true \implies (f \in S' \vee (f \textit{ holds-at } t+1) \in D_l)\big) \wedge$
   $\big(H(f,t+1) = false \implies (\neg f \in S' \vee (\neg f \textit{ holds-at } t+1) \in D_l\big)$.

2.3 Conclusion

Dans ce chapitre nous avons donné un aperçu de ce qu'est la modélisation et la représentation de l'action et du changement. Nous avons commencé par décrire un des modèles les plus courants, le système de transition d'états. Dans ce modèle, l'évolution du monde peut être vue comme une suite d'états qui s'enchaînent les uns après les autres au fur et à mesure que des actions se produisent.
Puis, nous avons présenté différents langages qui permettent de représenter un tel modèle en adoptant une structure permettant de présenter de façon concise la syntaxe et la sémantique d'un langage. Nous avons commencé par présenter le PDDL qui nous a permis d'avoir un aperçu global. Puis, nous avons présenté STRIPS, $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$, le Calcul des Situations et le Calcul des Évènements.
Nous avons vu que le Calcul des Situations a l'avantage d'être très expressif mais difficile à être utilisé et implémenté. $\mathscr{A}$, $\mathscr{A}_c$, $\mathscr{B}$, $\mathscr{E}$ et l'ADL [PEDNAULT, 1989], fragment déterministe du PDDL, sont plus facilement utilisables et implémentables, mais aucun ne permet à la fois la cooccurrence d'évènements et les préconditions disjonctives. Comme nous le découvrirons dans le chapitre 3 et 5, ces éléments sont fondamentaux si nous voulons pouvoir traiter

la causalité dans toute sa complexité. Pour avoir les deux il faut soit s'intéresser au PDDL+ ou à $\mathscr{C}$. Toutefois, ces versions très expressives permettent de représenter les actions non déterministes et duratives ce qui amène d'autres difficultés. Nous verrons dans le chapitre 6 un langage permettant de trouver un point intermédiaire idéal aussi bien pour modéliser et représenter le raisonnement éthique que le raisonnement causal.

Chapitre 3

État de l'art : causalité effective

> « *[First formulation :] we may define a cause to be an object followed by another, and where all objects, similar to the first, are followed by objects similar to the second. [Second formulation :] Or, in other words, where if the first object had not been, the second had never existed.* »
>
> HUME [1748]

Sommaire

Ce chapitre se veut un aperçu de ce qu'est la causalité. Cette notion joue un rôle essentiel dans notre compréhension du monde, ce qui explique sûrement qu'il soit possible de la retrouver dans la pensée de nombreux philosophes tout au long de l'histoire de la philosophie, depuis les philosophes de l'antiquité grecque jusqu'à nos jours [RUSSELL, 2011]. Mais cette notion n'intéresse pas uniquement les philosophes, elle fait l'objet d'études en psychologie, droit, mathématiques et informatique entre autres. Lorsque dans cette thèse nous parlerons du *domaine de la causalité*, il s'agira de l'ensemble de la communauté pluridisciplinaire s'intéressant à la causalité.

Une relation causale est une relation binaire qui lie une cause à une conséquence. Une telle relation nous permet de trouver une explication rationnelle aux phénomènes que nous observons. Deux types de causalité peuvent être distingués : la causalité générale et la causalité effective. Le deuxième est celui qui nous intéresse le plus et pour lequel cette thèse propose des contributions.
La *causalité générale* (« type causality » en anglais) peut être vue comme la découverte de lois régissant le monde dans lequel nous vivons. C'est en quelque sorte ce que la science cherche à déterminer. Par exemple, la seconde loi de Newton permet d'établir la relation qu'il existe entre l'accélération d'un objet et sa masse. Les travaux de chercheurs dans diverses disciplines ont permis d'établir l'impact de l'action humaine sur le dérèglement climatique. D'autres travaux ont montré que conduire en état d'ébriété augmentait le risque d'avoir un accident. Une fois établies, ces lois nous permettent non seulement de mieux comprendre le monde, mais aussi de prendre des décisions en ayant une meilleur idée des conséquences que nos actions peuvent avoir. La connaissance de ces lois ouvrent la porte à un raisonnement a priori.
La *causalité effective* (« actual causality » en anglais) peut être vue comme la détermination des facteurs qui se sont effectivement produits dans une situation donnée et qui ont produit une conséquence. Autrement dit, il s'agit de déterminer quelle partie d'une loi causale a été appliquée dans un cas précis. Par exemple, le fait qu'une pierre ait percuté la bouteille posée sur la table peut être reliée à la façon dont Suzy l'a lancée, utilisant en partie la seconde loi de Newton. Il se trouve que la bouteille contenait un produit chimique et que l'impact de la pierre fait que ce produit se déverse dans un lac. L'impact sur l'écosystème dans la semaine qui a suivi peut alors être déterminé en utilisant les travaux déterminant l'effet de ce produit chimique dans les organismes d'eau douce. Lorsque la police a interrogé Suzy au sujet de son action, celle-ci a rétorqué être affectée par l'accident de voiture qu'avait eu sa sœur la veille. Rentrant d'une soirée en état d'ébriété, les sens de cette dernière n'ont pas été au rendez-vous et elle ne s'est pas arrêtée à temps devant la porte du garage de chez elle. Pouvoir établir une relation causale dans des situations précises nous permet par exemple d'évaluer la responsabilité juridique d'un individu par rapport à un préjudice, ou le statut déontique d'une action dans certaines théories morales. Cette capacité peut être assimilée à un raisonnement a posteriori.

Dans le cadre concret de cette thèse, la causalité générale est essentielle dans la modélisation et la représentation de l'action et du changement. En effet, d'une façon ou d'une autre, tous les langages présentés dans le chapitre 2 demandent de représenter les préconditions ou les effets d'une action ou d'un ensemble d'actions. Cette information n'est rien d'autre que la représentation d'une loi causale. Nous sommes inévitablement des utilisateurs de la causalité générale, mais nous n'y contribuons pas.
La causalité effective est un concept essentiel pour cette thèse dans le domaine de l'éthique

computationnelle qui cherche à formaliser notre capacité en tant qu'êtres rationnels à évaluer moralement une action. En effet, la plupart des théories morales que nous avons vu dans le chapitre 1 s'appuient d'une façon plus ou moins explicite sur cette notion. Les théories conséquentialistes comme celles de la section 1.3.1 sont celles où le lien est le plus fort et le plus évident. Viennent ensuite les théories qui demandent un raisonnement sur les moyens et les fins comme celle de la section 1.3.2. Finalement, moins évident mais ayant un lien quand même, les théories reposant sur un code de conduite peuvent très bien inclure des règles morales touchant aux conséquences des actions. Cela concerne donc les théories comme celles de la section 1.2.1. S'intéresser à l'éthique computationnelle dans sa globalité passe donc nécessairement par s'intéresser à la causalité effective. Comme pour la causalité générale, nous sommes inévitablement des utilisateurs de la causalité effective. Toutefois, pour surprenant que cela puisse paraître, malgré l'omniprésence de cette notion et le nombre de travaux pluridisciplinaires s'étant penchés sur la question, il n'existe pas de consensus sur une définition. Dans cette thèse nous essayons d'y contribuer.

Ce chapitre est divisé en deux sections. La section 3.1 présente dans les grandes lignes l'histoire de la causalité contemporaine. Une attention particulière sera consacrée à deux points importants : les problèmes de surdétermination sur lesquels reposent la plupart des débats encore existants dans le domaine; l'importance de séparer la notion de causalité effective et de responsabilité. La section 3.2 expose les défis qui restent à traiter dans le domaine, notamment si nous voulons pouvoir utiliser les résultats du domaine pour l'éthique computationnelle.

3.1 La causalité effective vue par la philosophie, le droit et l'informatique

Dans cette section nous donnons un aperçu global de ce qu'est la causalité effective et des problématiques successives qui ont abouti au domaine comme nous le connaissons aujourd'hui. S'agissant d'un domaine particulièrement pluridisciplinaire, il ne ferait pas de sens que le contenu de cet aperçu ne le soit pas. C'est d'autant plus le cas que dans la suite de cette thèse une partie des choix réalisés réponde spécifiquement à des questions philosophiques. Cet aperçu est donc construit à l'aide de travaux en philosophie, en droit et en informatique.
Dans ce chapitre, lorsque nous parlons des travaux sur la causalité effective en informatique, nous les plaçons dans le contexte du domaine pluridisciplinaire de la causalité. Nous avons fait le choix de ne pas rentrer dans les détails formels de chaque approche, car ce n'est pas par rapport à ces détails que nous pensons que le choix de l'approche à adopter en éthique computationnelle doit être fait. En tout cas, pas au stade actuel du domaine de la causalité. Qui plus est, KUEFFNER [2021] propose une comparaison détaillée et très complète des différents travaux en informatique sur la causalité effective. Nous invitons les lecteurs qui voudraient approfondir leur connaissance sur ce sous-ensemble du domaine à s'y référer.

Nous utilisons un exemple qui est déroulé tout au long du chapitre et qui est complété par d'autres si besoin. Nous avons choisi de prendre un exemple qui traite de pollution. Il peut être utilisé pour illustrer les deux grands types de problèmes de surdétermination.

Exemple 3.1 [pollution]**.** *Un village situé le long d'une rivière abrite n familles. L'eau potable utilisée par les habitants du village provient d'une usine de potabilisation qui puise l'eau dans la rivière, elle-même provenant d'un lac situé en amont. Cependant, la capacité de cette usine est limitée, elle ne peut traiter l'eau que si elle a un indice de pollution strictement inférieur à un seuil. Lorsque l'eau de la montagne atteint le lac, l'indice de pollution est nul. Il existe deux sources potentielles de pollution du lac : (i) les eaux usées industrielles d'une usine qui produit des enceintes connectées pour un célèbre site d'achat en ligne et qui donne du travail à au moins un membre de chaque famille du village; (ii) les eaux usées industrielles d'une usine qui produit des médicaments vitaux pour k patients et qui a complètement automatisé sa ligne de production. En temps normal, les eaux usées de cette usine de médicaments ne polluent pas le lac car elles sont traitées rigoureusement par une station d'épuration avant d'y être déversées. Nous supposons que les déversements sont nécessaires au fonctionnement des deux usines et que la gestion de leur production est assurée par un agent.*

Exemple 3.2 [cas simple de pollution]**.** *Nous considérons le scénario dans lequel l'agent responsable de la gestion lance uniquement la production de l'usine de médicaments. Il s'avère que la station d'épuration des eaux usées de l'usine de médicaments est en panne depuis le début du mois. Les eaux usées déversées augmentent l'indice de pollution jusqu'au seuil. Dans ce scénario, les habitants du village se retrouvent sans eau potable.*

Exemple 3.3 [cas préemptif de pollution]**.** *Nous considérons le scénario dans lequel l'agent responsable de la gestion lance la production de l'usine d'enceintes connectées. Les eaux usées déversées augmentent l'indice de pollution jusqu'au seuil. Il s'avère que la station d'épuration des eaux usées de l'usine de médicaments est en panne depuis le début du mois. Quelques heures après le lancement de l'usine d'enceintes connectées, l'agent lance la production de l'usine de médicaments. Toutefois, les eaux usées de cette production ne peuvent pas être déversées car suite à l'augmentation du seuil de pollution du lac les autorités bloquent tout rejet dans le lac. Dans ce scénario, les habitants du village se retrouvent sans eau potable.*

L'exemple 3.3 est un cas de surdétermination préemptive. Du fait que la station d'épuration des eaux usées de l'usine de médicaments est en panne, la production d'enceintes et la production de médicaments sont individuellement suffisantes pour produire le préjudice aux habitants. En effet, chacun des déversements est suffisant pour élever le niveau de pollution de la rivière au seuil requis pour causer le préjudice. Toutefois, le fait que ce soient les déversements de l'usine de médicaments qui produisent le préjudice des habitants est préempté par l'antériorité des déversements de l'usine d'enceintes.

Exemple 3.4 [cas duplicatif de pollution]**.** *Nous restons dans le même contexte de l'exemple 3.3, sauf que dans ce scénario, l'agent lance la production des deux usines simultanément. Les eaux usées déversées augmentent l'indice de pollution jusqu'à deux fois le seuil. Les habitants du village situé le long de la rivière sont également privés d'eau potable dans ce scénario.*

L'exemple 3.4 est un cas de surdétermination duplicative. Nous somme dans la même situation que dans le cas de préemption, sauf que les déversements des deux usines se font simultanément. Leurs effets se combinent pour produire le préjudice conjointement.

Cette section est présentée en deux parties. La section 3.1.1 donne un aperçu global de ce qu'est la causalité effective et la section 3.1.2 présente deux problématiques importantes dans le domaine : la surdétermination et la distinction entre causalité et responsabilité.

3.1.1 Histoire du domaine de la causalité

Dans cette section nous présentons dans les grandes lignes l'histoire de la causalité contemporaine depuis HUME [1748] jusqu'à nos jours. Dans la section 3.1.1.1 nous présentons la conception de Hume de la causalité. Sur celle-ci reposent les deux types d'approches principales qui existent aujourd'hui. Cette section s'intéresse au développement des approches par régularité, dominantes depuis Hume jusqu'à la deuxième moitié du vingtième siècle. Dans la section 3.1.1.2 nous présentons le schisme créé par LEWIS [1973] en introduisant les approches contrefactuelles. Nous discutons ensuite de la vision de Halpern, approche dominante aujourd'hui et héritière de la conception de Lewis de la causalité. Dans la section 3.1.1.3 nous présentons comment les approches par régularité ont évolué pour devenir des approches par inférence, répondant ainsi aux critiques qui leur étaient adressées et se plaçant en concurrence avec les approches contrefactuelles.

3.1.1.1 De Hume à Mackie, les approches par régularité

La conception de la causalité introduite par Hume est à l'origine des deux types d'approches principales qui existent aujourd'hui. Nous allons dans cette section nous intéresser aux approches par régularité [ANDREAS et GUENTHER, 2021]. Selon cette vision de la causalité indissociable de l'empirisme de l'époque, une cause est régulièrement suivie par ses effets. Cette approche est définie par HUME [1748] dans une première formulation de ce qu'est la causalité pour lui.

> [First formulation :] an object precedent and contiguous to another, and where all the objects resembling the former are plac'd in like relations of priority and contiguous to those objects, that ressemble the latter.

Notez l'utilisation du terme « objet » pour faire références aux causes et aux conséquences. Celui-ci est intentionnellement large car il englobe des éléments de différente nature. Sa vision peut être résumée en trois points. Un objet c est une cause de l'objet e ssi :

1. c est contiguë dans l'espace et dans le temps à e ;
2. c précède e temporellement ;
3. tous les objets du même type que c sont suivis d'objets du même type que e.

Nous retrouvons dans cette dernière condition l'idée de régularité. À cela il est important d'ajouter que la conception de la causalité de Hume et de la plupart de ses successeurs adoptent le principe de raison suffisante de LEIBNIZ [1710] selon lequel : « jamais rien n'arrive sans qu'il y ait une cause ou du moins une raison déterminante, c'est-à-dire qui puisse servir à rendre raison a priori pourquoi cela est existant plutôt que non existant et pourquoi cela est ainsi plutôt que de toute autre façon ». L'adhésion à ce principe a de nombreuses implications. Celle qui va nous intéresser lorsque nous proposerons une approche en causalité effective pour l'éthique computationnelle est que, comme tout a une cause, il est théoriquement possible de remonter la chaîne causale à partir d'un objet jusqu'à la création de l'univers.

Exemple 3.2 [suite]**.** *Pour illustrer les propos de cette section il est utile d'adopter la notation logique utilisée couramment par les approches par régularité. Nous considérons les prédicats suivants : le préjudice des habitants du village p, le seuil de pollution de la rivière étant*

atteint s, le déversement des eaux industrielles provenant de l'usine d'enceintes o_e et de médicaments o_m, puis le fait que la station d'épuration des eaux usées de l'usine de médicaments soit hors service $\neg se$. Nous avons les formules suivantes : $s \leftrightarrow p$ et $o_e \vee (o_m \wedge \neg se) \leftrightarrow s$.
Nous pouvons imaginer que pour Hume, le seuil de pollution de la rivière étant atteint s est une cause du préjudice des habitants p si : il y a une proximité physique et dans le temps entre la rivière polluée et l'usine de potabilisation qui puise l'eau dans la rivière ne pouvant plus la traiter ; le fait que la rivière soit polluée précède bien temporellement la faille de l'usine de potabilisation ; et à chaque fois qu'un seuil similaire de pollution est atteint, il est possible d'observer une défaillance similaire dans les usines de potabilisation.

Cette définition proposée par Hume pose de nombreux problèmes, mais quatre apparaissent comme plus importants. Le premier est l'utilisation du terme « objets du même type » qui n'est pas clairement défini et qui semble pourtant jouer un rôle important dans cette définition.
Le deuxième est que cette définition semble fonctionner pour des causes individuelles (prédicats) mais pas pour des situations où plusieurs objets doivent être réunis (conjonctions) ou pour des situations où plusieurs objets sont suffisants individuellement à produire la conséquence (disjonctions).
Le troisième problème est l'existence d'objets exceptionnels qui peuvent causer d'autres objets, mais leur nature exceptionnelle semble ne pas rentrer dans l'idée de régularité. ANDREAS et GUENTHER [2021] donne l'exemple de l'extinction des dinosaures. En supposant que ce soit bien une météorite qui soit la cause de leur extinction, est-il possible d'établir cette relation causale par régularité sachant qu'il s'agit d'un évènement exceptionnel pouvant difficilement se reproduire ?
Le quatrième problème est qu'il existe des objets qu'il serait possible de relier causalement selon cette définition, mais qui ne le sont pas réellement. ANDREAS et GUENTHER [2021] donnent l'exemple classique du coq qui chante tous les jours juste avant le lever du soleil. Nous ne voulons pourtant pas dire que le chant du coq cause le lever du soleil. Si deux objets sont des conséquences d'une cause commune, ils peuvent satisfaire les conditions de Hume sans être reliés causalement.

Des avancées significatives pour ces approches sont faites par MILL [1843]. Il résout partiellement le deuxième problème mentionné ci-dessus en proposant une définition permettant de gérer les situations où plusieurs objets doivent être réunis (conjonctions). Mill s'appuie sur l'idée qu'une régularité témoigne de la présence de lois naturelles et qu'une relation causale est le résultat de l'application de ces lois. Ayant cette vision d'ensemble, il ajoute à la version de Hume l'idée que lorsqu'il est question de causes, il n'est pas suffisant d'identifier les objets qui par leur présence ont produit la conséquence ; il faut également prendre en compte les objets devant être absents. Cette vision de cause « suffisante » le pousse à considérer les ensembles comme la cause. Pour Mill une cause est l'ensemble d'objets positifs et négatifs suffisants et nécessaires. La définition qu'il propose est la suivante : soit l'ensemble d'objets positifs C et l'ensemble d'objets négatifs D. L'union de ces ensembles est la cause de *e* ssi :

1. les objets dans C sont contiguës dans l'espace et dans le temps à *e* et il n'y a pas d'objets dans D qui le soient ;
2. les objets dans C précèdent *e* temporellement ;
3. il existe une loi de la nature pour laquelle la présence de tous les objets de C est suivie

par la présence d'objets du même type que e, pour autant qu'aucun objet dans D ne soit présent.

Exemple 3.2 [suite]**.** *Nous pouvons imaginer que pour Mill, le déversement des eaux industrielles provenant de l'usine de médicaments o_m, puis le fait que la station d'épuration des eaux usées de l'usine de médicaments soit hors service $\neg se$ est une cause du seuil de pollution de la rivière étant atteint s si : il y a une proximité physique et dans le temps entre le déversement des eaux industrielles provenant de l'usine de médicaments* $C = \{o_m\}$ *et le fait que la rivière soit polluée ; il n'y a pas de proximité physique ou temporelle entre une station d'épuration pour les eaux usées de l'usine de médicaments et le fait que la rivière soit polluée* $D = \{\neg se\}$ *; le déversement des eaux industrielles provenant de l'usine de médicaments précède bien temporellement le fait que la rivière soit polluée ; il existe une loi de la nature de la forme $o_m \wedge \neg se \leftrightarrow s$.*

Même si plus aboutie, cette définition proposée par Mill ne résout pas tous les problèmes. Premièrement, le problème lié à l'utilisation du terme « objets du même type » qui n'est pas clairement défini n'a pas été résolu.
Cette définition adopte une vision plus large qui lui permet de gérer les situations où plusieurs objets doivent être réunis (conjonctions). Toutefois, cette définition ne couvre pas les situations où plusieurs ensembles d'objets sont suffisants individuellement à produire la conséquence (disjonctions). En effet, pour Mill une cause est l'ensemble d'objets positifs et négatifs suffisants et nécessaires, double condition qui ne peut être satisfaite lorsque plusieurs ensembles d'objets sont suffisants car aucun n'est vraiment nécessaire.
Le problème de l'existence d'objets exceptionnels est résolu en remplaçant l'idée de régularité dans la troisième condition de Hume par l'existence d'une loi de la nature. En effet, alors que dans la définition de Hume il était question d'un mélange entre causalité effective et générale, Mill se débarrasse de ce problème en considérant qu'il s'agit d'une question de causalité générale. Le rôle de celle-ci est de déterminer les lois de la nature. En l'occurrence $o_m \wedge \neg se \leftrightarrow s$ dans l'exemple ci-dessus. Une fois celles-ci établies, il est possible de raisonner sur la causalité effective. Pourvu qu'il existe une loi de la nature déterminant toutes les conditions menant à l'extinction des dinosaures, il est possible de déterminer si la météorite qui est tombée fait partie ou non de l'ensemble des causes effectives de l'extinction des dinosaures.
Le quatrième problème n'est pas résolu. Deux objets étant des conséquences d'une cause commune, peuvent toujours être reliés. Comme nous le verrons par la suite, il s'agit principalement d'un problème lié à l'utilisation de la logique et non un problème des définitions elles mêmes.

Des avancées significatives pour ces approches sont faites par MACKIE [1980]. Il fait une proposition essentielle pour pouvoir gérer les situations où plusieurs ensembles d'objets sont suffisants individuellement à produire la conséquence (disjonctions). Comme Mill, Mackie s'appuie également sur l'idée qu'il existe des lois naturelles et qu'une relation causale est le résultat de l'application d'une telle loi. De plus, il va également militer pour la nécessité d'inclure dans ces lois aussi bien les objets positifs que négatifs. Par contre, il ajoute qu'une loi naturelle est une formule sous forme normale disjonctive (FND), i.e. $(c_{1,1} \wedge \cdots \wedge c_{1,n}) \vee \cdots \vee (c_{k,1} \wedge \cdots \wedge c_{k,n}) \leftrightarrow e$. Sous cette forme, chaque ensemble de conditions $\{c_{i,1}, \ldots, c_{i,n}\}$ correspondant à chacun des éléments disjoints $(c_{i,1} \wedge \cdots \wedge c_{i,n})$ est un ensemble minimalement suffisant et non nécessaire à e. Chaque élément $c_{i,j}$ de ces ensembles minimalement suffisants est non suffisant mais sa présence est nécessaire pour

que l'ensemble $\{c_{i,1}, \ldots, c_{i,n}\}$ soit effectivement suffisant. Notez qu'il existe une différence pour Mackie entre un « ensemble suffisant » et un « *ensemble effectivement suffisant* ». Alors que le premier fait référence à la loi naturelle, donc du point de vue de la causalité générale, le deuxième fait référence à un situation précise où tous les éléments de l'ensemble suffisant sont effectivement présents et donc nous pourrions dire que celui-ci a été effectivement suffisant à produire la conséquence. La deuxième différence la plus importance est que, bien qu'il s'intéresse comme Mill aux ensembles de causes minimalement suffisants, Mackie soutient que les éléments dans ces ensembles peuvent être considérés des causes individuellement. Un objet positif ou négatif c est une cause de l'objet e ssi :

1. c est un élément insuffisant mais non redondant d'un ensemble non nécessaire mais suffisant. Il dira que c satisfait la « Insufficient but Non-redundant part of an Unecessary but Sufficient Condition (INUS) » ;
2. tous les membres d'au moins un ensemble suffisant contenant c sont présents dans la situation étudiée. c fait donc partie d'un ensemble effectivement suffisant ;
3. au moins un élément de tous les ensembles suffisants ne contenant pas c est absent de la situation étudiée.

Exemple 3.2 [suite]**.** *Nous pouvons imaginer que pour Mackie, le déversement des eaux industrielles provenant de l'usine de médicaments o_m, puis le fait que la station d'épuration des eaux usées de l'usine de médicaments soit hors service $\neg se$ est une cause du seuil de pollution de la rivière étant atteint s si : o_m et $\neg se$ satisfont la condition INUS (ce qui est le cas puisqu'il s'agit d'éléments insuffisants mais non redondants de $(o_m \wedge \neg se)$ qui forment un ensemble non nécessaire mais suffisant) ; o_m et $\neg se$ font partie d'un ensemble effectivement suffisant ; au moins un élément de tous les ensembles suffisants ne contenant pas o_m ou $\neg se$ est absent de la situation étudiée (ce qui est le cas puisque le seul autre ensemble suffisant est $\{o_e\}$). Notez que c'est la première définition où nous pouvons considérer la loi $o_e \vee (o_m \wedge \neg se) \leftrightarrow s$ dans son entièreté car pour cela il faut pouvoir traiter la disjonction.*

Comme pour la version proposée par Mill, celle de Mackie est plus aboutie mais ne résout pas tous les problèmes. Dans la définition de Mackie tout repose sur l'existence d'une loi causale et plus aucun mention n'est faite à l'idée d'« objets du même type ». Ce problème là est donc résolu.
Cette définition permet de gérer les situations particulières ou plusieurs objets doivent être réunis (conjonctions) et une partie des situations où plusieurs ensembles d'objets sont suffisants individuellement à produire la conséquence (disjonctions). Il est bien question ici d'uniquement une partie car l'ajout de la dernière condition introduit deux problèmes. Le premier est que dans le cas simple où l'agent responsable de la gestion lance uniquement la production de l'usine d'enceintes, o_e ne serait pas une cause car l'objet ne satisfait pas la condition INUS. Étant donné que $\{o_e\}$ est un singleton, o_e n'est pas un élément insuffisant.
Le deuxième problème est dans le cas de l'exemple 3.4 où l'agent responsable de la gestion lance simultanément la production de l'usine d'enceintes et de médicaments. En effet, la présence de la dernière condition fait qu'aucun objet ne serait considéré comme une cause. Étant donné que les deux ensembles suffisants $\{o_e\}$ et $\{o_m, \neg se\}$ sont effectivement suffisants, il n'y a pas au moins un élément de tous les ensembles suffisants ne contenant pas c qui est absent.
Le troisième problème, celui lié à l'utilisation de la logique, n'est toujours pas résolu. Toutefois, Mackie insiste sur le fait que sa définition n'est pas fausse pour autant, il rétorque qu'il

manque un élément au langage, la « priorité causale ». c est prioritaire causalement à e si un agent aurait en principe pu prévenir e en ne réalisant pas c ou en empêchant que c se produise. Il n'est pas possible de prévenir le lever du soleil en empêchant le coq de chanter. Cette notion de priorité causale est une façon pour Mackie d'introduire une information directionnelle rendant inconcevable de confondre la cause et la conséquence. Par ce moyen il essaye de résoudre la vrai problématique qui se trouve dans l'utilisation de la logique, et plus exactement l'utilisation de l'implication matérielle. La suffisance causale n'est pas la même chose que la suffisance logique car en causalité nous voulons pouvoir aller uniquement dans une direction, la contraposée de l'implication n'est pas souhaitable.

3.1.1.2 De Lewis à Halpern, les approches contrefactuelles

Nous allons dans cette section nous intéresser aux approches contrefactuelles [MENZIES et BEEBEE, 2020]. Selon cette vision de la causalité, la notion de nécessité doit être au premier plan, une cause est quelque chose qui fait la différence : « We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects—some of them, at least, and usually all—would have been absent as well » [LEWIS, 1973]. Cette deuxième approche est également définie par HUME [1748] dans une deuxième formulation de ce qu'est la causalité pour lui. Comme il est possible de le voir dans la citation du début de chapitre, cette deuxième formulation est proposée juste après la première.

> [Second formulation :] Or, in other words, where if the first object had not been, the second had never existed.

Cette deuxième voie n'est pas explorée par Hume et restera inexplorée jusqu'à la seconde moitié du vingtième siècle. Il faudra attendre LEWIS [1973] pour que cela change. Une des raisons avancées pour expliquer cela [MENZIES et BEEBEE, 2020] est la difficulté que représentait définir ce qu'était une contrefactuelle. Qui dit contrefactuelle dit raisonnement sur des situations hypothétiques, des situations ne s'étant pas réalisées. Cette problématique ne s'inscrivait donc pas dans la tradition empiriste de l'époque.
C'est les progrès dans la théorie des mondes possibles reposant sur la sémantique de KRIPKE [1963] qui permet cette exploration. En effet, Lewis définit ce qu'il considère comme une contrefactuelle en s'appuyant sur la sémantique des mondes possibles. Il utilise notamment des comparaison de similarité entre mondes, où un monde est considéré plus proche du monde de référence qu'un second si le premier ressemble plus au monde de référence que le second. Il définit alors la dépendance causale de la façon suivante : étant donné c et e deux évènements distincts s'étant produits, e dépend causalement de c ssi, si c ne s'était pas produit alors e ne se serait pas produit non plus.
Trois remarques peuvent être faites sur cette définition par rapport à celles que nous avons vu dans la section 3.1.1.1. La première est que c et e sont considérés comme des évènements. Il est important pour la suite de préciser que LEWIS [1973] indique bien qu'il s'agit d'un choix de simplification et que d'autres choses peuvent-être reliées par une relation causale : « events are not the only things that can cause or be caused ». La deuxième remarque est que LEWIS [2000] ajoute à sa définition le fait qu'il peut aussi bien s'agir d'évènements que de leur absence. c et e peuvent donc être des occurrences d'évènements ou des non occurrences d'évènements. Finalement, LEWIS [1973] interdit les raisonnements contrefactuels allant à contresens du temps. Il est autorisé de se poser la question qu'en serait-il de e si c

ne s'était pas produit, mais il n'est pas possible d'imaginer qu'en serait-il de c si e ne s'était pas produit. Cette interdiction est justifiée par une comparaison de similarité entre mondes. C'est la façon pour Lewis d'introduire l'information directionnelle rendant inconcevable de confondre la cause et la conséquence.

Exemple 3.2 [suite]**.** *Nous pouvons imaginer que pour Lewis, le seuil de pollution de la rivière étant atteint s dépend causalement du déversement des eaux industrielles provenant de l'usine de médicaments o_m, tous deux des évènements distincts s'étant produits, si : si o_m ne s'était pas produit, alors s ne se serait pas produit non plus. C'est bien le cas pour cet exemple simple.*

Si pendant deux siècles les approches par régularité ont été considérées comme dominantes, durant les cinquante ans qui nous séparent de la publication de LEWIS [1973] ça a été au tour des approches contrefactuelles. Ce schisme créé par Lewis est dû aussi bien à l'aspect satisfaisant intuitivement que la nécessité est une notion centrale en causalité, qu'à la forte critique qu'il fait des approches par régularité. Cette critique porte sur trois points principalement. Le premier repose sur les cas où c identifié comme cause de e serait en réalité un effet de e. Le deuxième repose sur le cas déjà mentionné où c identifié comme cause de e serait en réalité un effet d'une cause commune avec e. Ces deux premiers points soulèvent des problèmes qui ont la même origine, l'utilisation de l'implication matérielle. Le troisième point repose sur le cas où c identifié comme cause de e serait en réalité une cause préemptée de e. Traiter ce cas revient à pouvoir traiter l'exemple 3.3.

Revenons au développement des approches contrefactuelles après Lewis. Il se trouve que les tentatives de définition reposant exclusivement sur la nécessité de la cause, autrement dit sur la dépendance causale aussi appelée dépendance contrefactuelle, échouent à gérer une partie des situations où plusieurs ensembles d'objets sont suffisants individuellement à produire la conséquence (disjonctions) [HALL et PAUL, 2003; MENZIES et BEEBEE, 2020]. Prenons pour exemple le *But-for test*, aussi appelé « Conditio Sine Qua Non » [SATOH et TOJO, 2006]. Ce test couramment utilisé en droit stipule que [WRIGHT, 1985] : « an act was a cause of an injury if and only if, but for the act, the injury would not have occurred ». Il s'agit ici d'une simple reformulation de la dépendance causale. L'erreur commise par ce test est de croire à une équivalence entre causalité et dépendance causale [BECKERS, 2021b].

Exemple 3.3 [suite]**.** *Prenons l'exemple 3.3 et appliquons le but-for test. Est-ce que le seuil de pollution de la rivière aurait été atteint s si le déversement des eaux industrielles provenant de l'usine d'enceintes o_e n'avait pas eu lieu ? Dans cet exemple la réponse est oui. En effet, dans ce scénario le déversement des eaux industrielles provenant de l'usine de médicaments o_m, puis le fait que la station d'épuration des eaux usées de l'usine de médicaments soit hors service $\neg s$e auraient fait qu'en l'absence de o_e, s se serait tout de même produit. o_e n'est pas une cause selon le but-for test, s n'en dépend pas causalement. Autrement dit, o_e n'est pas nécessaire à s. Si nous appliquons le même raisonnement à o_m ou $\neg s$e nous obtenons exactement la même situation, aucun des deux n'est une cause de s. Nous nous retrouvons alors dans une situation problématique où s n'a pas de causes. Le but-for test ne gère donc pas les cas de préemption. Si nous appliquons ce même test au cas de duplication de l'exemple 3.4 nous obtenons le même résultat, s n'aurait pas de causes.*

À partir de la publication de LEWIS [1973] jusqu'à nos jours, de nombreux travaux en philosophie, en droit, en mathématiques et en informatique ont essayé de proposer une

définition satisfaisante reposant principalement sur la dépendance causale. Ce sont les définitions proposées par HALPERN [2016] qui semblent s'imposer aujourd'hui comme les plus acceptées [KUEFFNER, 2021]. La dernière définition qu'à proposé HALPERN [2015] est le résultat d'un long processus itératif qui a commencé par la représentation de la vision de Lewis dans le cadre d'équations structurelles (SEF) (« structural equations framework » en anglais) proposé par PEARL et NEUBERG [2000]. Pour les raisons déjà énoncées précédemment, nous ne rentrerons pas dans les détails formels des définitions proposées par Halpern. Toutefois, pour les besoins propres à cette thèse nous présenterons les quatre éléments qui caractérisent ces définitions, et toutes celles s'en inspirant [HALL, 2007; HITCHCOCK, 2001, 2007; WESLAKE, 2015; WOODWARD, 2005], comme présentés par BECKERS [2021b].
Le premier élément est le formalisme utilisé. Halpern modélise toutes ses définitions en SEF. Ce formalisme présente trois avantages pour le raisonnement causal : il se prête bien au raisonnement contrefactuel, les évènements sont des variables non nécessairement binaires et il intègre par sa conception l'information directionnelle rendant inconcevable de confondre la cause et la conséquence. En effet, dans ce formalisme la valeur des nœuds parents détermine la valeur des nœuds enfants et l'inverse est proscrit.
Le deuxième élément repose sur la notion de *dépendance causale* mentionné précédemment. Si un évènement e est dépendant causalement d'un évènement c, alors l'occurrence de c implique l'occurrence de e. Contrairement au but-for test qui établit une équivalence entre causalité et dépendance causale, pour Halpern l'existence de dépendance causale est suffisante pour parler de causalité mais elle n'est pas nécessaire. Les définitions de Halpern sont donc plus complexes que le but-for test car elles ajoutent d'autres conditions à la dépendance causale.
Le troisième élément est la notion de *contrefactuel causalement* qui va représenter la condition nécessaire pour qu'il y ait causalité. Si c est une cause effective de e dans un scénario, alors c doit pouvoir prendre une autre valeur dans un scénario contrefactuel où avec cette valeur c n'est pas suffisante causalement pour e.
Le quatrième élément est l'*interventionnisme*. Celui-ci permet de relier dépendance causale et causalité : c est une cause effective de e ssi e est dépendant causalement de c étant donné une intervention. Une intervention est une opération permettant de fixer arbitrairement la valeur d'un ensemble d'évènements autres que c et e dans le scénario en respectant un ensemble de conditions. Dans les définitions de Halpern ces conditions n'exigent pas de devoir respecter la cohérence des équations. Par exemple, il est possible de fixer la valeur d'un nœud enfant x dans des scénarios contrefactuels où la valeur de ses nœuds parents sont censés donner une autre valeur à x que celle fixée. La façon dont l'interventionnisme permet de relier dépendance causale et causalité est décrite par BECKERS [2021b] en utilisant le formalisme SEF de la façon suivante :

> Interventionism : They all share the assumption [HP-style definitions] that the relation between counterfactual dependence and causation takes on the following form : C = c causes E = e iff E = e is counterfactually dependent on C = c given an intervention $\vec{X} \leftarrow x$ that satisfies some conditions P. The divergence between these definitions is to be found in the condition P that should be satisfied.

Exemple 3.4 [suite]**.** *Prenons l'exemple 3.4 et appliquons un raisonnement reposant sur la vision commune aux définitions de Halpern. La figure 3.1 illustre la représentation de cet exemple en SEF.*
Pour HALPERN *[2016], les ensembles* $\{o_e = 1, o_m = 1\}$ *et* $\{o_e = 1, se = 0\}$ *sont les causes effectives de* s*. Pour trouver cela, il se pose deux questions : (i) est-ce que le seuil de pollution de la rivière aurait été atteint* s *si le déversement des eaux industrielles provenant de l'usine d'enceintes* o_e *n'avait pas eu lieu et que le déversement des eaux industrielles provenant de l'usine de médicaments* o_m *n'avait pas eu lieu ? (ii) est-ce que le seuil de pollution de la rivière aurait été atteint* s *si le déversement des eaux industrielles provenant de l'usine d'enceintes* o_e *n'avait pas eu lieu et que la station d'épuration de l'usine de médicaments n'était pas hors service ? La réponse à ces questions est non, le seuil de pollution de la rivière n'aurait pas été atteint dans ces deux cas. En effet, il se place dans les cas où : (i)* $o_e = 0$, $o_m = 0$, $se = 0$ *et donc* $s = 0$ *; (ii)* $o_e = 0$, $o_m = 1$, $se = 1$ *et donc* $s = 0$*. Ces questions visent à déterminer s'il existe une dépendance causale. Dans les deux cas, cette dépendance causale existe. Seuls les ensembles satisfont ces conditions, aucun des éléments ne le fait individuellement et donc ils ne peuvent pas être considérés comme des causes individuellement.*

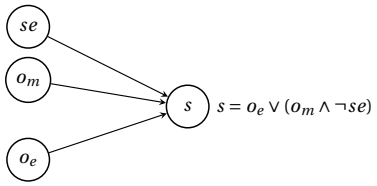


FIGURE 3.1 – Illustration de la représentation de l'exemple 3.4 en SEF.

Exemple 3.3 [suite]**.** *Prenons l'exemple 3.3 et appliquons un raisonnement reposant sur la vision commune aux définitions de Halpern. La figure 3.2 illustre la représentation de cet exemple en SEF. Cette représentation est obtenue après avoir ajouté des variables supplémentaires comme le fait classiquement* HALPERN *[2016] pour résoudre les problèmes de préemption, comme dans un de ces exemples les plus utilisés, le cas de Suzy et Billy que nous détaillons dans l'exemple 3.5. La première variable ajoutée est* l_e *indiquant que les eaux usées de l'usine d'enceintes ont pu être déversées dans le lac. L'équation associée à cette variable est* $l_e = o_e$*. La deuxième variable ajoutée est* l_m *indiquant que les eaux usées de l'usine de médicaments ont pu être déversées dans le lac. L'équation associée à cette variable est* $l_m = o_m \wedge \neg se \wedge \neg l_m$*. Ces deux variables sont ajoutées pour capturer le fait que les eaux usées de l'usine de médicaments ne peuvent pas être déversées si celles de l'usine d'enceintes le sont en premier.*
Pour HALPERN *[2016], l'ensemble* $\{o_e = 1\}$ *est la cause effective de* s*. Pour trouver cela, il se pose la question : est-ce que le seuil de pollution de la rivière aurait été atteint* s *si le déversement des eaux industrielles provenant de l'usine d'enceintes* o_e *n'avait pas eu lieu ? La réponse à cette question est oui, le seuil de pollution de la rivière aurait été atteint dans ce cas. En effet, il se place dans les cas où :* $o_e = 0$, $l_e = 0$, $o_m = 1$, $se = 0$, $l_m = 1$ *et donc* $s = 1$*. Cette question vise à déterminer s'il existe une dépendance causale. Sans interventionnisme, il n'existe pas de dépendance causale, comme lorsque nous avons appliqué le but-for test à l'exemple. Toutefois, avec l'interventionnisme, il décide de fixer la valeur de* $l_m \leftarrow 0$ *malgré que les valeurs de* $o_e = 1$, $se = 0$, *et* $l_e = 0$ *sont censées donner* $l_m = 1$*. Il justifie cette valeur en disant que dans la situation réelle, les eaux usées de l'usine de médicaments n'ont pas pu être déversées dans le*

lac. Grâce à l'interventionnisme, il se place dans le cas où : $o_e = 0$, $l_e = 0$, $o_m = 1$, $se = 0$, $l_m = 0$ *et donc* $s = 0$. *Dans ce cas là, cette dépendance causale existe.*

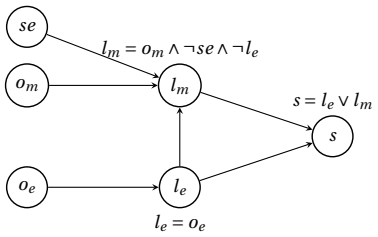


FIGURE 3.2 – Illustration de la représentation de l'exemple 3.3 en SEF.

Plus récemment, BATUSOV et SOUTCHANSKI [2018] et KHAN et LESPÉRANCE [2021] ont travaillé sur une modélisation en Calcul des Situations d'une définition de causalité effective reposant sur la vision de HALPERN [2015]. Cette définition stipule qu'un évènement e cause directement une formule φ si c'est le dernier évènement qui a changé la valeur de la formule de fausse à vraie. Les chaînes causales sont ensuite formées par rétro propagation. Plus précisément, en identifiant la cause directe e' qui a fait en sorte que e puisse se produire et que e rende effectivement φ vrai, et ainsi de suite. Cette définition s'appuie fortement sur l'hypothèse de non cooccurrence du Calcul des Situations ce qui l'empêche de gérer certains cas de surdétermination.
Les approches Seeing To it That (STIT) comme celles proposées par LORINI et collab. [2014] ou ABARCA et BROERSEN [2022] font également partie de cette famille où la nécessité est centrale. Du fait de leur approche modale, ces travaux STIT intègrent facilement des aspects épistémiques. Cette prédisposition fait que la plupart de ces travaux sont plus centrés sur la responsabilité que la causalité, question hors du cadre de cette thèse. De plus, les approches STIT classiques se concentrent sur la relation entre l'agent et les états du monde, plaçant les évènements dans un second plan.

3.1.1.3 Approches par inférence, suite des approches par régularité

Nous allons dans cette section nous intéresser aux approches par inférence [ANDREAS et GUENTHER, 2021]. Il s'agit d'une famille d'approches qui continuent dans la lignée de ce qui avait été fait par les approches par régularité vues dans la section 3.1.1.1. Il est possible de décrire ces théories comme un raffinement des théories par régularité. En effet, elles résolvent le problème lié à l'utilisation de la logique provenant de l'inadéquation de l'implication matérielle pour représenter la causalité. Cela est fait en modélisant les définitions dans des langages reposant sur la logique, mais qui disposent de moyens pour définir la suffisance causale. Cette suffisance causale introduit l'information directionnelle rendant inconcevable de confondre la cause et la conséquence.
Les approches causales par inférence partagent une structure commune. Un objet c y est considéré une cause de e ssi :

1. c et e sont présents dans la situation étudiée ;
2. e peut être inféré de c dans une théorie basée sur une logique appropriée.

Par le biais d'une série de travaux dans le domaine du droit, WRIGHT [1985, 1988, 2011] pose une des dernières pièces manquantes à l'édifice initié par Hume deux siècles et demi

plus tôt. Il reprend l'idée du test INUS de Mackie et règle le problème des quelques cas de disjonction qui résistaient encore aux approches par régularité en proposant un nouveau test. Pour WRIGHT [2011] :

> A condition *c* was a cause of a consequence *e* if and only if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of *e*.

Dit plus simplement, une condition c est une cause d'une condition e ssi elle satisfait le test Necessary Element for the Sufficiency of a Set (NESS). Ce test repose sur trois éléments essentiels. Le premier est la suffisance causale des causes, à ne pas confondre avec la suffisance logique. D'après WRIGHT [2011], la première, contrairement à la dernière, intègre l'aspect temporel de la causalité faisant que cause et conséquence ne peuvent jamais être confondues : « The successional nature of causation is incorporated in the concept of causal sufficiency, which is defined as the complete instantiation of all the conditions in the antecedent of the relevant causal law ». En plus de cet aspect purement temporel, nous retrouvons ici l'idée d'un ensemble effectivement suffisant.
Le deuxième élément essentiel est la nécessité faible des causes. Nous sommes loin de la nécessité forte d'autres théories qui posent tant de problèmes pour les cas disjonctifs, où en l'absence de la cause la conséquence ne peut pas être vraie. Ici la nécessité est dite faible car elle est subordonnée à la suffisance ; la cause n'est pas nécessaire à la conséquence directement, mais au fait que toutes les conditions de l'ensemble suffisant auquel elle appartient soient vraies : « a causally relevant factor need merely be necessary for the sufficiency of a set of conditions sufficient for the occurrence of the consequence, rather than being necessary for the consequence itself ».
Le troisième élément essentiel est l'effectivité des causes. Nous sommes en causalité effective, il n'est pas envisageable de considérer des causes des évènements ne s'étant pas produits. Wright parle bien d'un ensemble de causes s'étant produites : « set of existing antecedent conditions ».

Exemple 3.3 [suite]**.** *Reprenons les exemples 3.3 et 3.4 qui posaient certains problèmes à la définition de Mackie et appliquons le test NESS. Dans ces exemples, à partir de la formule* $o_e \vee (o_m \wedge \neg se) \leftrightarrow s$*, nous obtenons deux ensembles suffisants à causer* s *:* $\{o_e\}$ *et* $\{o_m, \neg se\}$*. Notez que nous parlons de suffisance par rapport à la formule, donc en causalité générale, et pas effectivement suffisants. Ils peuvent l'être, mais ce n'est pas toujours le cas. Cela dépend de l'exemple dans lequel nous nous situons. Dans l'exemple 3.4 les deux ensembles ont toutes leurs conditions présentes et ont donc été effectivement suffisants. Aussi bien* o_e*,* o_m *et* $\neg se$ *sont des éléments nécessaires à la suffisance effective de leurs ensembles respectifs. Pour Wright, ces trois conditions sont donc des causes de* s*.*
Dans l'exemple 3.3, uniquement l'ensemble $\{o_e\}$ *a toutes ses conditions présentes à un moment donné. La suffisance effective de* $\{o_m, \neg se\}$ *a été préemptée par celle de l'ensemble* $\{o_e\}$*.* o_e *est un élément nécessaire à la suffisance effective de cet ensemble. Pour Wright,* o_e *est la seule cause de* s*.*

Placer les travaux de Wright comme appartenant aux approches par inférence n'est possible qu'en étudiant ses travaux en détail. Le fait que Wright utilise exactement les mêmes outils formels que Hume, Mill et Mackie fait qu'à premier abord ses travaux ne devraient

pas être classés dans la catégorie des approches par inférence. Cela permet d'expliquer que, malgré que des auteurs influents dans les approches contrefactuelles comme PEARL et NEUBERG [2000] aient accepté que le test NESS traduise bien l'intuition de ce qu'est la causalité, sa proposition ait reçu les mêmes critiques que les approches par régularité qui l'ont précédé. La solution qu'apporte Wright n'est pas présentée formellement, mais à l'aide de raisonnements décrits en langage naturel. Ces éléments présentés en langage naturel apportent une solution à la façon des théories par inférence. Dans l'énonciation du test NESS, il est évident que Wright voit que les éléments qu'il avance résolvent les problèmes des approches par régularité. Toutefois, étant un juriste n'ayant pas nécessairement une connaissance poussée de la logique, il est concevable qu'il n'est pas identifié le besoin d'enrichir le formalisme utilisé. De plus, Wright est attaché à ce formalisme car associé à la conception de Mill et Mackie qu'il partage selon laquelle l'enquête causale consiste à essayer de déterminer quelles lois causales ont été appliquées dans le cas particulier par les conditions qui ont eu lieu dans cette occasion spécifique. Wright va même plus loin en ajoutant qu'en causalité effective nous nous intéressons à une partie seulement des lois causales, soit parce que notre connaissance sur la loi est incomplète, soit parce que la spécification détaillé de tous les éléments composant la loi serait très pénible, voir irréalisable. Tous ces éléments font que Wright est à cheval entre les approches par régularité et les approches par inférence. D'un côté, son formalisme le place dans la première famille d'approches. De l'autre, le raisonnement qu'il développe et la mise en avant du besoin de parler de suffisance causale le placent dans la deuxième famille d'approches.
La solution de l'exemple 3.3 montre la dualité dans laquelle se trouve Wright. Le raisonnement permettant de trouver le résultat inclut une analyse temporelle qui n'est pas intégrée au langage utilisé. Nous savons que la suffisance de $\{o_m, \neg se\}$ n'est pas effective que par la description qui est faite de l'exemple.

Suite à la proposition de Wright, divers travaux se sont intéressés au test NESS et ont proposé d'aboutir ce qu'il avait commencé en le modélisant. BAUMGARTNER [2013] propose une définition du test NESS en utilisant les même outils formels que Hume, Mill, Mackie et Wright. Toutefois, il ajoute des contraintes de minimalité pour résoudre les problèmes de ces approches.
BOCHMAN [2018a,b] propose une formalisation du test NESS en Causal Calculus, une logique appropriée pour parler d'inférence en causalité. Il étend la logique propositionnelle classique en ajoutant une relation causale permettant de lier deux propositions. Cette relation causale est comme une implication sans contraposée et sans réflexivité. Si nous avions la relation « o_e cause s » et que cette relation de cause était une implication matérielle, accepter sa contraposée reviendrait à dire que « $\neg s$ cause $\neg o_e$ ». Autrement dit, cela reviendrait à accepter que le fait que le seuil de pollution n'ait pas été atteint est une cause du non déversement des eaux usées. Si la relation de cause était une implication matérielle, cette relation serait réflexive et nous accepterions des relations de la forme s cause s. Le Causal Calculus est une combinaison des travaux de TURNER [1997], dont $\mathscr{B}$ vu en 2.2.4 est une sous partie, et la « logic of causal rules » introduite par BOCHMAN [2004].
BERREBY et collab. [2018] proposent une définition de causalité effective qui revient à un test NESS simplifié dans ce que BOURGNE et collab. [2021] appellent un « fragment STRIPS du Discrete Event Calculus ». Le langage proposé peut être vu comme une implémentation du langage de description d'action $\mathscr{E}$ vu en 2.2.7. Il est possible de parler ici d'une version du test NESS simplifiée car ce langage ne permet pas de représenter des préconditions dis-

jonctives, tous les problèmes classiques de la causalité ne peuvent donc pas être traités.
LeBlanc et collab. [2019] proposent une démarche très similaire à celle de Berreby et collab. [2018] mais dans le langage de description d'action $\mathscr{AL}$ introduit par Baral et Gelfond [2000]. $\mathscr{AL}$ est à $\mathscr{B}$ ce que $\mathscr{A}_c$ est à $\mathscr{A}$, i.e. une extension permettant de traiter la cooccurrence d'actions. Cette version ne gère pas les disjonctions non plus.
Finalement, Beckers [2021b] propose une implémentation du test NESS en SEF. Il montre ainsi que cette définition peut être représentée dans le langage de prédilection des approches contrefactuelles. Tous ces travaux récents replacent les approches par inférence en concurrence avec les approches contrefactuelles.

3.1.2 Problématiques dans le domaine de la causalité

Ce parcours historique terminé, parlons de deux problématiques importantes dans le domaine de la causalité. La section 3.1.2.1 traite de surdétermination, des cas où plus d'une cause est suffisante à elle seule à produire une conséquence. Sur ces problèmes reposent la plupart des débats encore existants dans le domaine. La section 3.1.2.2 montre l'importance d'établir une distinction claire entre causalité et responsabilité. Bien séparer ces concepts permet de comprendre et donc résoudre une partie importante des problématiques posées par la surdétermination.

3.1.2.1 La surdétermination, ou la pomme de la discorde

Maintenant que nous avons un aperçu global de ce qu'est la causalité effective, nous allons nous intéresser dans cette section à la surdétermination, des cas particulièrement importants en causalité car sur eux reposent la plupart des débats encore existants dans le domaine. Il s'agit de cas où plus d'une cause était suffisante à elle seule à produire une conséquence, comme dans les exemples 3.3 et 3.4, ou dans l'exemple emblématique de Suzy et Billy très souvent utilisé par Halpern [2016].

Exemple 3.5 [Suzy et Billy]**.** *Suzy et Billy lancent simultanément une pierre sur la même bouteille. La pierre de Suzy frappe en premier et brise la bouteille, une fraction de secondes plus tard la pierre de Billy passe à l'emplacement de la bouteille avant qu'elle n'ait été percutée. La pierre de Billy aurait brisée la bouteille si elle avait été là. Il n'est pas possible de déduire que le lancer de Suzy est la cause avec un simple but-for test, i.e. en regardant simplement ce qui se serait passé si elle n'avait pas lancé.*
La figure 3.3 illustre la représentation classique de cet exemple en SEF par Halpern [2016]. Cette représentation est obtenue après avoir ajouté des variables supplémentaires comme pour l'exemple d'application de la définition de Halpern à l'exemple 3.3 dans la section 3.1.1.2. ST *correspond à « Suzy throws »,* BT *correspond à « Billy throws »,* SH *correspond à « Suzy hits »,* BH *correspond à « Billy hits » et* BS *correspond à « bottle shatters ».*

Deux raisons principales expliquent l'importance des cas de surdétermination dans le domaine de la causalité. La première est que de nombreuses situations dont il est important de connaître les causes sont presque inévitablement des cas de surdétermination. Mentionnons par exemple le réchauffement climatique, la pollution d'un site protégé, la délocalisation d'un site industriel, l'inégalité de représentation en politique entre les femmes et les hommes, l'existence de déserts médicaux, une perte économique pour une filière agricole ou le suicide d'une personne.

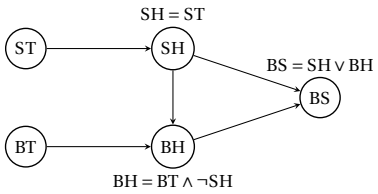


FIGURE 3.3 – Suzy and Billy throwing a rock example.

La deuxième raison est qu'une partie importante des débats encore ouverts sur la notion de causalité ont à voir de près ou de loin avec la surdétermination, et cela que ce soit en philosophie, psychologie, droit, mathématiques ou informatique. Lorsqu'une nouvelle définition de causalité est proposée, elle est la plupart du temps validée en étant confrontée à des exemples complexes. Ces exemples sont tous des exemples de surdétermination. Il existe différents types de cas de surdétermination et selon le cas de surdétermination dont il est question, les relations causales attendues ne sont pas les mêmes. Il existe un désaccord sur les relations causales attendues étant donné un type de cas de surdétermination donné.

Étudions à présent les types de cas de surdétermination les plus communément discutés. À notre connaissance, ces catégories n'ont été définies que de manière générale en utilisant le langage naturel. Voici pour chaque type de surdétermination une description extraite de travaux dans le domaine accompagnée d'un exemple :

— surdétermination symétrique/duplicative (« symmetric/duplicative » en anglais) : une situation où « a factor other than the specified act would have been sufficient to produce the injury in the absence of the specified act, but its effects [...] combined with or duplicated those of the specified act to jointly produce the injury » [WRIGHT, 1985].

Exemple 3.6 [symétrique/duplicative]**.** *Deux assassins versent une dose létale de poison dans la boisson de la victime [*HITCHCOCK*, 2007] ;*

— surdétermination imitative (« trumping » en anglais) : une situation où « the pre-empted factor runs its whole course–a course that normally produce some event *e*–yet that factor does not cause *e* on this occasion because some pre-empting cause 'trumps' it » [HALL et PAUL, 2003].

Exemple 3.7 [imitative]**.** *Un bateau sur un fleuve est obligé de s'arrêter car la rivière est bloquée. Non seulement le pont* A *s'est effondré dans la rivière à quelques mètres du bateau, le pont* B *s'est également effondré un peu plus loin dans la rivière bloquant également le passage [*WRIGHT*, 2011] ;*

— surdétermination préemptive précoce (« early preemption » en anglais) : une situation où « the pre-emption of the pre-empted causal sequence occurs before the completion of the completed causal sequence » [WRIGHT, 2011].

Exemple 3.8 [préemptive précoce]**.** *Un assassin* A *empoisonne la gourde d'un voyageur dans le désert, gourde étant sa seule source d'eau. Puis un assassin* B *vide la gourde avant même que le voyageur puisse y boire. Le voyageur est retrouvé mort déshydraté quelques jours plus tard [*BOCHMAN*, 2018b] ;*

— surdétermination préemptive tardive (« late preemption » en anglais) : une situation où « the pre-emption of the pre-empted causal sequence occurs at the same time as (or after) the completion of the completed causal sequence » [WRIGHT, 2011].

Exemple 3.9 [préemptive tardive]**.** *Deux feux de forêt sont déclenchés et chacun peut à lui seul détruire une maison. Il se trouve que l'un des deux feux atteint la maison en premier et brûle la maison dans son entièreté juste avant que le deuxième feu n'atteigne la maison [*BATUSOV *et* SOUTCHANSKI*, 2018].*

Le fait qu'il n'existe pas de définition formelle de ces différents types de problèmes se prête à une variété d'interprétations et donc de confusions. Par exemple, BAUMGARTNER [2013] retrace l'histoire des cas de surdétermination préemptive et montre que différentes interprétations existent :

> In recent years, there has been some variance in the literature as to what exactly the difference between early and late preemption amounts to. According to the understanding which hearkens back to LEWIS [1987] [...], the difference is that 'in cases of early preemption, the backup process is cut off before the effect occurs, whereas in cases of late preemption, the process is cut off by the effect itself' [HITCHCOCK, 2007]. By contrast, e.g. HALL et PAUL [2003] hold that the characteristic feature of early preemption is that a process is interrupted by another process, whereas in cases of late preemption no interruption takes place, rather, the preempted process just does not run to completion.

La citation ci-dessus donne lieu à trois remarques. La première est qu'un autre type de préemption tardive peut être envisagée. La spécificité de la variante proposée par HALL et PAUL [2003] est que, du point de vue de la représentation de l'action et du changement, cette variante ne peut être correctement représentée que par des processus duratifs. En effet, si « no interruption takes place » mais que le « process just does not run to completion », la seule explication est qu'un évènement duratif commence et ne parvient jamais à produire l'effet escompté parce que celui-ci est rendu vrai par un autre processus plus rapide. Nous appelons cette variante surdétermination préemptive tardive durative. Toutefois, ce cas ne peut être traité que dans une approche reposant sur une représentation de l'action et du changement capable de tenir compte des durées, ce n'est le cas d'aucune approche causale dont nous ayons connaissance.
La deuxième remarque est que les descriptions de surdétermination préemptive tardive et imitative semblent décrire les mêmes cas. Il ressort des discussions de la littérature que la surdétermination imitative est le cas le plus mal défini.
Finalement, notons que pour décrire ces cas les termes utilisés renvoient à l'existence de ce que nous pourrions appeler des chemins causaux. En effet, nous retrouvons dans la description des termes comme « (completed) causal sequence », « completion », « process », « backup process », « cut off » et « interrupted ». À notre connaissance, la notion de chemin causal a déjà été définie dans le domaine [BAUMGARTNER, 2013; BECKERS, 2021b], mais jamais pour traiter formellement des cas de surdétermination.

Il est possible de dire que s'intéresser à la causalité dans toute sa complexité passe nécessairement par s'intéresser à la surdétermination. Lorsqu'il est question de surdétermination, il est inévitablement question de préconditions disjonctives et très souvent de cooccurrence d'évènements. Ainsi, dès lors qu'une proposition causale est faite, si elle souhaite

traiter le problème dans toute sa complexité, elle doit nécessairement s'appuyer sur une représentations du monde contenant les deux éléments.

3.1.2.2 Causalité effective n'est pas responsabilité

Dans cette section nous allons parler de l'importance d'établir une distinction claire entre causalité et responsabilité. Dans une série d'articles influents, WRIGHT [1985, 1988] démontre à quel point une enquête causale est essentielle dans le processus de détermination de la responsabilité juridique. Il souligne la différence fondamentale entre la causalité et la responsabilité, la première notion est purement factuelle alors que la deuxième est subjective, elle est sujette à des aspects normatifs. Ce faisant, il montre que l'enquête causale doit impérativement être factuelle et donc indépendante de tout aspect normatif. Cette distinction peut également être retrouvée dans d'autres travaux comme dans la taxonomie de VINCENT [2011] qui différencie la « responsabilité causale » et la « responsabilité du résultat ». La distinction entre causalité et responsabilité est fondamentale, bien séparer ces concepts est la clé de voûte qui permet à Wright de résoudre une partie importante des problématiques posées par la surdétermination.

Exemple 3.10 [causalité et responsabilité]**.** *Reprenons l'exemple 3.8 où un assassin* A *empoisonne la gourde d'un voyageur dans le désert, gourde étant sa seule source d'eau. Puis un assassin* B *vide la gourde avant même que le voyageur puisse y boire. Le voyageur est retrouvé mort déshydraté quelques jours plus tard [*BOCHMAN*, 2018b].*
Il existe formellement quatre réponses à ce problème. La première consisterait à dire qu'aucun des deux assassins n'est une cause de la mort du voyageur. Cette première réponse est celle que donnerait le but-for test. Il est communément admis qu'il s'agit d'une preuve que celui-ci ne fonctionne pas, plus que d'une solution à envisager sérieusement.
La deuxième consisterait à dire qu'uniquement l'assassin A *est une cause de la mort du voyageur. Une façon de justifier cette deuxième réponse consiste à dire que même si le voyageur n'est pas mort empoisonné, l'assassin* A *l'a condamné. L'assassin* B *n'est pas une cause car le voyageur était déjà condamné, il n'a pas changé le résultat final. Il est même possible d'aller plus loin et dire que* B *n'est pas une cause car il serait possible de considérer que son intention était de sauver le voyageur car il savait que l'eau avait été empoisonnée.*
La troisième consisterait à dire qu'uniquement l'assassin B *est une cause de la mort du voyageur. Pour justifier cette troisième réponse il suffit de se contenter des faits et dire que le voyageur est retrouvé mort déshydraté. Il y a bien un chemin causal reliant l'action de* B *à la mort du voyageur. L'assassin* A *n'est pas une cause puisque, certes le chemin causal de l'empoisonnement a été initié, mais celui-ci n'a pas abouti car il a été interrompu par celui initié par* A.
La quatrième consisterait à dire qu'aussi bien l'assassin A *que* B *sont des causes de la mort du voyageur. Ils ont tout deux commis un acte mettant en danger la vie du voyageur.*
L'enquête causale étant un processus factuel, il n'existe qu'une unique réponse possible. Par contre, si la question n'est pas quelle est la cause de la mort du voyageur mais qui en est responsable juridiquement, alors là il est possible d'envisager toutes les solutions, même la première. Les différentes positions correspondront à des aspects normatifs différents. Par exemple, cela pourrait dépendre de la législation du pays dans laquelle le cas est jugé. Ou, cela pourrait également dépendre des croyances et des intentions attribuées aux différents agents de l'exemple.

WRIGHT [1985] décrit le processus permettant de déterminer si un individu est juridiquement responsable d'un préjudice. Ce processus comporte trois étapes :

— enquête sur la conduite délictueuse (« tortious-conduct inquiry ») : où sont identifiées les conduites du défendeur qui pourraient potentiellement impliquer une responsabilité juridique (intentionnelle, négligente, dangereuse, etc.).

Exemple 3.10 [suite]**.** *Reprenons les quatre cas que nous avons envisagé ci-dessus mais en nous intéressant à l'enquête sur la conduite délictueuse au lieu de la causalité effective. Le premier cas consiste à dire qu'aucun des deux assassins n'a eu une conduite délictueuse. Imaginons que l'assassin* A *arrive à prouver qu'il a mis du poison mais pas suffisamment pour que le voyageur meurt. En plus, il explique qu'il savait qu'un autre assassin ferait le travail si lui ne le faisait pas et donc qu'il a agit pour sauver le voyageur. Il est possible que du point de vue de certains agents sa conduite ne soit pas considérée comme délictueuse. Maintenant imaginons que l'assassin* B *arrive à prouver que vider la gourde était un accident ou qu'il avait vu ce qu'avait fait l'assassin* A *et qu'il souhaitait sauver le voyageur. Dans le premier cas, après avoir prouvé qu'il s'agissait bien d'un accident et que cet accident ne relève pas dans cette situation d'un acte négligent, il est possible que du point de vue de certains agents sa conduite ne soit pas considérée comme délictueuse. Dans le deuxième cas, après avoir prouvé que son intention était bien de sauver le voyageur et qu'il n'y avait pas de façon plus sure de le faire, il est possible que du point de vue de certains agents sa conduite ne soit pas considérée comme délictueuse.*
Le deuxième cas consiste à dire qu'uniquement l'assassin A *a eu une conduite délictueuse. Il est possible que du point de vue de certains agents le fait d'empoisonner la boisson soit considéré en toute circonstances comme une action intentionnelle et dangereuse, quelle que soit la dose introduite. Dans ce cas, la conduite de l'assassin* A *serait bien considérée comme délictueuse. Pour dire que ce n'est pas le cas de celle de* B *il suffit de prendre les exemples du premier cas.*
Le troisième cas consiste à dire qu'uniquement l'assassin B *a eu une conduite délictueuse. Il est possible que du point de vue de certains agents le fait de vider la gourde d'un voyageur dans le désert soit considéré en toute circonstances comme une action intentionnelle et dangereuse, quelle que soit l'intention. Ces agents peuvent par exemple considérer qu'il existait une meilleure alternative, malgré ce que dit l'assassin* B*. Dans ce cas, la conduite de l'assassin* B *serait bien considérée comme délictueuse. Pour dire que ce n'est pas le cas de celle de* A *il suffit de prendre les exemples du premier cas.*
Le quatrième cas consiste à dire qu'aussi bien l'assassin A *que* B *ont eu une conduite délictueuse. Il suffit pour cela de combiner le deuxième et troisième cas de façon adéquate.*
Il s'agit là de toutes sortes de considérations qui peuvent polluer le raisonnement causal. Chacun de ces cas est une étape permettant de justifier une décision finale quant à la responsabilité juridique de chacun des assassins. Toutefois, il est important de garder à l'esprit que cette décision est subjective car attachée à des choix propres aux agents qui délibèrent. Cette subjectivité est inévitable dans le cadre plus large de la responsabilité ;

— enquête causale (« causal inquiry ») : où est évalué si les conduites délictueuses identifiées ont réellement contribué à causer le préjudice, c'est-à-dire si elles peuvent être considérées comme des causes du préjudice.

Exemple 3.10 [suite]**.** *Seul l'assassin* B *est une cause de la mort du voyageur. Comme mentionné précédemment, l'enquête causale doit se contenter d'étudier les faits. Dans ce cas là, le voyageur est retrouvé mort déshydraté. Il y a bien un chemin causal reliant l'action de* B *à la mort du voyageur. L'assassin* A *n'est pas une cause puisque, certes le chemin causal de l'empoisonnement a été initié, mais celui-ci n'a pas abouti car il a été interrompu par celui initié par* A*. Le fait que l'assassin* B *soit la cause ne veut pas nécessairement dire qu'il est responsable juridiquement. Inversement, le fait que l'assassin* A *ne soit pas la cause, ne veut pas nécessairement dire qu'il n'a aucune responsabilité juridique. Ce dernier peut par exemple être reconnu comme responsable d'une tentative d'homicide ;*

— enquête sur la proximité de la cause (« proximate-cause inquiry ») : où d'autres facteurs sont examinées, afin d'évaluer s'ils atténuent ou éliminent la responsabilité juridique du défendeur pour le préjudice.

Exemple 3.10 [suite]**.** *Reprenons les quatre cas que nous avons envisagé ci-dessus mais en nous intéressant à l'enquête sur la proximité de la cause. Le premier cas consiste à dire qu'aucun élément ne permet d'atténuer ou éliminer la responsabilité juridique d'un des deux assassins. Imaginons que durant le procès des preuves fortes soient amenées montrant qu'aucun des deux assassins ne savait de l'existence d'un autre danger pour le voyageur. Il est possible que du point de vue de certains agents cela fasse que ni la responsabilité de l'assassin* A*, ni celle de l'assassin* B *ne soit atténuée. Sans cette connaissance, il est impossible pour les assassins de justifier que leur intention respective était de sauver le voyageur.*
Le deuxième cas consiste à dire qu'il existe uniquement des éléments qui permettent d'atténuer ou d'éliminer la responsabilité juridique de l'assassin A*. Il est possible que du point de vue de certains agents le fait que le voyageur ne soit pas mort par empoisonnement atténue sa responsabilité.*
Le troisième cas consiste à dire qu'il existe uniquement des éléments qui permettent d'atténuer ou d'éliminer la responsabilité juridique de l'assassin B*. Il est possible que du point de vue de certains agents le fait que le voyageur était condamné au moment où l'assassin* B *a vidé la gourde atténue sa responsabilité.*
Le quatrième cas consiste à dire qu'il existe des éléments qui permettent d'atténuer ou d'éliminer la responsabilité juridique de l'assassin A *et* B*. Il suffit pour cela de combiner le deuxième et troisième cas de façon adéquate.*
De nouveau, il s'agit là de toutes sortes de considérations qui peuvent polluer le raisonnement causal. Chacun de ces cas est une étape permettant de justifier une décision finale quant à la responsabilité juridique de chacun des assassins. Toutefois, il est important de garder à l'esprit que cette décision est subjective car attachée à des choix propres aux agents qui délibèrent.

De ces trois étapes, seule la deuxième est entièrement indépendante des choix politique et donc factuelle. Elle permet de déterminer si une action est à l'origine du préjudice. Les deux autres étapes sont soumises à des considérations normatives qui déterminent les causes qui donneront lieu à une responsabilité juridique.

L'observation de Wright est que les notions de causalité et responsabilité sont trop souvent confondues car la frontière entre ces notions est mince. Très souvent dans le langage

commun, lorsque nous entendons « la cause » nous pensons en réalité à « la cause responsable ». Cette confusion est quelque part compréhensible, dans la vie quotidienne nous sommes plus souvent confrontés à raisonner en termes de responsabilité car nous vivons dans des sociétés où les normes sont omniprésentes, qu'elles soient formelles comme les normes institutionnelles ou étatiques, ou informelles comme les normes sociales. Établir une distinction claire entre causalité et responsabilité est indispensable pour la tâche qui incombe au domaine de la causalité.

3.2 Analyse des besoins pour une formalisation de la causalité effective

Dans cette section nous exposons les défis qu'il reste à traiter dans le domaine de la causalité, notamment si nous voulons avoir une approche causale commune pour l'éthique computationnelle. La plus grande partie des contributions de cette thèse sont une proposition de solution à ces défis.

Pour plus de clarté nous décomposons la causalité en deux parties. D'un côté ce que nous appelons la *causalité positive* qui permet de s'interroger sur les causes d'une partie de l'état du monde ou d'évènements se produisant. Par exemple, la causalité positive nous permet de déterminer les causes d'un lac ayant atteint un seuil de pollution. De l'autre ce que nous appelons la négation dans la relation causale, ou *causalité négative* pour simplifier, qui permet de s'interroger aussi bien sur les causes derrière le fait qu'un évènement ne se soit pas produit, que des conséquences que peut avoir la non occurrence d'un évènement. Par exemple, la causalité négative concerne le cas où une maître-nageuse empêche un enfant de se noyer, ou au contraire omet de sauver l'enfant.

Nous identifions quatre défis principaux. Le premier consiste à avoir une approche de causalité positive commune adaptée à l'éthique computationnelle. Cela passe tout d'abord par une définition adaptée qui sépare clairement causalité et responsabilité. De cette façon nous nous assurons que l'approche utilisée convient à toutes les théories morales qui pourraient être formalisées. Bien que nécessaire, une définition adaptée n'est pas suffisante. Cela passe également par une modélisation de cette définition dans un formalisme adapté.
Le deuxième défi concerne une partie de la causalité négative, il s'agit du cas où nous souhaitons connaître les causes pour lesquelles une condition ne s'est pas produite. S'intéresser à ces causes fait sens en éthique computationnelle, nous voulons être en mesure de pouvoir raisonner sur le fait qu'un agent ait agi pour empêcher le mal ou le bien de se produire. Dans l'exemple 3.1, si nous voulons évaluer l'action d'un agent qui répare la station d'épuration des eaux usées de l'usine de médicaments il semble important de pouvoir relier cette action au fait que dans certains scénarios elle empêche le seuil de pollution d'être atteint.
Le troisième défi concerne la deuxième partie de la causalité négative, le cas où nous souhaitons connaître les conséquences d'une condition ne s'étant pas produite. S'intéresser à ces conséquences fait sens en éthique computationnelle, nous voulons être en mesure de pouvoir raisonner sur le fait qu'un agent aurait pu agir pour empêcher le mal ou causer le bien, mais ne l'a pas fait. Dans l'exemple 3.1, si l'agent en charge de faire la maintenance de la station d'épuration des eaux usées de l'usine de médicaments ne l'a pas faite et que c'est cette négligence qui est à l'origine de son dysfonctionnement, il semble important de pouvoir relier cette omission au fait que dans certains scénarios elle participe au fait que le seuil de pollution soit atteint.

Finalement, le quatrième défi concerne la transitivité de la causalité et plus exactement le fait qu'elle soit nécessaire pour évaluer la portée que peut avoir une action en reliant l'action à ses conséquences indirectes. En effet, l'alternative envisageable serait de relier directement les actions à leurs effets directs comme indirects, mais une telle modélisation souffrirait du problème de ramification mentionné dans la section 2.2. Dans l'exemple 3.1, une fois déterminé quelle est la cause du fait que le seuil de pollution ait été atteint, nous souhaitons pouvoir utiliser cette information pour déduire les causes du préjudice des habitants.

Cette section est divisée en trois parties. Dans la section 3.2.1 nous montrons en quoi les approches de causalité positive existantes ne sont pas satisfaisantes pour les besoins de l'éthique computationnelle. Cela est fait en deux étapes, chacune correspond à un aspect du premier défi. Dans la section 3.2.2 nous parlons des débats dans le domaine concernant la causalité négative. Encore une fois, cela est fait en deux parties, une première section traite le deuxième défi et la deuxième partie le troisième défi. Finalement, dans la section 3.2.3 nous évoquons les débats dans le domaine autour de la transitivité, quatrième et dernier défi.

3.2.1 Approches existantes pas tout à fait satisfaisantes pour une formalisation

Dans cette section nous montrons pourquoi les approches existantes ne sont pas adaptées à l'éthique computationnelle. Pour cela nous considérons que toute approche peut être décomposée en deux parties : la définition de la causalité et le formalisme utilisé pour modéliser cette définition. En l'occurrence, il est tout à fait possible de modéliser le test NESS en équations structurelles [BECKERS, 2021b] ou le but-for test dans le formalisme proposé par BOCHMAN [2018b]. Cette séparation nous permet de ne pas commettre l'erreur de confondre les problématiques qui relèvent de la définition de causalité avec celles qui relèvent du formalisme utilisé. Dans la section 3.2.1.1 nous abordons la question de la définition de causalité effective la plus appropriée. Dans la section 3.2.1.2 nous parlons des limites existantes reliées aux formalismes.

3.2.1.1 Le choix de la définition

Dans cette section nous montrons pourquoi le besoin d'une séparation claire entre causalité et responsabilité implique une préférence en éthique computationnelle pour les définitions des approches par inférence à celles des approches contrefactuelles.

Justifions d'abord l'idée qu'il existe un besoin de séparer clairement causalité et responsabilité en éthique computationnelle. Dans la section 3.1.2.2 nous avons vu que cette distinction est nécessaire en droit. Comme en droit, en éthique la causalité n'est qu'une étape pour arriver au résultat final. Si d'un côté la causalité est une étape pour déterminer s'il existe une responsabilité juridique, de l'autre c'est une étape pour déterminer le statut déontique d'une action. Selon la théorie morale, la causalité joue un rôle plus ou moins important. Comme pour les aspects normatifs en responsabilité juridique, le choix de la théorie morale fait en sorte que certaines causes sont pertinentes ou non dans l'évaluation éthique. En l'occurrence, pour l'utilitarisme de l'acte vu dans la section 1.3.1.1 toutes les conséquences d'une action doivent être prises en compte de la même façon, alors que pour la théorie du droit naturel vue dans la section 1.3.2 l'intention de l'agent et la proportionnalité entre le bien et le mal causé font que certaines conséquences sont plus ou moins influentes

dans l'évaluation de l'action. Chaque théorie morale défend donc des aspects normatifs qui lui sont propres, ces choix découlent des fondements de la théorie. Mélanger causalité et responsabilité revient à intégrer dans l'enquête causale des aspects normatifs propre à une vision particulière. Si cette enquête causale est utilisée dans la formalisation d'une théorie morale ne partageant pas cette vision, alors la formalisation obtenue ne correspondra pas à la théorie souhaitée. Formaliser l'utilitarisme de l'acte en appliquant une enquête causale qui ne considère que les conséquences d'une action qui sont intentionnelles aura pour résultat un processus qui ne sera certainement pas de l'utilitarisme de l'acte. Si nous voulons avoir une approche de causalité positive commune adaptée à l'éthique computationnelle, celle-ci doit nécessairement veiller à respecter la séparation entre causalité et responsabilité. Cette conclusion n'est pas restreinte à la causalité positive, elle s'applique à la causalité dans son ensemble.

Comme nous l'avons vu dans la section 3.1, nous avons le choix entre deux types de définitions : celles des approches contrefactuelles présentées dans la section 3.1.1.2 ou celles des approches par inférence présentées dans la section 3.1.1.3. Nous commençons par montrer pourquoi les définitions contrefactuelles ne sont pas satisfaisantes pour nos besoins en éthique computationnelle.

L'option naïve consiste à prendre le but-for test qui considère qu'il y a une équivalence entre causalité et dépendance causale. Autrement dit, cette approche se base uniquement sur la notion de nécessité en considérant qu'une cause doit absolument être nécessaire à la conséquence. Cette définition est simple et efficace dans de nombreux cas. La preuve en est, elle a longtemps été utilisée en droit et les quelques travaux en éthique computationnelle qui considèrent la causalité l'utilisent [LINDNER et BENTZEN, 2018; LINDNER et collab., 2017]. Toutefois, comme nous l'avons vu dans la section 3.1.1.2, cette définition ne gère aucun cas de surdétermination. S'agissant de cas importants que nous voulons pouvoir traiter en éthique computationnelle, cette définition ne correspond pas à nos besoins.

Nous nous tournons donc vers les définition de HALPERN [2016], aujourd'hui dominantes parmi les approches contrefactuelles. En effet, celles-ci intègrent l'idée de dépendance causale mais y ajoutent des éléments supplémentaires de façon à pouvoir traiter les cas de surdétermination. Comme exposé dans la section 3.1.1.2, un des éléments principaux rajoutés est l'interventionnisme par lequel sont reliées dépendance causale et causalité. Une intervention est une opération permettant de fixer arbitrairement la valeur d'un ensemble d'évènements dans le scénario en respectant un ensemble de conditions. Dans les définitions de Halpern ces conditions n'exigent pas de devoir respecter la cohérence des équations, ce qui introduit un aspect non factuel plutôt problématique pour nos besoins en éthique computationnelle. Qui plus est, cela semble même être problématique dans le domaine de la causalité effective puisque des approches contrefactuelles décident de se passer de l'interventionnisme [BECKERS, 2021b; BECKERS et VENNEKENS, 2018] et que HALPERN [2018] lui-même semble être gêné de devoir recourir à cet outil :

> if I fix BH [Billy hits] to zero here, I am sort of violating the way the world works. [...] I am contemplating counterfactuals are inconsistent with the equations but I seem to need to do that in order to get things to work out right. Believe me, we tried many other definitions.

Outre les éléments non factuels intégrés dans l'enquête causale, il semble que faire usage de l'interventionnisme implique d'intégrer des aspects normatifs. En effet, comme l'explique clairement BECKERS [2021b], les variations entre les différentes définitions du style

Halpern [HALL, 2007; HITCHCOCK, 2001, 2007; WESLAKE, 2015; WOODWARD, 2005] reposent principalement sur le choix de conditions que doivent respecter les interventions : « The divergence between these definitions is to be found in the condition P that should be satisfied ». L'interventionnisme, essentiel aux définitions de style Halpern, est incompatible avec le besoin d'une séparation claire entre une enquête causale factuelle et les aspects subjectifs propres à la responsabilité. Ces définitions réussissent à gérer les cas de surdétermination, mais au prix de l'aspect factuel de l'enquête causale. Elles ne sont donc pas adaptées à nos besoins.

Les définitions des approches contrefactuelles ne semblent pas convenir comme base pour une approche de causalité positive commune adaptée à l'éthique computationnelle. En plus des cas que nous venons de voir, des travaux récents dans le domaine en viennent même à remettre en question l'idée qu'il soit nécessaire qu'il existe une quelconque dépendance causale entre la cause et la conséquence et donc que la contrefactuelle doivent intervenir dans la définition de la causalité. ABRAMS [2022] considère que la dépendance causale est au plus une heuristique utile : « Perhaps but-for is merely an effective heuristic for detecting genuine causation, but genuine causation itself is not a matter of counterfactual dependence ». Wright est en désaccord avec l'idée que l'enquête causale demande la construction de mondes contrefactuels, comme le demande tout raisonnement reposant sur la dépendance causale. Si les deux avis précédents proviennent du droit, il est possible de trouver des avis similaire en informatique. BOCHMAN [2018b] considère que la contrefactuelle n'a pas sa place dans les définitions de causalité : « [...] contrary to the currently dominant opinions, counterfactuals (at least on their standard understanding) cannot serve as a ground neither for (causal) laws, nor even for actual causation ».

Ayant montré que les définitions contrefactuelles ne sont pas satisfaisantes pour nos besoins en éthique computationnelle, il nous reste à évaluer si les définitions par inférence le sont. Plus précisément, nous nous intéressons au test NESS qui semble s'imposer comme la définition dominante.
D'un point de vue purement causal, les approches par inférence sont aussi satisfaisantes que les approches contrefactuelles. Qui plus est, comme l'indique BAUMGARTNER [2013], elles ne font pas appel à des raisonnements controversés comme l'interventionnisme : « I am going to argue that [a regularity/inference theorie] performs at least as well as modern counterfactual accounts. Furthemore, contrary to the latter, a regularity theory achieves its goal by implementing uncontroversial and straightforward conceptual and technical ressources ».
Maintenant, du point de vue de l'éthique computationnelle, le test NESS semble répondre à nos attentes. D'un côté, cette définition permet de gérer les cas de surdétermination [BAUMGARTNER, 2013; WRIGHT, 1985, 1988, 2011] ; de l'autre, elle le fait sans renoncer à la factualité nécessaire dans l'enquête causale. Cela est possible principalement grâce au fait qu'elle place la suffisance au premier plan et qu'elle y subordonne la nécessité. Une cause ne doit pas être nécessaire à la conséquence, il suffit qu'elle soit nécessaire à la suffisance d'un ensemble de causes pour produire la conséquence. Le test NESS arrive donc très bien à gérer les cas de surdétermination où aucune cause n'est individuellement nécessaire à la conséquence et cela sans passer par des raisonnements hypothétiques. Qui plus est, comme l'explique très bien WRIGHT [2011], le test NESS gère également les cas où aucune cause n'est individuellement suffisante à la conséquence, des cas omniprésents lorsqu'il est question de décisions d'êtres humains comme en droit ou en éthique :

> The NESS account's ability to identify conditions that were neither strongly necessary nor independently strongly sufficient as causes applies to and is especially useful for accounting for human decisions and actions, which often are based on multiple reasons, none of which may have been – or can be proven to have been – strongly necessary or independently strongly sufficient for the particular decision or action.

Le test NESS a la capacité à gérer les cas de surdétermination, et de causalité positive en générale, de façon factuelle. Par conséquent, elle assure une séparation entre causalité et responsabilité qui font de cette définition celle qui semble convenir le mieux comme base pour une approche de causalité positive commune adaptée à l'éthique computationnelle.

3.2.1.2 Le choix du formalisme

Dans cette section nous montrons que les formalismes des approches de causalité existantes ne sont pas satisfaisants pour une approche de causalité positive commune adaptée à l'éthique computationnelle. Cela se doit principalement au fait qu'ils ne sont pas aptes à représenter les subtilités nécessaires en éthique computationnelle, et qu'ils peuvent être à l'origine de confusions dans les cas les plus complexes de la causalité.

Outre le fait que les différentes variantes de l'exemple 3.1 permettent d'illustrer différents types de surdétermination, cet exemple peut également mettre en évidence la complexité du raisonnement éthique. En effet, de nombreux facteurs rendent le problème plus complexe. Tout d'abord, la valeur produite par l'activité polluante est différente. D'une part, la production d'un bien dit de confort, d'autre part, la production d'un médicament essentiel à la survie des individus. Nous pourrions ajouter que la fabrication d'enceintes fait vivre la région en donnant du travail à une grande partie des habitants du village, alors que l'usine de médicaments a complètement automatisé sa chaîne de production. Nous pourrions également ajouter que les rejets d'une usine sont autorisés, alors que ceux de l'autre ne le sont pas. Au final, tous ces facteurs peuvent avoir une influence significative sur le statut déontique des actions à évaluer. L'éthique ne peut se passer des subtilités des problèmes. Cependant, toute cette richesse ne peut être explorée que si l'on dispose d'une expressivité suffisante pour la représenter et d'un mécanisme permettant d'établir des relations de causalité suffisamment complexes.
Pour faire une analogie avec la métrologie, notre capacité à décrire la connaissance que nous avons du contexte définit la résolution, c'est-à-dire la plus petite variation de grandeur qu'un instrument peut mesurer. Il est inutile d'avoir un modèle causal très complexe et de multiples théories éthiques avec beaucoup de nuances si nous n'avons pas les moyens de représenter les petites variations qui font toute la différence dans les problèmes éthiques.

L'approche dominante pour représenter les problèmes de causalité effective est d'utiliser le cadre des équations structurelles proposé par PEARL et NEUBERG [2000]. Ce cadre n'est pas adapté à l'éthique computationnelle. En effet, cette approche a ses limites car elle ignore certaines nuances importantes pour le raisonnement éthique. En l'occurrence, dans ce cadre aucune distinction n'est faite entre états du monde et évènements. C'est le cas également des formalismes utilisés par BAUMGARTNER [2013]; BECKERS [2021b]; BOCHMAN [2018b]; WRIGHT [2011]. Notamment dans les SEF, toutes les variables sont considérées comme des évènements. Dans l'exemple 3.5 sur Suzy et Billy, les variables sont « Suzy/Billy lance », « Suzy/Billy touche » et « bouteille se brise ». Par conséquent, lorsque la causalité ef-

fective est abordée, la plupart des exemples concernent des relations causales entre deux occurrences d'évènements. En éthique les agents et leurs actions sont au centre de la réflexion. Cependant, il n'est pas possible de réduire le monde à cela. La distinction entre les états du monde et les évènements semble profondément ancrée dans le raisonnement éthique. Évaluer un état du monde revient à lui attribuer une valeur, alors qu'évaluer une action revient à déterminer son statut déontique. Sans distinction entre états du monde et évènements ni les théories axées sur la valeur présentées dans la section 1.3, ni les théories axées sur la vertu présentées dans la section 1.4 n'existeraient.

Il semble qu'une partie de la confusion qui subsiste dans le domaine de la causalité provienne de l'ambiguïté due au manque d'expressivité des formalismes utilisés. D'un côté, il apparaît que la causalité comme l'éthique sont très sensibles à la représentation qui est faite des problèmes. Le simple fait que l'approche dominante dans le domaine ait besoin de modifier la représentation naïve du problème de Suzy et Billy, probablement l'exemple le plus utilisé, pour être en mesure de le résoudre en est une preuve. BAUMGARTNER [2013] pousse l'analyse plus loin en montrant plusieurs exemples qu'il représente de différentes façons afin de montrer l'importance des variables qui sont choisies pour représenter le problème. D'un autre côté, il n'est pas rare de trouver des assertions montrant la difficulté de trouver la représentation adéquate d'un problème [HITCHCOCK, 2007] : « What constitutes an appropriate model is a tricky affair, more a matter of art than science ». Difficulté due à l'absence de règles indiquant clairement la procédure à suivre, mais aussi au manque d'expressivité du formalisme. En effet, en causalité effective il est souvent question de problèmes dynamiques [BATUSOV et SOUTCHANSKI, 2018; HALPERN, 2000] que les approches classiques ne sont pas en mesure de représenter convenablement. HOPKINS et PEARL [2007] et BOCHMAN [2018a] sont conscients de ces limites lorsqu'ils écrivent :

> Structural causal models are excellent tools for many types of causality-related questions. Nevertheless, their limited expressivity render them less than ideal for some of the more delicate causal queries, like actual causation. These queries require a language that is suited for dealing with complex, dynamically changing situations. [HOPKINS et PEARL, 2007]
>
> Despite the established connection between the two causal formalisms [causal calculus and causal models], there are obvious differences in the respective objectives of these theories, as well as in the required expressive means. Thus, the restrictions appearing in our definition of a causal Pearl theory make such theories completely inadequate for describing dynamic domains in reasoning about action and change. [BOCHMAN, 2018a]

L'utilisation de certains formalismes peut expliquer une partie de la confusion qui subsiste dans le domaine de la causalité autour des cas de surdétermination. En plus de ne pas faire de distinction entre états du monde et évènements, les formalismes cités précédemment n'incluent pas la temporalité dans le raisonnement. Pourtant, comme nous l'avons vu dans la section 3.1.2.1, lorsqu'il est question de décrire les cas de surdétermination, il est commun de trouver des notions appartenant au champ sémantique du temps. En fin de compte, certains désaccords ne porteraient pas réellement sur la définition de la causalité, mais sur la manière dont ils sont formalisés.
Comme montré dans la section 3.1.2.1, il existe différents types de cas de surdétermination et selon le cas de surdétermination dont il est question, les relations causales attendues ne sont pas les mêmes. Deux types de désaccords peuvent donc être identifiés, soit il existe un

désaccord sur les relations causales attendues étant donné un type de cas de surdétermination donné, soit le désaccord repose sur le type de cas de surdétermination qui est traité. Dans le premier cas nous dirons qu'il s'agit d'un *désaccord causal fondamental*, la divergence se situe dans la définition même de ce qu'est la causalité. Dans le deuxième cas nous dirons qu'il s'agit d'un *désaccord causal non fondamental*, la divergence est sur le problème qui est traité et la façon dont il est formalisé. Si chaque proposition traite un type de cas différent, il n'est pas surprenant qu'elles puissent arriver à des conclusions différentes. En revanche, cela ne veut pas dire que ces propositions ont une conception différente de la causalité. Ce deuxième cas de figure où le débat est plus sur la forme que sur le fond est difficilement résoluble sans un formalisme qui permette de distinguer clairement les divergences.

Les langages d'action et du changement semblent être une alternative prometteuse pour résoudre les problématiques qui subsistent dans le domaine de la causalité. En l'occurrence, le besoin d'avoir une distinction entre états du monde et évènements apparaît également dans le domaine de la causalité. En se basant sur plusieurs travaux influents dans le domaine [Collins et collab., 2004; Fumerton et Kress, 2001; Hall, 2004; Strevens, 2007; Thomson, 2008], Wright [2011] reconnaît ce besoin : « More generally, a proper theory of events almost certainly must count as such things that we ordinarily would classify as states or standing or background conditions ». Lorsque Lewis [1973] choisi de s'intéresser uniquement aux évènements, il précise qu'il s'agit d'un choix de simplification et que d'autres choses peuvent-être reliées par une relation causale : « events are not the only things that can cause or be caused ». Les SEF et autres formalismes très abstraits ont permis d'isoler la problématique de la causalité de celle de la représentation de l'action et du changement. Comme nous l'avons vu dans le chapitre 2, cette problématique n'est pas simple à résoudre non plus. Cette séparation a sûrement été bénéfique à l'identification des problèmes qui étaient propres à chaque domaine. Toutefois, aussi bien le temps que la distinction entre états du monde et évènements semblent être des éléments enracinés dans la façon dont nous raisonnons sur l'évolution du monde et semblent donc utiles pour discuter des subtilités de la causalité. McCarthy et Hayes [1969] soulignent « practical systems require epistemologically adequate systems in which those facts [commonsense concepts] which are actually ascertainable can be expressed ». Nous pensons qu'il ne peut être que bénéfique de sortir de la simplification des SEF et autres formalismes classiques en causalité et voir comment cela peut aider à résoudre certaines problématiques encore existantes en causalité. La publication ces dernière années de travaux en causalité qui reposent sur des formalismes comme le Calcul des Situations [Batusov et Soutchanski, 2018] ou des langages de description d'action [Berreby et collab., 2018; Bourgne et collab., 2021; LeBlanc et collab., 2019] semble indiquer que cette vision est partagée. Toutefois, il s'agit là de premières étapes qui ne peuvent pas gérer la plupart des cas de surdétermination. En effet, aucune de ces approches n'intègre à la fois les préconditions disjonctives et la cooccurrence d'évènements, deux éléments essentiels pour pouvoir représenter ces problèmes comme nous l'avons vu dans la section 3.1.2.1. En terme d'expressivité nécessaire, Batusov et Soutchanski [2018] défendent l'idée que pour traiter les problèmes en causalité effective, le formalisme choisi doit inclure également la notion de temps et les évènements naturels :

> It is clear that a broader definition of actual cause requires more expressive action theories that can model not only sequences of actions, but can also include explicit time and concurrent actions. Only after that one can try to analyze some

> of the popular examples of actual causation formulated in philosophical literature. Some of those examples sound deceptively simple, but faithful modelling of them requires time, concurrency and natural actions.

Avant de clore cette section, il est pertinent de préciser qu'adopter les représentations présentées dans le chapitre 2 pour la causalité présente des bénéfices d'implémentation. Une fois la description d'action décrite, il est possible de raisonner sur tous les scénarios imaginables avec cette description. Par exemple, si nous avons la description d'action nécessaire à l'exemple 3.1, celle-ci permet de raisonner aussi bien sur le cas simple de l'exemple 3.2, le cas préemptif de l'exemple 3.3 et le cas duplicatif de l'exemple 3.4. Mieux encore, des éléments d'une description d'action peuvent être utilisés pour d'autres problèmes. Si nous décrivons l'évènement fermer la porte de façon très générale pour un problème donné, il peut tout à fait être utilisé dans un problème différent où cet évènement intervient. Avec suffisamment de rigueur et d'énergie, il est envisageable de construire une description d'action qui permettrait de traiter de nombreux problèmes.
Les SEF et autres formalismes classiques en causalité ne permettent pas cela. La représentation classique de Suzy et Billy avec uniquement trois variables correspondrait au cas duplicatif de l'exemple 3.4. Il n'y est pas possible sans reformulation d'intégrer le fait que Suzy lance avant ou simplement plus fort, qui en fait un cas préemptif comme dans l'exemple 3.3. Comme nous l'avons vu précédemment dans ce chapitre, pour gérer cela HALPERN [2016] se voit obligé d'ajouter des variables au problème. Les SEF ne font pas totalement abstraction de la temporalité, celle-ci n'est pas explicite mais elle est implicite dans les variables choisies pour représenter le problème et la façon dont elles sont reliées. Un modèle correspond donc à une version d'un problème, s'intéresser à une variante demande de reformuler le problème, ce qui complique considérablement l'implémentation.

3.2.2 La négation dans la relation causale

Dans cette section nous parlons des débats dans le domaine concernant la négation dans la relation causale. Au début du chapitre nous avons défini une relation causale comme une relation binaire qui lie une cause à une conséquence. Lorsque nous parlons de négation dans la relation causale, deux cas sont envisageables. Le premier est celui où la négation est dans la conséquence. Dans l'exemple 3.1, cela revient à chercher les causes au fait que le seuil de pollution n'ait pas été atteint. Le deuxième cas est le cas où la négation est dans la cause. Dans l'exemple 3.1, cela revient à chercher les conséquences du fait que l'agent en charge de faire la maintenance de la station d'épuration des eaux usées de l'usine de médicaments ne l'ait pas faite.
Cette section est donc décomposée en deux parties. Dans la section 3.2.2.1 nous traitons le deuxième défi correspondant au cas où la négation est dans la conséquence. Dans la section 3.2.2.2 nous discutons du troisième défi correspondant au cas où la négation est dans la cause.

3.2.2.1 Conséquence négative, ou empêcher

Dans cette section nous allons parler du cas où la négation est dans la conséquence. Ce cas correspond à ce que nous avons appelé le deuxième défi, arriver à déterminer les causes pour lesquelles une condition ne s'est pas produite. La question de ces causes se pose en éthique computationnelle, nous voulons être en mesure de pouvoir raisonner sur

le fait qu'un agent ait agi pour empêcher le mal ou le bien de se produire. Dans l'exemple 3.1, si nous voulons évaluer l'action d'un agent qui répare la station d'épuration des eaux usées de l'usine de médicaments il semble important de pouvoir relier cette action au fait que dans certains scénarios elle empêche le seuil de pollution d'être atteint.

Le raisonnement pour déterminer les causes pour lesquelles une condition s'est produite est différent de celui pour déterminer les causes pour lesquelles une condition ne s'est pas produite. Cette différence pourrait indiquer qu'il s'agit d'une notion différente à celle de la causalité positive, ce qui mérite d'être étudié. Le premier raisonnement consiste à étudier un processus causal qui a abouti, toute la problématique est de savoir quels éléments ont permis cet aboutissement. Le second repose sur l'étude d'un processus causal qui n'a pas abouti. Pour qu'un tel échec ait lieu, il faut qu'au moins une condition nécessaire dans chaque ensemble de conditions suffisantes ne soit pas vérifiée. Par exemple, si dans l'exemple 3.8 le voyageur dans le désert meurt soit par empoisonnement, soit par déshydratation, chacun des cas étant suffisants pour causer la mort, il faut que l'empoisonnement n'ait pas lieu et que le voyageur ne se retrouve pas sans eau. La problématique est de savoir quels éléments sont à l'origine de l'échec de chacun de ces cas. Ce premier niveau de raisonnement plus élémentaire est ce que WRIGHT [2011] appelle « simple prevention », il intervient lorsque l'action que nous souhaitons évaluer empêche quelque chose de se produire, il est la cause d'une absence dans le monde. Par exemple, si le garde du corps du voyageur substitue la fiole de poison de l'assassin A par un liquide inoffensif nous avons envie de dire qu'il empêche l'empoisonnement. De la même façon, si le garde du corps du voyageur a pris avec lui des réserves d'eau, ce qui fait en plus que l'assassin B abandonne sa mission et ne vide pas la gourde, nous avons envie de dire qu'il empêche la déshydratation. Mais ce premier niveau ne semble pas être suffisant lorsque nous pensons à la relation empêcher, quelque part ce qui nous intéresse réellement dans ce cas est de déterminer les raisons pour lesquelles le processus causal aboutissant à la mort du voyageur a échoué. WRIGHT [2011] parle alors de « double prevention », un raisonnement qui intervient lorsque l'action que nous souhaitons évaluer interromps un processus causal qui, s'il n'avait pas été interrompu, aurait eu des conséquences sur le monde. Dans ce cas, nous avons envie de dire que l'action que nous évaluons est la cause que ces conséquences n'aient pas eu lieu. En l'occurrence, dans l'extension de l'exemple 3.8 développée ci-dessus, nous avons envie de relier l'action du garde du corps avec le fait que le voyageur ne meurt pas lors de son voyage. Cette intuition largement partagée montre bien que la notion d'empêcher mérite d'être étudiée.

Cette notion a été traitée dans de nombreux travaux [BAUMGARTNER, 2013; BERREBY et collab., 2018; COLLINS, 2000; DOWE, 2001; HITCHCOCK, 2007; MOORE, 2019; WRIGHT, 2011]. Un simple aperçu de ces travaux permet de comprendre que cette notion est complexe à traiter. Comme pour la causalité positive, les cas de surdétermination suscitent des débats dans le domaine. En l'occurrence, dans l'extension de l'exemple 3.8 que nous avons proposé, il n'est finalement pas si évident de savoir quelle action relier avec le fait que le voyageur ne meurt pas lors de son voyage. Certes, avoir remplacé la fiole a fait que l'eau dans la gourde n'était pas empoisonnée, mais ayant des réserves d'eau rien ne dit que dans le cas où la gourde aurait été empoisonnée le voyageur serait mort. De même, avoir pris des réserves d'eau a fait que le voyageur ne puisse pas mourir déshydraté, mais ayant remplacé la fiole empoisonnée avec une fiole avec un produit inoffensif, rien ne dit que dans le cas où son garde du corps n'aurait pas pris les réserves d'eau le voyageur serait mort. Nous

nous trouvons ici face à un cas de surdétermination classique où aucune des actions du garde du corps n'est nécessaire individuellement à empêcher la mort du voyageur. WRIGHT [2011] suggère d'appliquer le test NESS comme pour la causalité positive, ce qui revient à dire qu'aussi bien avoir remplacé la fiole, qu'avoir pris des réserves d'eau ont empêché la mort du voyageur.

La notion d'empêcher semble faire intervenir plus fortement la notion de nécessité que dans la causalité positive. Cette intuition explique que dans certains cas, l'application du test NESS ne semble pas satisfaisante pour les résoudre. Nous allons illustrer cela à l'aide d'un exemple utilisé par COLLINS [2000].

Exemple 3.11 [la balle et la fenêtre]**.** *Une balle se dirige dans la direction de* B *qui s'apprête à l'attraper. Cependant,* A *étant plus proche et réagissant plus rapidement que* B*, il fait un bond sur sa gauche et attrape la balle juste devant l'endroit où la main de* B *attendait pour intercepter la balle. La balle allait droit vers une fenêtre, si elle n'avait pas été interceptée elle l'aurait brisée.*

Comme tout cas de surdétermination, aucune des deux actions n'était nécessaire à la conséquence, ici empêcher la fenêtre d'être brisée. En effet, si A n'avait pas attrapé la balle, B l'aurait fait et vice versa. Le fait que B s'apprête à attraper la balle a quelque part rendu la contribution de A superflue, la balle n'aurait pas pu atteindre la fenêtre. D'après cette vision, il est possible de considérer que A n'a pas réellement empêché la fenêtre d'être brisée mais uniquement B. Si nous adoptons un point de vue purement factuel comme pour la causalité positive, nous avons plutôt envie de dire que c'est uniquement A qui a empêché la fenêtre d'être brisée, c'est son processus causal qui a abouti, celui de B peut être vu comme préempté. Étudions une variante de ce problème présentée par MCDERMOTT [1995] :

Exemple 3.12 [la balle, la fenêtre et le mur]**.** *Une balle se dirige dans la direction de* A *qui fait un bond sur sa gauche et attrape la balle. Derrière lui un mur de briques solides et plus loin une fenêtre.*

La situation dans cette variante semble être la même, A attrape la balle alors que derrière lui quelque chose l'aurait de toute façon interceptée. Dans l'exemple 3.11 il s'agissait de B, dans l'exemple 3.12, B a été remplacé par un mur. Malgré cette symétrie, le point de vue purement factuel consistant à dire que A seul a empêché la fenêtre d'être brisée car c'est uniquement son processus causal qui a abouti est plus difficile à entendre. Cela semble venir du fait qu'il est plus compliqué d'imaginer le monde hypothétique où le mur n'est pas là que le monde hypothétique où B n'essaye pas d'intercepter le balle. Mais que veut dire qu'il est « plus compliqué d'imaginer » un monde hypothétique qu'un autre ? Cette mesure semble difficilement pouvoir être complètement indépendante d'aspects normatifs propres à un individu ou un groupe d'individus.
Cet exemple est loin d'être le seul à poser ces problèmes. Imaginons le cas où l'agent commanditaire du meurtre du voyageur paye l'assassin A pour l'empoisonner, puis qu'il paye l'assassin B pour tuer l'assassin A. Même si factuellement il a empêché que l'autre assassin puisse empoisonner la gourde du voyageur, il semble contre intuitif de dire qu'il est la cause que le voyageur soit encore en vie. Dans la variante où l'assassin A aurait aussi été embauché par un autre agent, en plus de celui qui a payé l'assassin A et B, il devient un peu plus acceptable de faire ce lien. Celui-ci l'est encore plus si l'agent qui embauche l'assassin A et l'assassin B sont distincts. HITCHCOCK [2007] propose un autre de ces problèmes. Il s'agit du

cas où un garde du corps verse un antidote inoffensif pour la santé humaine dans le café de l'agent qu'il protège. Puis, Buddy verse un poison létal dans le café, mais les effets de celui-ci se voient annulés par la présence de l'antidote. On considère dans cet exemple, pour des raisons qui lui appartiennent, que Buddy n'aurait pas empoisonné le café si le garde du corps n'avait pas en premier versé l'antidote. L'agent boit son café et ne meurt pas. À nouveau, nous sommes dans une situation où factuellement avoir versé l'antidote a empêché le poison de faire effet et l'agent à l'origine de cette action est également à l'origine de la menace. Mais ici la situation est plus complexe car ce n'est pas juste le même agent qui est à l'origine du danger et de son absence, c'est la même action. En effet, l'exemple précise que c'est le fait de verser l'antidote qui fait que Buddy verse le poison. Pour Hitchcock nous ne pouvons pas dire que le garde du corps ait empêché la mort de l'agent qu'il protège. BAUMGARTNER [2013] propose des structures de problèmes qu'il montre équivalentes, mais qui pourtant suscitent une intuition causale différente.

Ces quelques exemples montrent que, contrairement au traitement de la causalité positive, une approche entièrement factuelle pour le traitement de ce cas de négation dans la causalité n'est pas satisfaisante. Il semblerait qu'il puisse exister une différence entre ces formes de causalité, différence qu'il est important de clarifier afin de mieux comprendre la causalité et pouvoir proposer une approche convenable [HITCHCOCK, 2007] :

> [...] there is some difference of opinion regarding whether cases of prevention and omission are cases of genuine causation. What we should demand of a theory of causation is not so much that it settle this disagreement in one way or the other, but that it identify the respects in which cases of prevention and omission both resemble and differ from pragmatic cases of causation.

En SEF l'occurrence d'un évènement et sa non occurrence ne sont pas sémantiquement différents. En fin de compte il s'agit juste d'une variable qui prend une valeur différente, mais la non occurrence peut très bien être représentée par la variable prenant comme valeur 0, mais aussi 2, ou 10000. Ce problème est soulevé par BAUMGARTNER [2013] :

> Controversial questions as to the ontological makeup of the instances of factors or as to what instantiates absences are deliberately ignored in the present context. To avoid these questions the structural equations framework has a very handy terminology on offer : both occurrences of and non-occurrences of events are simply understood as random variables taking one of their respective values. Thus, alternatively, factors can be seen as binary variables that take value 1 whenever a token of the corresponding type occurs and the value 0 whenever no such token occurs.

Il en est de même dans les formalismes classiques en causalité proches de la logique BAUMGARTNER [2013]; BOCHMAN [2018a], sauf que les valeurs possibles sont limitées à 0 ou 1. Il n'y a pas une distinction formelle entre un fluent présent, une occurrence d'un évènement, ou une non occurrence d'un évènement, comme c'est le cas dans les formalismes avec pour sémantique un STEE. Proposer une définition de causalité positive dans ces formalismes revient finalement à proposer une définition pour tous les cas de causalité.
L'avantage des langages de description d'action comme ceux que nous avons présenté dans le chapitre 2 est qu'il est possible d'isoler les différents cas de causalité pour les étudier séparément. En effet, il est possible de faire une distinction forte entre ce que veut dire causer la véracité d'un fluent, causer l'occurrence d'un évènement et causer la non occurrence

d'un évènement. Encore une raison de penser que les langages d'action et du changement semblent être une alternative prometteuse pour résoudre les problématiques qui subsistent dans le domaine de la causalité.

3.2.2.2 Cause négative, ou l'omission

Dans cette section nous allons parler du cas où la négation est dans la cause. Ce cas correspond à ce que nous avons appelé le troisième défi, arriver à déterminer les conséquences d'une non occurrence d'un évènement. Ces conséquences ont une importance toute particulière en éthique computationnelle, nous voulons être en mesure de pouvoir raisonner sur le fait qu'un agent aurait pu agir pour empêcher le mal ou causer le bien, mais ne l'a pas fait, il a omis de la faire. Dans l'exemple 3.1, si l'agent en charge de faire la maintenance de la station d'épuration des eaux usées de l'usine de médicaments ne l'a pas faite et que c'est cette négligence qui est à l'origine de son dysfonctionnement, il semble important de pouvoir relier cette omission au fait que dans certains scénarios elle participe au fait que le seuil de pollution soit atteint.

Cette question ne se pose pas que dans le cadre de l'éthique computationnelle, elle a été soulevée et traitée de différentes façons dans le domaine de la causalité [HALL et PAUL, 2003] : « [. . .] there are other examples that a comprehensive treatment would need to explore—most especially, examples that highlight issues involving the causal status of omissions ». L'exemple suivant est un de ceux le plus souvent utilisés pour parler d'omission.

Exemple 3.13 [le jardinier]**.** *Imaginons une copropriété qui souscrit un contrat avec un jardinier pour qu'il s'occupe de l'entretien de leurs espaces verts, ce qui inclus leur arrosage. Une magnifique vigne orne un espace proche de l'entrée principale par laquelle les habitants passent tous les jours. Durant plusieurs semaines d'affilée, le jardinier omet d'arroser la vigne qui finit par mourir.*

À première vue, la plupart des approches arrivent aisément à traiter ce problème avec la même définition que pour la causalité positive et donc sans distinction entre une action et une omission. Du côté des approches contrefactuelles le but-for test suffit pour trouver que l'omission du jardinier est une cause de la mort de la vigne. En effet, si le jardinier n'avait pas omis de l'arroser, celle-ci ne serait pas morte a priori. Il y a bien une dépendance causale. Le test NESS permet également de résoudre ce problème. Nous pouvons imaginer qu'un des ensembles suffisant pour produire la conséquence inclut un processus biochimique menant à la mort de la plante lorsqu'elle n'a pas assez d'eau et l'absence du jardinier qui arrose la plante, toutes deux des conditions nécessaires à cet ensemble suffisant. L'omission du jardinier est une condition nécessaire à la suffisance d'un ensemble, et donc une cause selon le test NESS.

En réalité, ces solutions ne sont pas tout à fait satisfaisantes et soulèvent des problématiques qui font penser que l'omission mérite d'être étudiée plus en détail. Pour l'exemple 3.13, l'argument qui est souvent utilisé contre ces solutions consiste à se demander pourquoi l'omission du jardinier serait une cause et non pas celle de tous les habitants de la copropriété. Après tout, les habitants passent tous les jours devant la vigne et ont parfaitement les moyens de l'arroser. À part des raisonnements sur la responsabilité s'appuyant sur le fait que c'était le devoir du jardinier, ou que les habitants ne savaient pas que la vigne n'était pas arrosée, aucun argument purement causal n'explique pourquoi leur omission serait ignorée.

Dans le cas où elle ne l'est pas, nous nous retrouvons dans un cas de surdétermination et donc aucune des omissions n'est nécessaire à la conséquence, ici la mort de la vigne. D'un côté, le but-for test indique qu'aucune omission n'est une cause, issue qui paraît inacceptable. De l'autre, le test NESS dit qu'aussi bien l'omission du jardinier que celle des habitants sont des causes. Cette deuxième issue paraît déjà plus acceptable. Toutefois, cette position s'avère être une pente glissante. Si nous acceptons que toutes les omissions des habitants doivent être prises en compte, qu'est-ce qui nous dit que nous devons nous arrêter là ? Nous pourrions aussi intégrer les omissions des riverains qui sont passés devant l'immeuble, et aller jusqu'à intégrer l'omission de tous les agents qui avaient la capacité de se rendre là où se trouve la copropriété dans le laps de temps où arroser la vigne l'aurait sauvée. Après tout, à part des raisonnements sur la responsabilité s'appuyant sur le fait que tous ces agents ne savaient pas que la vigne était en danger, ou qu'ils n'ont aucun devoir envers les affaires de cette copropriété, aucun argument purement causal n'explique pourquoi leur omission serait ignorée.

L'omission mérite d'être étudiée indépendamment de la causalité positive car une approche entièrement factuelle pour le traitement de ce cas de négation dans la causalité n'est pas satisfaisante. Il semblerait qu'il puisse exister une différence entre ces formes de causalité, différence qu'il est important de clarifier afin de mieux comprendre la causalité et pouvoir proposer une approche convenable. En effet, le problème posé par l'exemple 3.13 n'est pas un cas isolé. Imaginons un médecin dans un hôpital qui omet de donner ses médicaments à un patient. L'absence de traitement entraîne des complications graves avec des séquelles pour le patient. Il se trouve que 100 autres médecins travaillent dans l'hôpital. Si nous considérons l'approche factuelle utilisée pour la causalité positive, il n'existe aucune raison d'ignorer l'omission de ces autres médecins et de tout le personnel en capacité de délivrer les médicaments.
Dans tous ces problèmes le raisonnement est que si un agent échoue à interrompre un processus, alors il devient une cause des conséquences de ce processus. Il s'agit là d'une des principales questions autour de l'omission. En reprenant les termes utilisés dans le domaine, il s'agit de déterminer dans quelles circonstances « fail to prevent » équivaut à causer. Si A décide de frapper B et que C intervient et bloque le coup, est-ce correct de dire que si C n'était pas intervenu alors il serait une cause du fait que B ait été frappé ? Cette formulation montre que les deux cas de négations dans la causalité sont étroitement liés. Éclaircir les liens entre ces cas ne peut être que bénéfique pour mieux comprendre le domaine de la causalité effective.

3.2.3 La transitivité

Dans cette section nous allons parler de la transitivité de la causalité. Ce cas correspond à ce que nous avons appelé le quatrième défi, arriver à évaluer la portée que peut avoir une action en reliant l'action à ses conséquences indirectes. Dans l'exemple 3.1, une fois déterminé quelle est la cause du fait que le seuil de pollution ait été atteint, nous souhaitons pouvoir utiliser cette information pour déduire les causes du préjudice des habitants.

Savoir si la causalité est transitive ou pas n'est pas une simple affaire. Cette question suscite de nombreux débats dans la communauté : « The debate about the transitivity of causation is not easily settled » [MENZIES et BEEBEE, 2020].
D'un côté, la nécessité d'avoir une relation causale transitive semble indispensable. En ef-

fet, cela permet de construire des chemins causaux qui paraissent indispensables au moins pour deux aspects. Le premier est que pour décrire les cas de surdétermination les termes utilisés renvoient à l'existence de chemins causaux, comme nous l'avons vu dans la section 3.1.2.1. Comprendre ces cas semble nécessairement passer par un raisonnement sur ces chemins. Deuxièmement, le monde dans lequel nous sommes pouvant être vu comme un cadre multi-agent où les évènements peuvent avoir des conséquences indirectes, il est nécessaire pour toute question d'imputabilité d'avoir un moyen de relier les actions à leurs conséquences indirectes. Lorsqu'un meurtre est commis par un tueur à gage, le processus judiciaire ne se limite pas à juger le tueur à gage, le commanditaire est également recherché même si ce n'est pas lui qui a commis le meurtre directement. La personne qui a vendu l'arme avec laquelle le meurtre a été commis peut également être recherchée. Sans transitivité, une alternative envisageable serait de relier directement les évènements à leurs conséquences directes comme indirectes, mais une telle modélisation souffrirait du problème de ramification mentionné dans la section 2.2.

D'un autre côté, il existe des situations où avoir une relation transitive peut sembler problématique. Prenons le cas proposé par BERREBY et collab. [2018].

Exemple 3.14 [cambriolage]**.** *Imaginons qu'un agent passe devant une maison dont la porte a été laissée grande ouverte par son unique habitant avant de partir travailler. Cet agent qui par son inexpérience aurait été incapable de rentrer dans la maison si la porte n'avait pas été ouverte, décide de la cambrioler.*

Dans cet exemple, le fait que la porte ait été ouverte a factuellement joué un rôle dans le cambriolage de la maison. Toutefois, il n'est pas tout à fait intuitif de dire que l'habitant de la maison est une cause du cambriolage. MOORE [2019] propose un problème similaire mais cette fois-ci dans un cas réel :

Exemple 3.15 [Regina v. Blaue]**.** *Un individu poignarde un autre. En raison de ses convictions religieuses, la victime refuse un traitement médical tout en sachant qu'un tel refus la tuera. La victime meurt.*

Comme précédemment, le geste de l'individu jugé a factuellement joué un rôle dans la mort de la victime. Malgré cela, il n'est pas tout à fait intuitif de dire qu'il est la cause de sa mort. La confusion entre causalité et responsabilité peut expliquer pourquoi ces cas semblent contredire le besoin de transitivité. Dans l'exemple 3.14, le fait que la porte ait été ouverte ne justifie pas l'acte du cambrioleur. En confondant causalité et responsabilité nous imaginons qu'établir cette relation causale voudrait dire que le cambriolage est quelque part la faute de l'habitant de la maison. Comme nous l'avons vu dans la section 3.1.2.2, ce n'est pas le cas. Il ne s'agit ici que de l'enquête causale, une des étapes pour déterminer la responsabilité. Dans l'exemple 3.15 la même confusion est à l'œuvre. En effet, prendre en compte les connaissances de l'agent semble influencer l'intuition que nous avons sur le problème. S'il est déterminé que l'individu jugé connaissait les convictions religieuses de la victime et donc savait qu'elle refuserait tout traitement, alors il semble plus facile de dire qu'il a causé sa mort. S'il est déterminé qu'il ignorait cela et qu'en plus la blessure infligée ne représentait aucun danger sous condition d'accepter quelques points de suture, alors c'est l'inverse.
Notez pour terminer que dans ces exemples où la transitivité est remise en question, il est souvent question de négation dans la causalité. Dans l'exemple 3.14 laisser la porte ouverte est une omission, il aurait pu empêcher le cambriolage en fermant la porte mais ne l'a pas

fait. Est-ce que cela veut dire qu'il est une cause? Dans l'exemple 3.15 nous retrouvons le même cas sauf que la temporalité est inversée, la possibilité d'empêcher la conséquence n'est pas avant l'acte criminel mais après. La victime aurait pu empêcher sa mort en acceptant le traitement mais ne l'a pas fait. La problématique de la transitivité est souvent liée aux cas de négation dans la causalité.

3.3 Conclusion

Dans ce chapitre nous avons donné un aperçu de ce qu'est la causalité effective, notion qui joue un rôle essentiel dans notre compréhension du monde. La causalité effective étant sujet de recherche en philosophie, psychologie, droit, mathématiques et informatique entre autres, l'aperçu que nous avons donné se devait d'être pluridisciplinaire.

Dans un premier temps, nous avons présenté les grandes lignes l'histoire de la causalité contemporaine depuis HUME [1748] jusqu'à nos jours. Nous avons vu que c'est sur la conception de Hume de la causalité que reposent les principales approches qui existent aujourd'hui. Premièrement, nous nous sommes intéressés au développement des approches par régularité, dominantes depuis Hume jusqu'à la deuxième moitié du vingtième siècle. Deuxièmement, nous sommes penchés sur le développement des approches contrefactuelles, famille dont fait partie l'approche dominante aujourd'hui. Troisièmement, nous nous sommes intéressés à l'évolution des approches par régularité en approches par inférence, répondant ainsi aux critiques qui leur étaient adressées et se plaçant en concurrence avec les approches contrefactuelles.
Une fois les grandes lignes de l'histoire de la causalité contemporaine introduites, nous nous sommes penchés sur deux points indispensables pour mieux comprendre le domaine. Pour commencer, nous avons parlé des problèmes de surdétermination, catégorie de problèmes sur lesquels reposent la plupart des débats encore existants dans le domaine. Puis, nous avons montré l'importance d'établir une distinction claire entre causalité et responsabilité car cela permet de comprendre et donc résoudre une partie importante des problématiques posées par la surdétermination.

Ensuite, nous avons exposé les défis qui restent à traiter dans le domaine de la causalité effective, notamment si nous voulons pouvoir utiliser les résultats du domaine pour l'éthique computationnelle. Le premier consiste à avoir une approche de causalité positive commune adaptée à l'éthique computationnelle. Le deuxième concerne une partie de la causalité négative, il s'agit de pouvoir raisonner sur les causes pour lesquelles un évènement ne s'est pas produit. Le troisième défi concerne la deuxième partie de la causalité négative, il s'agit de pouvoir connaître les conséquences d'une non occurrence d'un évènement. Finalement, le quatrième défi concerne la transitivité de la causalité et plus exactement le fait qu'elle soit nécessaire pour évaluer la portée que peut avoir une action en reliant l'action à ses conséquences indirectes. Cette partie était particulièrement importante car la plus grande partie des contributions de cette thèse sont une proposition de solution à ces défis. Nous avons décomposé cela en trois parties.
Dans la première nous avons montré que les approches de causalité positive existantes ne sont pas satisfaisantes pour les besoins de l'éthique computationnelle, soit parce que la définition de causalité ne convient pas, soit parce que le formalisme dans lequel elle est modélisée n'est pas adapté. Nous en avons conclu que le test NESS est la définition qui semble convenir le mieux comme base pour une approche de causalité positive commune adaptée

à l'éthique computationnelle. En effet, celle-ci a la capacité à gérer les cas de surdétermination, et de causalité positive en générale, de façon factuelle, ce qui assure une séparation entre causalité et responsabilité. Nous avons également conclu que les langages d'action et du changement présentés dans le chapitre 2 semblent être une alternative prometteuse pour résoudre les problématiques qui subsistent dans le domaine de la causalité. En effet, les formalismes des approches de causalité existantes ne sont pas aptes à représenter les subtilités nécessaires en éthique computationnelle et peuvent être à l'origine de confusions dans les cas les plus complexes de la causalité.
Dans la deuxième partie nous avons parlé des deux cas possibles lorsqu'il est question de négation dans la relation causale : le cas où la négation est dans la conséquence, relié à la notion d'empêcher; le cas où la négation est dans la cause, relié à la notion d'omission. Nous avons vu que ces notions sont complexes à traiter, et qu'elles méritent d'être étudiées indépendamment de la causalité positive. En effet, nous avons vu que le raisonnement à appliquer semble différent, ce qui pourrait indiquer qu'il s'agit d'une notion différente à celle de la causalité positive. Notamment, l'approche complètement factuelle préconisée pour le traitement de la causalité positive ne semble pas être satisfaisant dans ces cas. Pour finir, nous avons vu que la capacité qu'offrent certains langages d'action et du changement à étudier séparément causalité positive et négation dans la causalité sont une raison supplémentaire de penser qu'ils sont une alternative prometteuse pour résoudre les problématiques qui subsistent dans le domaine de la causalité.
Dans la troisième et dernière partie nous avons montré que la transitivité de la causalité semble indispensable si nous voulons arriver à évaluer la portée que peut avoir une action en reliant l'action à ses conséquences indirectes. Nous avons également vu que les débats que la transitivité suscite dans la communauté peuvent être en partie dûs à la confusion entre causalité et responsabilité. Toutefois, cela n'explique pas tout. Nous avons évoqué le fait que cette problématique est liée aux cas de négation dans la causalité. L'intrication de tous ces problèmes complexes ne facilite pas leur résolution.

Deuxième partie

Contributions

Chapitre 4

Contribution : modélisation des théories morales dans un cadre commun

« *No blame or praise may be assigned without some account of causal relationship between an agent and an outcome.* »

BEEBEE et collab. [2009]

Sommaire

Ce chapitre se veut une contribution au domaine de l'éthique computationnelle par la proposition d'un cadre commun pour la formalisation du raisonnement éthique. Dans ce cadre nous identifions et proposons une modélisation des concepts essentiels et des principales théories morales définis en éthique normative. Dit autrement, nous définissons mathématiquement les concepts et les processus qui interviennent dans les raisonnements philosophiques sur la prise de décision considérant des aspects éthiques. Le contenu de ce chapitre est le résultat d'un travail en collaboration étroite avec Guillaume Gervois avec qui je partage le goût prononcé pour la clarification de concepts.

La branche normative de l'éthique computationnelle a pour objectif de formaliser le raisonnement éthique en vue d'intégrer des principes éthiques dans des systèmes de prise de décision. Ces principes sont généralement issus de la philosophie. Toutefois, le langage utilisé en philosophie n'est pas le même qu'en informatique. De fait, intégrer ces principes philosophiques nécessite de les traduire vers une représentation informatique formelle. Cette traduction n'a d'intérêt que si elle reste fidèle à l'essence du principe que nous cherchons à traduire, mais cette fidélité est difficile à évaluer. Par exemple, s'il est clair dans le conséquentialisme que c'est uniquement à partir de ses conséquences qu'une action peut être évaluée, le concept de conséquence n'y est pas clairement défini. De plus, comme montré dans le chapitre 3, il n'existe pas aujourd'hui d'approche causale universelle, il y a ainsi plusieurs formalisations possibles du conséquentialisme.

Le cadre que nous proposons a pour objectif de dépasser la difficulté d'obtenir une traduction fidèle. Nous avons pour cela étudié le livre de TIMMONS [2012] qui décrit en détail un large panel de théories en éthique normative. Ce livre est pertinent pour proposer une modélisation de par son niveau de détail plus élevé que la plupart des introductions à l'éthique.
Les théories ont été modélisées en utilisant une architecture modulaire qui permet d'identifier clairement tous les processus y intervenant. Nous distinguons d'un côté le squelette clairement défini des théories, de l'autre les processus qui ne le sont pas, telle l'identification des conséquences d'une action. De cette façon, pour s'assurer d'être fidèle à une théorie donnée, toute formalisation peut reprendre la modélisation du squelette que nous proposons, puis proposer une version des processus indéfinis. Si l'aspect indéfini des processus prête à une certaine liberté, leur formalisation doit toutefois respecter l'intuition initiale de la théorie. Identifier clairement les choix réalisés permet ainsi de comparer différentes implémentations d'une théorie.
Afin de faciliter cette comparaison, nous avons choisi d'inscrire nos modélisations au sein du cadre commun pour le raisonnement éthique. Ce cadre contient tous les concepts essentiels à l'éthique normative tout en étant suffisamment général pour englober les formalisations déjà proposées dans le domaine. Dans le meilleur des cas, ces formalisations sont des instantiations de notre cadre, sinon ce dernier facilite la comparaison en offrant une structure générale de la théorie morale. Cet avantage qu'offre le cadre que nous proposons est mis en valeur par une étude comparative des travaux en éthique computationnelle normative.

Ce chapitre est divisé en quatre sections. La section 4.1 présente les entrées et sorties du cadre que nous proposons. Les entrées sont un contexte et un ensemble de décisions possibles, et la sortie un tri des décisions justes. D'autres sorties sont envisageables, mais celle choisie correspond au cas le plus courant en éthique normative. Dans la section 4.2 sont proposées les différentes théories morales modélisées dans notre cadre. Nous modéli-

sons aussi bien des théories axées sur le devoir, comme le commandement divin, que des théories axées sur la valeur, comme l'utilitarisme de l'acte. Puis, la section 4.3 contient une étude comparative des travaux en éthique computationnelle normative s'appuyant sur la structure proposée par le cadre commun. Finalement, la section 4.4 montre que la causalité est une brique nécessaire à la formalisation de la plupart des théories morales, et donc à l'éthique computationnelle. Les travaux présentés dans ce chapitre ont fait l'objet en 2024 d'une présentation à la journée commune des GT ACE (Aspects computationnels de l'éthique) et TADJ (Théorie algorithmique de la décision et des jeux) sur les aspects computationnels de l'éthique pour la décision individuelle ou collective et les jeux, réunissant les GDR RADIA et RO.

4.1 Entrées et sorties du cadre commun

Dans cette section nous présentons les entrées et sorties du cadre commun pour le raisonnement éthique que nous proposons. Ce cadre doit contenir tous les concepts essentiels à l'éthique normative tout en étant suffisamment général pour englober les formalisations déjà proposées dans le domaine. Ci-dessous un exemple que nous utilisons pour illustrer nos propos tout au long du chapitre. Il s'agit d'une version enrichie du problème du trolley.

Exemple 4.1 [L'aiguilleuse, la biologiste et le franc tireur]**.** *Un trolley avance sur une voie menant à une trifurcation où se trouve un aiguillage. Sur la voie de gauche se trouve un franc tireur, sur la voie centrale nous trouvons cinq civils et celle de droite est libre. Le franc tireur vise une biologiste qui se repose dans la forêt en lisière des voies. Au niveau de la trifurcation se trouve une aiguilleuse. Il se trouve que l'aiguilleuse est amoureuse de la biologiste, elle souhaite donc qu'elle reste en vie, mais elle ne sait pas que le franc tireur a été mandatée pour l'assassiner.*
De par sa fonction dans l'entreprise ferroviaire, l'aiguilleuse peut déterminer quelle voie le trolley va emprunter en actionnant le levier d'aiguillage. Sans intervention de sa part le trolley prendra la voie centrale et se dirigera droit vers les cinq civils. En poussant le levier elle peut diriger le trolley vers la voie de gauche où se trouve le franc tireur et en le tirant elle peut le diriger vers la voie vide située sur la droite. En tant que passionnée de ferroviaire, elle a pour habitude de diffuser en direct sur ses réseaux sociaux le passage des trolleys. Pour cela, elle utilise un drone qui permet de visualiser les trois voies simultanément. Elle contrôle le début de la diffusion depuis son téléphone. Toutefois, le fait de tirer le levier lui demande d'utiliser ses deux mains, ce qui l'empêche dans ce cas de lancer la diffusion.
Enfin, précisons que les individus ne peuvent pas quitter les voies sans une aide extérieure et que toute collision avec le trolley est mortelle.

Dans le chapitre 1 nous avons vu que l'éthique est sensible au contexte. Autrement dit, le statut déontique d'une action dépend en partie de faits non moraux reliés au contexte. Traiter cet exemple en éthique computationnelle demande donc d'avoir une représentation du monde dans laquelle le raisonnement éthique s'inscrit. Les entrées de notre cadre sont les informations de base nécessaires à la construction de cette représentation. Dans la section 4.1.1 nous présentons les entrées qui concernent l'action et le changement. Plus précisément, nous définissons ces entrées qui sont le contexte et les décisions. Pour que cette contribution puisse englober les formalisations déjà proposées dans le domaine, les entrées choisies sont celles d'un STEE très général. Nous le notons $\mathscr{S}_e$. Nous évitons de faire

des choix autres que ceux qui semblent indispensables pour représenter le raisonnement éthique. Dans la section 4.1.2 nous introduisons les entrées qui concernent les théories du bien, nécessaires spécifiquement aux théories axées sur la valeur. Comme présenté dans le chapitre 1, il s'agit ici d'indiquer ce qui est considéré comme ayant une valeur positive ou négative intrinsèquement. Finalement, dans la section 4.1.3 nous déterminons la sortie du cadre commun, i.e. l'évaluation des décisions. Comme présenté dans le chapitre 1, le plus courant en éthique normative est qu'évaluer moralement des actions revienne à déterminer leur statut déontique, i.e. déterminer si ces actions sont justes ou non. Cette opération peut être assimilée à un tri des actions justes.

4.1.1 Contexte et décisions, la base d'une représentation du monde éthique

Nous avons vu dans le chapitre 2 que la structure classique d'un système de transition d'état est d'avoir d'un côté l'état du monde défini comme une collection de variables le décrivant et de l'autre les transitions entre états définies par un ensemble d'actions. Les entrées de notre cadre sont le contexte et un ensemble de décisions possibles.

Commençons par parler du contexte. Comme mentionné précédemment, le statut déontique d'une action dépend en partie de faits non moraux liés au contexte. Ces faits non moraux peuvent être de nature très différente selon la théorie morale. Ils vont de faits tangibles comme des propriétés physiques du monde, à des faits intangibles comme l'état mental des agents. En l'occurrence, pour avoir une représentation de l'exemple 4.1 qui convienne à toutes les théories morales de la section 4.2, nous avons besoin de représenter la position des agents, mais aussi les connaissances de l'aiguilleuse, ses désirs et ses capacités.

Nous considérons le contexte χ comme étant une collection de fluents. Nous notons $\mathbb{F}_e$ l'ensemble des fluents. Notez l'utilisation de l'indice e permettant de distinguer cet ensemble de fluents de ceux dans les chapitres qui suivront dans cette thèse. En effet, dans les chapitres suivants nous nous concentrons principalement sur des aspects causaux qui ne nécessitent pas de modèle d'agent et donc de fluents épistémiques.

L'objectif pratique d'une théorie morale est d'évaluer les décisions des agents. Clarifions ce que nous entendons par décisions. Les agents peuvent réaliser des actions pour modifier le contexte χ dans lequel ils se trouvent. Nous notons $\mathbb{A}$ l'ensemble d'actions. Il se trouve que des actions peuvent être réalisées conjointement, et peuvent ne pas produire les mêmes conséquences que si elles étaient réalisées individuellement. Ayant vu dans les chapitres 1 et 3 que les subtilités du problème sont essentielles en éthique, il serait erroné d'évaluer la réalisation de ces actions individuellement. Ce qui est censé être évalué, ce sont les ensembles d'actions qu'un agent effectue conjointement ; nous appelons ces ensembles des décisions.

La distinction que nous proposons entre actions et décisions n'est pas nouvelle, elle peut être retrouvée dans d'autres travaux en éthique computationnelle [Bonnemains et collab., 2018; Lindner et collab., 2017]. En l'occurrence, Gelfond et Lifschitz [1998] définissent formellement l'ensemble des décisions de la façon suivante :

$$\mathbb{D} = \{i : \mathbb{A} \rightarrow \{1, 0\}\}.$$

La définition proposée suppose que toutes les décisions sont envisageables, bien que dans certains contextes, les agents puissent être incapables d'exécuter certaines décisions. Par

exemple, l'aiguilleuse est incapable de lancer la diffusion sans un téléphone chargé. De même, il n'est pas possible de pousser et tirer le levier, ou de tirer et lancer l'émission conjointement. Pour l'objectif pratique de la théorie morale, évaluer des décisions irréalisables n'est pas pertinent. Il est donc nécessaire de raisonner exclusivement sur les décisions qui respectent un ensemble de contraintes C.

Définition 4.1 [*Ensemble de décisions* $\mathbb{D}$]**.** *Étant donné un ensemble d'actions* $\mathbb{A}$ *et un ensemble de contraintes concernant la faisabilité des décisions* C*, l'ensemble de décisions* $\mathbb{D}$ *est défini comme :*

$$\mathbb{D} = \left\{ i : \mathbb{A} \to \{1,0\} \mid [\mathrm{C}]^i = 1 \right\}.$$

Pour être en mesure de raisonner sur des ensembles, nous introduisons une notation ensembliste pour une décision $d = \{a | i(a) = 1\} \cup \left\{\overline{a} | i(a) = 0\right\}$. Nous ferons référence à la décision consistant à ne réaliser aucune action comme l'*omission d'agir*. Cette notion ne doit pas être confondue avec l'*omission d'une décision* qui elle consiste à réaliser une alternative parmi celles existantes dans l'ensemble de décisions. Étant donné une décision $d \in \mathbb{D}$, omettre de réaliser d correspond à réaliser une décision $d' \in \overline{d}$, où $\overline{d} = \mathbb{D} \setminus \{d\}$ est l'ensemble des alternatives.

Exemple 4.1 [suite]**.** *Formellement, l'ensemble d'actions dans notre exemple est :*

$$\mathbb{A} = \left\{ tirer, pousser, diffuser \right\}.$$

L'ensemble de décisions réalisables $\mathbb{D}$ *respectant les contraintes énoncées dans l'exemple est :*

$$\left\{ \left\{ tirer, \overline{pousser}, \overline{diffuser} \right\}, \left\{ \overline{tirer}, pousser, diffuser \right\}, \left\{ \overline{tirer}, pousser, \overline{diffuser} \right\}, \left\{ \overline{tirer}, \overline{pousser}, diffuser \right\}, \left\{ \overline{tirer}, \overline{pousser}, \overline{diffuser} \right\} \right\}.$$

Dans notre exemple l'omission d'agir correspond à la décision $\left\{ \overline{tirer}, \overline{pousser}, \overline{diffuser} \right\}$*. Puis, si nous considérons la décision* $d = \left\{ \overline{tirer}, pousser, diffuser \right\}$*, omettre de réaliser cette décision correspond à réaliser une décision* d' *appartenant à l'ensemble des alternatives* $\overline{d}$*.*

Notez que la notion d'omission ne peut être clairement définie dans notre exemple qu'en utilisant la notion de décision. Parler d'omission d'un action est une notion ambiguë. Supposons un scénario où un agent réalise $d = \left\{ \overline{tirer}, pousser, diffuser \right\}$. Que voudrait dire omettre de pousser le levier ? Une décision parmi celles contenant $\overline{pousser}$? La décision obtenue en retirant uniquement $pousser$ de d qui serait $\left\{ \overline{tirer}, \overline{pousser}, diffuser \right\}$? Ou l'omission d'agir ?

La notion d'omission est donc une nouvelle occasion de défendre l'idée que le raisonnement doit porter sur les décisions plutôt que sur les actions. Cela peut paraître un choix divergent car en éthique il est la plupart du temps question d'actions. Toutefois, l'étude des exemples discutés en éthique normative permet de constater qu'il est souvent question de situations où l'agent a le choix entre des alternatives exclusives entre elles, comme les décisions, ce qui leur permet d'échapper à la problématique soulevée ci-dessus. La même chose peut être observée dans l'annexe A qui est une compilation d'exemples utilisés dans différents travaux en éthique computationnelle. Le cadre simplifié des expériences de pensée

est ce qui semble rendre cela possible. Lorsque dans l'exemple il peut y avoir des problématiques de combinaisons d'actions, l'énoncé du problème fait en sorte d'éliminer ces questions en présentant un nombre limité d'alternatives. De plus, ces alternatives ne sont pas présentées comme une combinaison d'actions élémentaires, et donc des décisions, mais comme les actions élémentaires du problème. Par exemple, il serait dit dans une version simplifiée de l'exemple 4.1 que l'aiguilleuse a deux choix : soit elle dévie le trolley vers la gauche ce qui aurait pour effet la mort du franc tireur mais sauverait les cinq civils et la biologiste, soit elle ne fait rien ce qui aurait pour effet la mort des cinq civils et la biologiste mais sauverait le franc tireur. Dans les deux cas, le tout serait diffusé en direct sur ses réseaux. Les variations dans la granularité de la représentation permettent de mieux faire ressortir les problématiques éthiques en question en les isolant d'autres facteurs. Toutefois, si en éthique computationnelle nous voulons formaliser le processus de raisonnement éthique, il est nécessaire d'avoir une représentation commune qui permette au moins de traiter plusieurs variantes d'un même problème, et idéalement des problèmes différents. Pour cela, il est impératif de raisonner sur les décisions et d'automatiser le processus permettant de les obtenir.

4.1.2 Théorie de la valeur, ou attribuer une valeur aux éléments du contexte

Dans cette section nous introduisons les entrées qui concernent spécifiquement les théories axées sur la valeur. Une théorie de la valeur Υ a deux objectifs, l'un théorique et l'autre pratique. Le premier est de définir pourquoi les fluents en général sont bons ou mauvais. Le second est de déterminer si un fluent spécifique est bon ou mauvais. Alors que certains Υ ne s'intéressent qu'à la distinction entre bien, mal et neutre, d'autres essaient d'aller plus loin et de déterminer dans quelle mesure les choses sont bonnes ou mauvaises. L'ensemble $\mathbb{V}_\Upsilon$ représente toutes les valuations possibles.

Définition 4.2 [*Théorie de la valeur Υ*]**.** *Étant donné un ensemble de fluents $\mathbb{F}_e$, une théorie de la valeur Υ est un processus indéfini qui évalue les fluents dans $\mathbb{F}_e$. Formellement :*

$$\Upsilon : \mathbb{F}_e \rightarrow \mathbb{V}_\Upsilon.$$

Les théories de la valeur présentées dans cette thèse utilisent deux valuations différentes. La première, la plus courante, peut être considérée comme un simple processus de classification des fluents en trois ensembles : $\mathbb{G}_\Upsilon$ l'ensemble des fluents considérés comme bons (« good » en anglais) par Υ, $\mathbb{B}_\Upsilon$ l'ensemble des fluents considérés comme mauvais (« bad » en anglais) par Υ, et $\mathbb{N}_\Upsilon$ l'ensemble des fluents considérés comme neutres (« neutral » en anglais) par Υ. Il s'agit du cas où $\mathbb{V}_\Upsilon = \{good, bad, neutral\}$. Le second est principalement utilisé par l'utilitarisme et attribue un nombre réel aux fluents ; plus le nombre est élevé, meilleur est le fluent. Il s'agit du cas où $\mathbb{V}_\Upsilon = \mathbb{R}$.

4.1.3 Théorie du juste, ou déterminer le statut déontique des décisions

Dans cette section nous nous déterminons la sortie du cadre commun. Soit $\mathbb{D}$ notre ensemble de décisions. Notre objectif est d'évaluer les éléments de $\mathbb{D}$ d'un point de vue éthique. Cela signifie que nous voulons déterminer si une décision est éthiquement requise, optionnelle ou interdite, où une action requise ou optionnelle est une action juste et où une action interdite ne l'est pas. Ainsi, désignons Θ une théorie morale qui a pour entrée une

décision et un contexte et qui donnera en sortie une évaluation éthique. Cette évaluation doit être comprise comme la sortie de la théorie du juste. L'ensemble $\mathbb{V}_\Theta$ représente toutes les évaluations possibles.

Définition 4.3 [*Théorie morale* Θ]**.** *Étant donné un ensemble de décisions* $\mathbb{D}$ *et un contexte* $\chi \in 2^{\mathbb{F}_e}$*, une théorie morale* Θ *est un processus qui évalue les décisions dans* $\mathbb{D}$*. Formellement :*

$$\Theta : \mathbb{D} \times 2^{\mathbb{F}_e} \rightarrow \mathbb{V}_\Theta.$$

Nous considérerons l'évaluation la plus courante en philosophie qui consiste à déterminer le statut déontique, i.e. à déterminer si les décisions dans $\mathbb{D}$ sont justes ou non d'un point de vue éthique. Nous désignons $\mathbb{R}_\Theta$ l'ensemble des décisions considérées comme justes (« right » en anglais) par Θ et $\mathbb{W}_\Theta$ l'ensemble des décisions considérées comme injustes (« wrong » en anglais) par Θ. Il s'agit du cas où $\mathbb{V}_\Theta = \{right, wrong\}$ Notez le choix de ne pas faire la distinction dans nos modélisations entre le requis et l'optionnel qui apparaît dans les principes tels que décrits par TIMMONS [2012]. Celui-ci a été fait par souci de clarté. En effet, nous constatons que pour la plupart des théories une décision est requise si c'est la seule juste, alors qu'elle est optionnelle s'il en existe plus d'une. La distinction peut donc facilement être réalisée une fois $\mathbb{R}_\Theta$ obtenu, sans devoir alourdir la formalisation des théories morales.

Une théorie morale Θ est dite cohérente si les ensembles de décisions justes et injustes sont disjoints. Formellement :

$$\mathbb{R}_\Theta \cap \mathbb{W}_\Theta = \emptyset.$$

4.2 Modélisation des différentes théories morales dans le cadre commun

Dans cette section nous proposons une modélisation des différentes théories morales vues dans le chapitre 1. Pour rappel, celles-ci ont été choisies de façon à donner l'aperçu le plus large possible des structures existantes. En reprenant ces théories nous espérons proposer une modélisation pour la plupart des structures existantes. Nous modélisons aussi bien des théories axées sur le devoir, que sur la valeur. Nous avons fait le choix de ne pas modéliser de théories axées sur la vertu car elles nous semblent plus adaptées à la branche descriptive de l'éthique computationnelle qu'à la branche normative à laquelle nous nous intéressons. De plus, malgré le fait que l'éthique de la vertu propose une procédure pour déterminer le statut déontique d'une action, elle reste particulièrement centrée sur l'agent et convient plutôt aux approches ascendantes que descendantes.

Nous avons choisi d'inscrire nos modélisations au sein d'un cadre commun pour le raisonnement éthique. Ce cadre contient tous les concepts essentiels à l'éthique normative tout en étant suffisamment général pour englober les formalisations déjà proposées dans le domaine. Nous modélisons les théories morales en nous basant sur une architecture modulaire avec une décomposition par processus comme dans [COINTE et collab., 2016]. Nous faisons en plus une distinction entre les processus qui sont clairement définis, pouvant être vus comme formant le squelette de la théorie morale, et les processus indéfinis dont une définition précise doit être proposée lorsqu'il est question de les formaliser. Le caractère indéfini des processus se retrouve grâce à une notation commune π. Tout processus indéfini

est noté π_{nom} où nom identifie le processus. Cette volonté de séparer les processus indéfinis de ceux ne l'étant pas pour clarifier la discussion peut être retrouvée sous forme de discussion dans [BONNEMAINS et collab., 2018].
Notez qu'une exception est faite pour la notation du processus indéfini Υ. En effet, contrairement aux autres processus indéfinis notés π_{nom}, celui-ci n'adopte pas cette convention. Cette exception s'explique par sa nature particulière qui fait de lui une entrée du système. Attribuer une valeur à chaque fluent est un processus, mais celui-ci précède le raisonnement que cherche à formaliser la branche normative de l'éthique computationnelle.

Dans la section 4.2.1 nous modélisons des théories axées sur le devoir. Plus précisément, nous modélisons la théorie du commandement divin dans la section 4.2.1.1, la théorie du relativisme moral dans la section 4.2.1.2 et la théorie morale de Kant dans la section 4.2.1.3. Dans la section 4.2.2 nous modélisons des théories axées sur la valeur. Plus précisément, nous modélisons l'utilitarisme de l'acte dans la section 4.2.2.1, l'utilitarisme espéré dans la section 4.2.2.2, le conséquentialisme satisfaisant dans la section 4.2.2.3, l'utilitarisme de la règle dans la section 4.2.2.4 et la théorie du droit naturel dans la section 4.2.2.5.

4.2.1 Théories axées sur le devoir

Cette section traite des théories morales où le concept de devoir est au moins aussi fondamental que les concepts de valeur. Pour rappel, pour déterminer le statut déontique d'une action, les théories appartenant à cette catégorie définissent directement ce qui est juste ou non, sans avoir besoin de faire intervenir les notions de bien ou de vertu.

Comme nous l'avons vu dans le chapitre 1, une partie des théories axées sur le devoir repose sur des codes de conduite. Il s'agit là d'un ensemble de normes morales qui dictent la façon juste d'agir. Au sein d'un code de conduite il est possible de trouver des normes morales de tout type. Certaines peuvent faire référence uniquement à des actions, « tu ne commettras pas de vol », d'autres peuvent par exemple faire référence à des désirs « tu ne convoiteras pas la maison de ton prochain ». Ces normes peuvent être très générales, « il est obligatoire de protéger la biosphère », ou plus précises « il est interdit d'utiliser des substances du moment où elles causent des dommages environnementaux à l'eau ». L'avantage de cette expressivité est qu'il est possible d'exprimer de nombreuses choses dans un code de conduite. Le désavantage est qu'appliquer un tel code peut s'avérer très difficile, il s'agit d'un processus qui reste la plupart du temps indéfini en éthique normative. Ce processus fait donc partie des propositions que peuvent faire les approches en éthique computationnelle.
Nous notons $\pi_{code}(d, \chi, \mathscr{C})$ le processus dont le rôle est de déduire d'un code de conduite $\mathscr{C}$ si une décision d doit être réalisée (Do), ne doit pas être réalisée (nDo), ou aucune des deux ($neither$), dans un contexte donné χ. Formellement :

$$\pi_{code} : \mathbb{D} \times 2^{\mathbb{F}_e} \times \mathscr{C} \rightarrow \{\mathrm{D}o, n\mathrm{D}o, neither\}.$$

4.2.1.1 Théorie du commandement divin

Cette section traite de la théorie du commandement divin présentée dans la section 1.2.1.1. Pour rappel, ce qui fait la spécificité de cette théorie est que le juste, le bien et le vertueux dépendent exclusivement de ce que ce dieu commande. Il est possible de voir la théorie du commandement divin comme un code de conduite dont les normes morales qui

le compose sont déterminées par le commandement divin. Le principe moral selon lequel une action est juste pour la théorie du commandement divin peut s'énoncer ainsi :

> TCD : une action est requise si et seulement si un dieu source et créateur de tout commande qu'elle soit réalisée. Une action est interdite si et seulement si ce dieu commande de ne pas la réaliser. Une action qui n'est ni requise, ni interdite est optionnelle.

Nous proposons de modéliser le principe moral derrière la théorie du commandement divin de la façon suivante :

Définition 4.4 [Θ_{tcd}]**.** *Soit $d \in \mathbb{D}$ une décision, χ un contexte, $\mathscr{C}$ le code de conduite commandé par un dieu et le processus indéfini π_{code}.*

— $d \in \mathbb{W}_{tcd}$ *si* $\pi_{code}(d, \chi, \mathscr{C}) = n\mathrm{D}o$;

— $d \in \mathbb{R}_{tcd}$ *sinon.*

Plusieurs remarques peuvent être faites à partir de cette modélisation. la première remarque est que cette théorie repose exclusivement sur le processus π_{code} consistant à appliquer un code de conduite. Sa particularité repose dans l'origine divine de ce code et les justifications de la théorie. Si nous faisons abstraction de l'origine divine du code de conduite, toute approche qui applique simplement un code de conduite supposé donné partage le même squelette que la théorie du commandement divin.
La deuxième remarque est que cette modélisation suppose que nous disposons du code de conduite fourni par le dieu en question. Il existe différents avis sur la façon dont ce code de conduite est transmis. Certains considèrent que c'est à travers les textes sacrés des religions, d'autres que c'est à travers les institutions terrestres qui représentent la religion. Mais ces avis ne sont pas les seuls, il peut en exister autant que de croyants sur terre. Ce point n'est pas un simple détail, il s'agit là d'une des principales difficultés que rencontre cette théorie.
La troisième remarque est que lorsque nous pensons à un code de conduite, nous pensons généralement à des normes sur les actions individuelles. Que se passe-t-il lorsque nous évaluons des actions réalisées conjointement? La façon d'agréger les évaluations individuelles n'a pas été définie clairement. Ce processus fait donc partie des propositions que doivent faire les approches en éthique computationnelle lorsqu'elles proposent une version de $\pi_{code}(d, \chi, \mathscr{C})$.
La quatrième remarque est que dans le cas où le code de conduite ne contient aucune information sur une action, celle-ci est considérée comme juste par défaut.

Exemple 4.1 [suite]**.** *Si l'aiguilleuse évalue sa décision avec Θ_{tcd} et que le code de conduite établi par le dieu auquel elle croit indique qu'il faut sauver à tout prix les êtres qui nous sont chers, alors il est requis de faire l'action* `pousser`*. Cette conclusion suppose que l'aiguilleuse est consciente du danger auquel est exposé la biologiste. Si par contre ce code de conduite indique que, dans des questions de vie ou de mort, il est interdit de faire une action qui change le cours des choses, alors il est interdit de faire l'action* `pousser` *ou* `tirer`*. Cela étant dit, reste l'inconnue de l'agrégation si nous voulons pouvoir choisir une des décisions, sachant que les décisions ont été définies comme un ensemble d'actions.*

4.2.1.2 Théorie du relativisme moral

Cette section traite de la théorie du relativisme moral présentée dans la section 1.2.1.2. Pour rappel, ce qui fait la spécificité de cette théorie est que le juste, le bien et le vertueux pour un individu dépendent exclusivement de la culture à laquelle cet individu appartient. Si deux individus appartiennent à des cultures adhérant à des codes de conduite différents, il est tout à fait possible que ce qui est juste pour l'un ne le soit pas pour l'autre. Il est possible de voir la théorie du relativisme moral comme un code de conduite dont les normes morales qui le compose sont déterminées par la culture à laquelle l'individu qui évalue appartient. Le principe moral selon lequel une action est juste pour la théorie du relativisme moral peut s'énoncer ainsi :

> TRM : une action est requise, interdite ou optionnelle pour des membres d'une culture si et seulement si un code de conduite de cette culture le stipule ainsi.

Comme la théorie du commandement divin, cette théorie repose principalement sur un code de conduite, elles partagent la même structure. Au premier abord, trouver le code de conduite déterminé par une culture semble plus accessible que trouver le code de conduite déterminé par un dieu. Il est possible d'imaginer des moyens de consultation qui permettraient d'approximer un tel code. Par exemple, il serait possible de demander à tous les membres d'une culture d'écrire un code de conduite et d'essayer d'extraire les normes morales les plus partagées. Toutefois, en essayant d'imaginer un processus il est facile de se rendre compte qu'il s'agit d'une tâche très complexe. En l'occurrence, comment déterminer à quelle culture appartient un individu ? Pour cela il faudrait déjà être capables de définir précisément des cultures. En outre, il n'est pas improbable de penser qu'une culture peut voir cohabiter plusieurs codes de conduite, ou qu'un individu peut appartenir à plusieurs cultures simultanément. Comme $\pi_{code}(d,\chi,\mathscr{C})$, déterminer le code de conduite de la culture à laquelle appartient un agent est un processus indéfini en éthique normative. Ce processus fait donc partie des propositions que peuvent faire les approches en éthique computationnelle.
Nous notons $\pi_{cocu}(\chi)$ le processus dont le rôle est de déterminer le code de conduite de la culture à laquelle appartient un agent. $cocu$ correspond à l'abréviation de « code of culture ». Formellement :

$$\pi_{cocu} : 2^{\mathbb{F}_e} \rightarrow \mathscr{C}.$$

Nous proposons de modéliser le principe moral derrière la théorie du relativisme moral de la façon suivante :

Définition 4.5 [Θ_{trm}]**.** *Soit $d \in \mathbb{D}$ une décision, χ un contexte et les processus indéfinis π_{code} et π_{cocu}.*

- $d \in \mathbb{W}_{trm}$ *si* $\pi_{code}(d,\chi,\pi_{cocu}(\chi)) = n\mathrm{D}o$;
- $d \in \mathbb{R}_{trm}$ *sinon.*

La modélisation proposée fait ressortir clairement que, contrairement à la théorie du commandement divin, la théorie du relativisme morale ne peut pas se résumer à π_{code}, elle requiert un processus supplémentaire visant à déterminer le code de conduite de la culture de l'individu. Il n'est donc pas possible de considérer que n'importe quelle approche qui applique un code de conduite est une formalisation du relativisme moral.

Exemple 4.1 [suite]**.** *Si l'aiguilleuse évalue sa décision avec Θ_{trm} et que le code de conduite établi par la culture à laquelle elle appartient indique qu'il faut sauver à tout prix les êtres qui nous sont chers, alors il est requis de faire l'action $pousser$. Si par contre ce code de conduite indique que, dans des questions de vie ou de mort, il est interdit de faire une action qui change le cours des choses, alors il est interdit de faire l'action $pousser$ ou $tirer$. Comme précédemment, il reste l'inconnue de l'agrégation si nous voulons pouvoir choisir une des décisions.*

4.2.1.3 Théorie morale de Kant

Cette section traite de la théorie morale de Kant présentée dans la section 1.2.2. Pour rappel, le principe moral qu'il propose pour définir le juste, le bien et la vertu est connu comme l'Impératif Catégorique. Ce qui caractérise un impératif catégorique est qu'il requiert qu'une action soit faite quels que soient les objectifs propres à l'agent. Un tel impératif ne peut exister que s'il existe un objectif imposable à tous pour l'incarner. Un tel objectif est dit être une fin en soi. Sous cette condition, un impératif catégorique possède l'inconditionnalité nécessaire pour être un principe moral. Kant propose l'autonomie pour incarner son Impératif Catégorique, i.e. la capacité inhérente à tout agent rationnel d'agir librement sur la base de la raison et indépendamment de ses désirs propres. Il appelle « humanité » cette autonomie propre aux agent rationnels. Plusieurs formulations de l'Impératif Catégorique ont été proposées par le philosophe, formulations qu'il revendique équivalentes. Pour rappel, nous nous intéressons à deux d'entre elles. Voici celle qui fait apparaître explicitement la notion d'humanité :

> HFS : agis toujours de manière à traiter l'humanité, aussi bien dans ta personne que dans la personne des autres, comme une fin et à ne t'en servir jamais comme d'un simple moyen. [KANT, 1785]

Deux parties composent ce principe moral, une première positive et une deuxième négative. La partie positive exige de traiter l'humanité comme une fin. Cela équivaut pour Kant à agir de sorte à promouvoir le perfectionnement de soi et le bonheur des autres. La partie négative interdit de se servir des autres comme de simples moyens. Bien que la signification exacte de ce que cela veut dire suscite de nombreuses discussions en philosophie, il semblerait qu'il existe un courant majoritaire qui associe cela à traiter les autres de façon à ce qu'ils ne puissent pas consentir rationnellement. Toutefois, déterminer si un agissement traite l'humanité comme une fin et ne s'en sert pas comme d'un simple moyen est un processus indéfini en éthique normative. Ce processus fait donc partie des propositions que peuvent faire les approches en éthique computationnelle.
Nous notons $\pi_{hfs}(d,\chi)$ le processus dont le rôle est de déterminer si un agissement traite l'humanité comme une fin et ne s'en sert pas comme d'un simple moyen. hfs correspond à l'abréviation de « humanité comme une fin en soi ». Formellement :

$$\pi_{hfs} : \mathbb{D} \times 2^{\mathbb{F}_e} \rightarrow \{1,0\}.$$

Nous proposons de modéliser le principe moral derrière cette formulation de l'Impératif Catégorique de la façon suivante :

Définition 4.6 [Θ_{hfs}]**.** *Soit $d \in \mathbb{D}$ une décision, χ un contexte et un processus indéfini π_{hfs}.*

- *$d \in \mathbb{R}_{hfs}$ si $\pi_{hfs}(d,\chi) = 1$;*
- *$d \in \mathbb{W}_{hfs}$ sinon.*

Kant propose deux autres versions de son Impératif Catégorique. Parmi les trois formulations existantes, celle qui suit est, selon TIMMONS [2012], celle qui répond le plus directement à l'objectif pratique de la théorie morale, ce qui fait d'elle la plus intéressante pour l'éthique computationnelle. De plus, cette deuxième formulation suggère un mécanisme d'évaluation des actions intéressant par sa construction.

> LU : agis uniquement d'après la maxime qui fait que tu puisses vouloir en même temps qu'elle devienne une loi universelle. [KANT, 1785]

Pour rappel, une maxime est une déclaration de la forme : je vais faire A, si certaines conditions sur l'état du monde S sont réunies, afin d'accomplir O. C'est une formalisation d'un état psychologique d'un agent indiquant ce qu'il désire faire. Adopter une maxime revient à agir selon cette maxime. Le principe moral LU indique que pour savoir si une action est juste moralement, un agent doit d'abord identifier la maxime selon laquelle cette action est réalisée. Comme nous l'avons vu dans la section 1.2.2, ce processus n'est pas anodin. En effet, selon la façon dont la maxime justifiant cette action est formulée, il est possible que le statut déontique d'une même action change. Ce que TIMMONS [2012] suggère est que le principe HFS peut être utilisé comme une heuristique pour savoir quels détails de l'action, des circonstances et de l'objectif peuvent apparaître dans la maxime. Toutefois, il s'agit là d'un des nombreux points de vue possibles. Déterminer la maxime selon laquelle une action est réalisée est un processus indéfini en éthique normative. Ce processus fait donc partie des propositions que peuvent faire les approches en éthique computationnelle. Nous notons $\pi_{maxi}(d,\chi)$ le processus dont le rôle est de déterminer la maxime selon laquelle une action est réalisée. $maxi$ correspond à l'abréviation de « maxime ». Formellement :

$$\pi_{maxi} : \mathbb{D} \times 2^{\mathbb{F}_e} \rightarrow \mathbb{D} \times 2^{\mathbb{F}_e} \times 2^{\mathbb{F}_e}.$$

Une fois la maxime selon laquelle une action est réalisée déterminée, la deuxième étape consiste à considérer ce que serait le monde si tous les êtres rationnels adoptaient cette maxime. Si dans cette situation hypothétique aucune incohérence apparaît, la maxime est vouée à être une loi universelle et l'action est requise. Dans le cas contraire, l'action est interdite. Déterminer si une maxime est universalisable est un processus indéfini en éthique normative. Ce processus fait donc partie des propositions que peuvent faire les approches en éthique computationnelle.
Nous notons $\pi_{lu}\left(\pi_{maxi}(d,\chi),\chi\right)$ le processus dont le rôle est de déterminer si une maxime est universalisable. lu correspond à l'abréviation de « loi universelle ». Formellement :

$$\pi_{lu} : \mathbb{D} \times 2^{\mathbb{F}_e} \times 2^{\mathbb{F}_e} \times 2^{\mathbb{F}_e} \rightarrow \{1,0\}.$$

Le principe moral selon lequel une action est juste pour la théorie morale de Kant peut s'énoncer ainsi :

> TMK : une action est requise si et seulement si la maxime selon laquelle la réalisation de l'action est omise n'est pas universalisable. Une action est interdite si et seulement si la maxime selon laquelle l'action est réalisée n'est pas universalisable. Une action qui n'est ni requise, ni interdite est optionnelle.

Nous proposons de modéliser le principe moral derrière la théorie morale de Kant de la façon suivante :

Définition 4.7 [Θ_{lu}]**.** *Soit $d \in \mathbb{D}$ une décision, χ un contexte et les processus indéfinis π_{maxi} et π_{lu}.*

— *$d \in \mathbb{R}_{lu}$ si $\pi_{lu}\left(\pi_{maxi}(d,\chi),\chi\right) = 1$;*

— *$d \in \mathbb{W}_{lu}$ sinon.*

Exemple 4.1 [suite]**.** *Supposons que l'aiguilleuse évalue son agir avec Θ_{lu}. Imaginons qu'elle découvre les plans du franc tireur et qu'elle décide de faire la décision $\left\{\overline{tirer}, pousser, \overline{diffuser}\right\}$. Imaginons que la maxime selon laquelle elle agit est la suivante : je vais aller jusqu'à tuer une personne, si cette personne agit de façon injuste et que sa mort est nécessaire, afin de sauver une personne qui m'est chère. Essayons de l'universaliser. Y aurait-il une incohérence dans le monde si tous les individus étaient prêts à tuer une personne si celle-ci agissait de façon injuste et que sa mort était nécessaire pour sauver une personne qui nous est chère ? Il semblerait que oui. Il est possible de considérer que le franc tireur est assassiné alors qu'il essaie juste de protéger la société de la biologiste qui conçoit une arme biologique pour la vendre sur le marché noir. Si nous adoptons le point de vue du collègue du franc tireur, il peut considérer que son collègue est assassiné injustement. De ce fait, l'universalisation de cette maxime impliquerait qu'il serait juste pour lui de tuer l'aiguilleuse. Le monde dans lequel la maxime est universalisée ne peut pas être souhaitable pour l'aiguilleuse ou n'importe quel agent qui tient à sa vie. La décision $\left\{\overline{tirer}, pousser, \overline{diffuser}\right\}$ est donc injuste.*

4.2.2 Théories axées sur la valeur

Cette section traite des théories morales où le concept de valeur est antérieur au concept de devoir. Elles partagent l'idée qu'il est possible, sans faire appel à la notion de juste, d'attribuer une valeur aux constituants des états du monde et donc comparer différents constituants ou même différents états du monde entre eux. Pour rappel, dans toutes ces théories morales, le concept de valeur est utilisé pour déterminer le statut déontique d'une action.

Comme nous l'avons vu dans le chapitre 1, une partie importante des théories axées sur la valeur sont conséquentialistes. Cela veut dire qu'elles reposent sur l'idée que la valeur des conséquences associées à une action est la seule chose à prendre en compte pour déterminer si une action est juste. À partir de là, un grand nombre de variantes peuvent être construites en faisant varier la théorie de la valeur ou la théorie du juste. En fonction de la variante, certaines précisions sont données sur ce qu'est considéré comme la conséquence d'une action. Cependant, la notion de conséquence n'est pas clairement définie, il n'est pas défini s'il s'agit d'une notion purement causale ou si l'intuition derrière inclue des notions de responsabilité. Imaginons le cas simple où il s'agirait d'une notion purement causale. Comme montré dans le chapitre 3, il existe diverses définitions de ce qu'est la causalité. Même dans ce cas simple le processus consistant à déterminer les conséquences associées à une action reste indéfini en éthique normative. Ce processus fait donc partie des propositions que peuvent faire les approches en éthique computationnelle.
Nous notons $\pi_{cons}(d,\chi)$ le processus dont le rôle est de déterminer les conséquences associées à une décision. $cons$ correspond à l'abréviation de « conséquences ». Formellement :

$$\pi_{cons} : \mathbb{D} \times 2^{\mathbb{F}_e} \rightarrow 2^{\mathbb{F}_e}.$$

4.2.2.1 Utilitarisme de l'acte

Cette section traite de l'utilitarisme de l'acte présenté dans la section 1.3.1.1. Pour rappel, l'utilitarisme de l'acte est sans aucun doute la théorie conséquentialiste la plus connue. Jeremy Bentham et John Stuart Mill sont considérés comme les plus grands exposants des bases de ce conséquentialisme. L'utilitarisme de l'acte peut être considéré comme un conséquentialisme hédoniste, agent-neutre, non priorisant, direct, effectif, universaliste et maximisant. Le principe moral selon lequel une action est juste pour l'utilitarisme de l'acte peut s'énoncer ainsi :

> UA : une action est requise si et seulement si son utilité est supérieure à l'utilité de toutes les actions alternatives. Une action est interdite s'il existe une action alternative dont l'utilité est supérieure à la sienne. Une action qui n'est ni requise, ni interdite est optionnelle.

Les théories utilitaristes parlent d'utilité pour faire référence à la valeur des conséquences d'une action. Elle s'obtient en sommant la valeur de toutes les conséquences négatives de l'action, puis en sommant la valeur de toutes les conséquences positives de l'action, pour enfin sommer ces deux résultats et ainsi obtenir la valeur totale de l'action, son utilité. L'utilitarisme de l'acte étant un conséquentialisme hédoniste, seules les expériences de plaisir et de souffrance peuvent avoir une valeur intrinsèque. En conséquence, l'utilité d'une action est la somme de la souffrance causée à laquelle s'ajoute la somme du plaisir causé. Le calcul de l'utilité est clairement défini. Mais celui-ci n'est possible que s'il est possible d'attribuer une valeur aux expériences de plaisir et de souffrance. Ainsi, pour formaliser l'utilitarisme de l'acte, nous avons besoin d'une théorie de la valeur permettant de déterminer la valeur intrinsèque de f qui est liée au plaisir et à la souffrance qui lui sont associés. Nous notons un tel processus Υ_{ua}, où ua désigne l'utilitarisme de l'acte. Dans le chapitre 1 nous avons vu qu'il n'existe pas une unique théorie de la valeur pour l'utilitarisme de l'acte ; aussi bien BENTHAM [1789] que MILL [1863] ont proposé leur propre théorie.

Nous proposons de modéliser le principe moral derrière l'utilitarisme de l'acte de la façon suivante :

Définition 4.8 [Θ_{ua}]**.** *Soit $d \in \mathbb{D}$ une décision, χ un contexte et les processus indéfinis π_{cons} et Υ_{ua}.*

- *$d \in \mathbb{R}_{ua}$ si $\forall d' \in \overline{d}, \sum_{f \in \pi_{cons}(d,\chi)} \Upsilon_{ua}(f) \geq \sum_{f \in \pi_{cons}(d',\chi)} \Upsilon_{ua}(f)$;*
- *$d \in \mathbb{W}_{ua}$ sinon.*

Notez ici que pour modéliser cette théorie il est indispensable de raisonner sur les alternatives. Il est donc nécessaire d'avoir une définition claire de ce que sont ces alternatives, et donc de parler de décisions et non d'actions.

Exemple 4.1 [suite]**.** *Nous nous plaçons dans le cas où l'aiguilleuse évalue sa décision avec Θ_{ua} et que la vie des individus a une valeur incommensurable par rapport à toute autre valeur, ce qui nous permet dans un premier temps d'ignorer les conséquences de l'action $diffuser$. Imaginons que toutes les décisions contenant l'action $pousser$ ont dans leur conséquences le fait que six vies soient sauvées, celles des cinq civils et de la biologiste dans la forêt, et la mort du franc tireur. Puis, imaginons que toutes les décisions contenant l'action $tirer$ ont dans leur conséquences le fait que six vies soient sauvées, celles des cinq civils et du franc tireur,*

et la mort de la biologiste. Finalement, imaginons que toutes les décisions contenant ni l'action $tirer$, ni l'action $pousser$, ont dans leur conséquences le fait que la vie du franc tireur soit sauvée et la mort de six individus, celle des cinq civils et de la biologiste.
Dans ce cas là, Θ_{ua} déterminerait dans un premier temps que les décisions ne contenant ni l'action $tirer$, ni l'action $pousser$, sont injustes. En effet, en prenant juste en compte les conséquences sur la vie des individus, ces décisions ont les utilités les plus basses. Avec cette même hypothèse, les décisions contenant l'action $pousser$ ou $tirer$ ont exactement la même utilité. Si nous considérons maintenant que l'action $diffuser$ a comme conséquence le fait qu'un nombre important de personnes vont voir ce qu'il se passe sur les voies, la décision consistant à $pousser$ et à $diffuser$ aura comme conséquences le fait que toutes ces personnes verront une situation traumatisante. Cette décision est donc considérée comme injuste. Les deux seules décisions considérées comme justes par Θ_{ua} sont alors :

$$\left\{tirer, \overline{pousser}, \overline{diffuser}\right\}, \left\{\overline{tirer}, pousser, \overline{diffuser}\right\}.$$

4.2.2.2 Utilitarisme espéré

Cette section traite de l'utilitarisme espéré mentionné dans la section 1.3.1. Pour rappel, l'utilitarisme espéré est une des alternatives proposées à l'utilitarisme de l'acte. La principale différence dans cette version de l'utilitarisme est qu'elle considère que toutes les conséquences ne doivent pas être prises en compte de la même façon. En effet, cette théorie avance que dans l'évaluation du statut déontique il est nécessaire de prendre en compte à quel point l'agent espérait que l'action ait la conséquence en question. Il est ici question des croyances de l'agent. L'utilitarisme espéré considère que pour agréger la valeur des conséquences espérées des actions il faut, avant de sommer le tout, multiplier la valeur de chaque conséquence par la probabilité que l'agent attribue à leur production effective. À nouveau, le calcul de l'utilité est clairement défini, à condition que nous disposions de tous les éléments. Malheureusement, la probabilité que l'agent attribue à la production effective d'une conséquence n'est pas quelque chose d'anodin à obtenir et la façon dont cela doit être fait reste indéfinie en éthique normative. Ce processus fait donc partie des propositions que peuvent faire les approches en éthique computationnelle.
Nous notons $\pi_{prob}(d, \chi, f)$ le processus dont le rôle est de déterminer la probabilité que l'agent attribue à la production effective d'une conséquence suite à la réalisation d'une décision. $prob$ correspond à l'abréviation de « probabilité ». Formellement :

$$\pi_{prob} : \mathbb{D} \times 2^{\mathbb{F}_e} \times \mathbb{F}_e \rightarrow \mathbb{V}_\Upsilon.$$

Nous proposons de modéliser le principe moral derrière l'utilitarisme espéré de la façon suivante :

Définition 4.9 [Θ_{ue}]**.** *Soit $d \in \mathbb{D}$ une décision, χ un contexte et les processus indéfinis π_{cons}, π_{prob} et Υ_{ua}.*

- *$d \in \mathbb{R}_{ue}$ si $\forall d' \in \overline{d}, \sum_{f \in \pi_{cons}(d,\chi)} \pi_{prob}(d, \chi, f)\Upsilon_{ua}(f) \geq \sum_{f \in \pi_{cons}(d',\chi)} \pi_{prob}(d', \chi, f)\Upsilon_{ua}(f)$;*
- *$d \in \mathbb{W}_{ue}$ sinon.*

Notez que nous considérons ici la version de l'utilitarisme espéré qui ne varie de l'utilitarisme de l'acte que par l'aspect des conséquences espérées. Pour être précis, ce que nous

modélisons dans cette section est un conséquentialisme hédoniste, agent-neutre, non priorisant, direct, espéré (au lieu d'effectif), universaliste et maximisant. Nous pouvons donc utiliser ici le processus Υ_{ua} correspondant à la théorie de la valeur de l'utilitarisme de l'acte. Nous ferrons de même pour les autres théories conséquentialistes. Toutefois, il est important de garder à l'esprit que différentes théories du bien peuvent être utilisées. Cela aura pour effet de modifier le conséquentialisme qui est modélisé. Plus exactement, en modifiant la théorie de la valeur les caractéristiques qui peuvent changer dans la description sont : hédoniste, agent-neutre et non priorisant.

Exemple 4.1 [suite]**.** *Nous nous plaçons dans le cas où l'aiguilleuse évalue sa décision avec* Θ_{ue} *et que la vie des individus a une valeur incommensurable par rapport à toute autre valeur, ce qui nous permet dans un premier temps d'ignorer les conséquences de l'action* $diffuser$*. Imaginons que toutes les décisions contenant l'action* $pousser$ *ont dans leur conséquences le fait que six vies soient sauvées, celles des cinq civils et de la biologiste dans la forêt, et la mort du franc tireur. Toutefois, la description de l'exemple 4.1 précise bien que l'aiguilleuse ne savait pas que le franc tireur menaçait la vie de la biologiste. Nous supposons que la probabilité que l'action* $pousser$ *lui sauve la vie est nulle d'après les croyances de l'aiguilleuse. Nous supposons également que la probabilité qu'un individu meurt suite au choc avec le trolley est de un d'après les croyances de l'aiguilleuse, tout comme la probabilité que les cinq civils soient sauvés si elle dévie le trolley. Nous pouvons donc dire que pour le calcul d'utilité toutes les décisions contenant l'action* $pousser$ *ont en réalité dans leur conséquences espérées le fait que les cinq vies civiles soient sauvées et la mort du franc tireur. Selon le même raisonnement, toutes les décisions contenant l'action* $tirer$ *ont dans leur conséquences espérées le fait que six vies soient sauvées, celles des cinq civils et du franc tireur. Finalement, toutes les décisions contenant ni l'action* $tirer$*, ni l'action* $pousser$*, ont dans leur conséquences espérées le fait que la vie du franc tireur soit sauvée et la mort des cinq civils.*
Dans ce cas là, Θ_{ue} *déterminerait dans un premier temps que les décisions ne contenant ni l'action* $tirer$*, ni l'action* $pousser$*, ou contenant uniquement l'action* $pousser$ *sont injustes. En effet, en prenant juste en compte les conséquences sur la vie des individus, ces décisions ont les utilités les plus basses. Avec cette même hypothèse, les décisions contenant l'action* $tirer$ *ont l'utilité la plus élevée.* $\left\{tirer, \overline{pousser}, \overline{diffuser}\right\}$ *est dans ce cas la seule décision considérée comme juste.*

4.2.2.3 Conséquentialisme satisfaisant

Cette section traite du conséquentialisme satisfaisant présenté dans la section 1.3.1.3. Pour rappel, le conséquentialisme satisfaisant est une autre alternative proposée à l'utilitarisme de l'acte. Il s'agit d'une version moins exigeante. Cette version du conséquentialisme adopte l'idée qu'une action est juste si ses conséquences sont suffisamment bonnes. En effet, même s'il existe des alternatives meilleures, du moment où les conséquences de l'action sont suffisamment bonnes, alors l'action est juste. Le principe moral selon lequel une action est juste pour le conséquentialisme satisfaisant peut s'énoncer ainsi :

> CS : une action est requise si c'est l'unique action dont utilité est supérieure ou égale au seuil d'utilité spécifié, ou si son utilité est supérieur à celle de toutes les alternatives et qu'aucune action n'a une utilité supérieure au seuil. Une action est interdite si son utilité est inférieure au seuil d'utilité spécifié et qu'il existe

des actions dont l'utilité est supérieure. Une action qui n'est ni requise, ni interdite est optionnelle.

Nous considérons ici la version du conséquentialisme satisfaisant qui ne varie de l'utilitarisme de l'acte que par l'aspect maximisant. Pour être précis, ce que nous modélisons dans cette section est un conséquentialisme hédoniste, agent-neutre, non priorisant, direct, effectif, universaliste et satisfaisant (au lieu de maximisant). Nous proposons de modéliser le principe moral derrière le conséquentialisme satisfaisant de la façon suivante :

Définition 4.10 [Θ_{cs}]**.** *Soit $d \in \mathbb{D}$ une décision, χ un contexte, τ le seuil d'utilité fixé et les processus indéfinis π_{cons} et Υ_{ua}.*

— *$d \in \mathbb{R}_{cs}$ si :*
  — $\left(\sum_{f \in \pi_{cons}(d,\chi)} \Upsilon_{ua}(f) \geq \tau\right)$*, ou*
  — $\left(\sum_{f \in \pi_{cons}(d,\chi)} \Upsilon_{ua}(f) < \tau\right) \wedge (d \in \mathbb{R}_{ua})$.
— *$d \in \mathbb{W}_{cs}$ sinon.*

Notez l'apparition de $\mathbb{R}_{ua}$ dans la définition de Θ_{cs}. Dans le cas où l'utilité de la décision n'est pas supérieure au seuil déterminé, elle peut tout de même être considérée comme juste si son utilité est supérieure à l'utilité de toutes les décisions alternatives. Dans ce cas très particulier ou l'utilité d'aucun décision ne satisfait le seuil, l'aspect maximisant de l'utilitarisme de l'acte est réintroduit ce qui permet d'obtenir au moins une décision juste dans tous les cas.
Comme précédemment, nous pouvons utiliser ici le processus Υ_{ua} correspondant à la théorie de la valeur de l'utilitarisme de l'acte. Toutefois, il est important de garder à l'esprit que si nous changeons la théorie de la valeur, il est également nécessaire de changer la théorie de la valeur permettant de déterminer si $d \in \mathbb{R}_{ua}$. Dans le cas contraire, nous nous retrouverions dans une configuration où deux théories du bien différentes sont utilisées pour une même théorie du juste : celle choisie pour Θ_{cs} qui viendrait remplacer Υ_{ua} et Υ_{ua} qui serait utilisée pour Θ_{ua}.

Exemple 4.1 [suite]**.** *Nous nous plaçons dans le cas où l'aiguilleuse évalue sa décision avec Θ_{cs} et que la vie des individus a une valeur incommensurable par rapport à toute autre valeur, ce qui nous permet dans un premier temps d'ignorer les conséquences de l'action $diffuser$. Imaginons que toutes les décisions contenant l'action $pousser$ ont dans leur conséquences le fait que six vies soient sauvées, celles des cinq civils et de la biologiste dans la forêt, et la mort du franc tireur. Puis, imaginons que toutes les décisions contenant l'action $tirer$ ont dans leur conséquences le fait que six vies soient sauvées, celles des cinq civils et du franc tireur, et la mort de la biologiste. Ensuite, imaginons que toutes les décisions contenant ni l'action $tirer$, ni l'action $pousser$, ont dans leur conséquences le fait que la vie du franc tireur soit sauvée et la mort de six individus, celle des cinq civils et de la biologiste. Finalement, imaginons que le seuil τ corresponde simplement au fait qu'il faut qu'il y ait plus de vies sauvées que de morts. Dans ce cas là, Θ_{cs} déterminerait que les décisions contenant uniquement l'action $pousser$ ou $tirer$ sont justes, alors que celles contenant ni l'action $tirer$, ni l'action $pousser$, sont injustes. En effet, ces dernières ont une utilité qui n'atteint pas le seuil et toutes les autres ont une utilité qui atteint le seuil. Les décisions justes sont alors :*

$$\left\{tirer, \overline{pousser}, \overline{diffuser}\right\}, \left\{\overline{tirer}, pousser, diffuser\right\}, \left\{\overline{tirer}, pousser, \overline{diffuser}\right\}.$$

4.2.2.4 Utilitarisme de la règle

Cette section traite de l'utilitarisme de la règle présenté dans la section 1.3.1.2. Pour rappel, l'utilitarisme de la règle est une autre alternative proposée à l'utilitarisme de l'acte. Ce conséquentialisme est indirect car il ne va pas évaluer les conséquences des actions individuelles faites dans un contexte donné pour déterminer si une action est juste ou non. Il va plutôt évaluer les conséquences qu'aurait l'adoption d'une norme morale par tous les individus. Cette version du conséquentialisme satisfait l'idée attrayante que certaines actions, comme tuer pour le plaisir, sont interdites par le type d'action qu'elles sont et donc en toutes circonstances. Le principe moral selon lequel une action est juste pour l'utilitarisme de la règles peut s'énoncer ainsi :

> UR : une action est requise, interdite ou optionnelle si et seulement si le code de conduite idéal le stipule ainsi.

Comme pour la théorie du commandement divin et du relativisme moral, cette théorie repose principalement sur un code de conduite ; elles partagent la même structure et le processus π_{code} y est également nécessaire. Toutefois, contrairement aux précédentes théories basées sur des codes de conduite, l'obtention du code de conduite à appliquer est justifié d'un point de vue conséquentialiste. En effet, il faut en premier regarder les conséquences de l'adoption par tous les individus d'une norme morale. À partir de cette opération élémentaire, ce conséquentialisme propose de construire un code de conduite dit « idéal ». Un tel code de conduite correspond à un ensemble de normes morales dont les conséquences de son adoption sont supérieures ou égales à celles de toute autre ensemble de normes morales. La construction d'un tel code n'est pas quelque chose d'anodin, la façon dont cela doit être fait reste indéfinie en éthique normative. Ce processus fait donc partie des propositions que peuvent faire les approches en éthique computationnelle.
Nous notons $\pi_{coid}(2^{\mathbb{F}_e}, \Upsilon)$ le processus dont le rôle est d'obtenir un code le conduite idéal pour une théorie de la valeur donnée. $coid$ correspond à l'abréviation de « code idéal ». Formellement :

$$\pi_{coid} : 2^{2^{\mathbb{F}_e}} \times \mathbb{V}_{\Upsilon}{}^{\mathbb{F}_e} \to \mathscr{C}.$$

Nous proposons de modéliser le principe moral derrière l'utilitarisme de la règle de la façon suivante :

Définition 4.11 [Θ_{ur}]**.** *Soit $d \in \mathbb{D}$ une décision, χ un contexte et les processus indéfinis π_{coid}, π_{code} et Υ_{ua}.*

- *$d \in \mathbb{W}_{ur}$ si $\pi_{code}\left(d, \chi, \pi_{coid}(2^{\mathbb{F}_e}, \Upsilon_{ua})\right) = n\mathrm{D}o$;*
- *$d \in \mathbb{R}_{ur}$ sinon.*

Notez que nous considérons ici la version de l'utilitarisme de la règle qui ne varie de l'utilitarisme de l'acte que par l'aspect direct. Pour être précis, ce que nous modélisons dans cette section est un conséquentialisme hédoniste, agent-neutre, non priorisant, indirect (au lieu de direct), effectif, universaliste et maximisant. Il est indirect car il évalue les conséquences qu'aurait l'adoption d'une norme morale par tous les individus et non les conséquences réelles de la décision évaluée.
Notez également que la modélisation proposée fait ressortir clairement que cette théorie est très proche du relativisme morale, leur squelette est le même. Comme cette dernière, l'utilitarisme de la règle ne peut pas se résumer à π_{code}, elle requiert un processus supplémentaire

visant à déterminer le code de conduite. Si pour le relativisme morale ce code dépendait de la culture de l'individu, dans l'utilitarisme de l'acte, l'obtention de celui-ci est justifié d'un point de vue conséquentialiste.

Exemple 4.1 [suite]**.** *Nous nous plaçons dans le cas où l'aiguilleuse évalue sa décision avec* Θ_{ur}*. Imaginons que la norme morale selon laquelle il faut toujours faire ce qui semble sauver le plus de vies à première vue est dans le code de conduite idéal. Imaginons que toutes les décisions contenant l'action* $pousser$ *ont à première vue dans leur conséquences le fait que les cinq vies civiles soient sauvées et la mort du franc tireur. Puis, imaginons que toutes les décisions contenant l'action* $tirer$ *ont à première vue dans leur conséquences le fait que six vies soient sauvées, celles des cinq civils et du franc tireur. Finalement, toutes les décisions contenant ni l'action* $tirer$*, ni l'action* $pousser$*, ont à première vue dans leur conséquences le fait que la vie du franc tireur soit sauvée et la mort des cinq civils. Dans ce cas là, il est requis de réaliser la décision contenant l'action* $tirer$ *:*

$$\left\{tirer, \overline{pousser}, \overline{diffuser}\right\}.$$

4.2.2.5 Théorie du droit naturel

Cette section traite de la théorie du droit naturel présentée dans la section 1.3.2. Pour rappel, ce qui fait la spécificité de cette théorie est la croyance en l'existence d'un ensemble objectif de principes moraux ancrés dans la nature humaine et donc d'autorité supérieure par leur source naturelle. D'AQUIN [1266] est considéré comme un des exposants les plus importants de ce courant de pensé. Le principe moral selon lequel une action est juste pour la théorie du droit naturel de Thomas d'Aquin peut s'énoncer ainsi :

> TDN : une action est requise si et seulement si ne pas la réaliser résulte en une violation directe d'au moins une des valeurs de base. Une action est interdite si et seulement si la réaliser résulte en une violation directe d'au moins une des valeurs de base. Une action qui n'est ni requise, ni interdite est optionnelle.

Thomas d'Aquin propose une théorie de la valeur perfectionniste comme base de sa théorie morale. Il s'agit de l'idée selon laquelle ce qui a de la valeur pour nous les humains sont les états du monde qui permettent le développement des capacités qui nous sont propres. Cela veut dire que plus un humain se rapproche d'un état de perfection, plus l'état du monde dans lequel cela se produit a de la valeur positive intrinsèque. Dans cette théorie les valeurs de base qui permettent un rapprochement de cette perfection sont la vie, la procréation, la connaissance et la sociabilité. Nous notons cette théorie de la valeur Υ_{tdn}, où tdn désigne la théorie du droit naturel.

En pratique, il est possible de considérer qu'une action est une violation directe d'une valeur de base si l'action ne peut pas être justifiée par la doctrine du double effet. Pour rappel, d'après ce principe, une action est juste si les conditions suivantes sont satisfaites :

1. *L'action ne doit pas être intrinsèquement interdite.* Cette première condition est une des composantes les plus distinctives de cette théorie, notamment par rapport aux autres théories axées sur la valeur. Il s'agit de l'absolutisme moral, idée selon laquelle certaines actions ne sont jamais justes quels que soient le contexte et les conséquences qui résulteraient de leur réalisation. Le statut déontique de ces actions est toujours le même, elles sont interdites. Ce premier point peut être rapproché de la notion de

code de conduite. Toutefois, il s'agit d'un code de conduite un peu particulier car il ne stipule que ce qui est interdit de faire et cela en regardant la nature intrinsèque des actions et non pas leur conséquences comme dans l'utilitarisme de la règle. Il s'agit là d'un code de conduite plus proche de celui vu dans la théorie du commandement divin.

2. *Le mal engendré par l'action ne doit pas être intentionnel.* C'est considéré comme étant le cas si : soit la conséquence étant mauvaise est une des fins de réaliser l'action, soit elle est un moyen pour arriver à une des fins de réaliser l'action. Mais que veut exactement dire qu'une conséquence est une fin ou un moyen? Par exemple, est-ce qu'une conséquence est un moyen du moment où elle fait partie de la chaîne causale reliant une action et une fin? Ou moyen et fin sont tous deux des concepts prenant en compte les croyances et désirs des agents? Une grande partie des désaccords autour de la doctrine du double effet concernent ce point précis. Déterminer si une conséquence est intentionnelle est un processus indéfini en éthique normative. Ce processus fait donc partie des propositions que peuvent faire les approches en éthique computationnelle.
   Nous notons $\pi_{inte}(d, f, \chi)$ le processus dont le rôle est de déterminer si une conséquence est intentionnelle. $inte$ correspond à l'abréviation de « intentionnelle ». Formellement :
   $$\pi_{inte} : \mathbb{D} \times \mathbb{F}_e \times 2^{\mathbb{F}_e} \to \{1, 0\}.$$

3. *Il y a une raison proportionnellement suffisante pour engendrer le mal.* C'est le cas s'il n'existe pas d'action alternative qui apporterait autant de bien sans apporter autant de mal et que le mal engendré n'est pas disproportionné par rapport au bien recherché. Contrairement aux théories conséquentialistes modélisées précédemment, il n'est pas clairement défini comment doit se faire la quantification du bien et du mal apporté. Il n'est pas non plus clairement défini comment évaluer si le mal engendré est disproportionné par rapport au bien recherché. Déterminer s'il y a une raison proportionnellement suffisante pour engendrer le mal est un processus indéfini en éthique normative. Ce processus fait donc partie des propositions que peuvent faire les approches en éthique computationnelle.
   Nous notons $\pi_{prop}\left(\pi_{cons}(d, \chi), \pi_{cons}(d', \chi), \Upsilon\right)$ le processus dont le rôle est de déterminer s'il y a une raison proportionnellement suffisante pour engendrer le mal. $prop$ correspond à l'abréviation de « proportionnelle ». Formellement :
   $$\pi_{prop} : 2^{\mathbb{F}_e} \times 2^{\mathbb{F}_e} \times \mathbb{V}_\Upsilon{}^{\mathbb{F}_e} \to \{1, 0\}.$$

Nous proposons de modéliser la doctrine du double effet de la façon suivante :

Définition 4.12 [*doctrine du double effet dde*]**.** *Soit $d \in \mathbb{D}$ une décision, χ un contexte, $\mathscr{C}$ un code de conduite et les processus indéfinis π_{code}, π_{cons}, π_{inte}, π_{prop} et Υ_{tdn}. La doctrine du double effet que nous notons dde indique que réaliser d n'est pas une violation directe d'une valeur de base ssi cette décision satisfait toutes les conditions suivantes :*

1. $\pi_{code}(d, \chi, \mathscr{C}) \neq n\mathrm{Do}$;
2. $\left\{f \in \pi_{cons}(d, \chi) | \Upsilon_{tdn}(f) = \mathbb{B}_{tdn} \wedge \pi_{inte}(d, f, \chi) = 1\right\} = \emptyset$;
3. $\forall d' \in \overline{d}, \pi_{prop}\left(\pi_{cons}(d, \chi), \pi_{cons}(d', \chi), \Upsilon_{tdn}\right) = 1$.

La doctrine du double effet peut être modélisée comme un processus dde dont le rôle est de déduire si une décision d est une violation directe d'une valeur de base $\mathrm{D}v$ ou ne l'est pas $n\mathrm{D}v$ dans un contexte donné χ. Formellement :

$$dde : \mathbb{D} \times 2^{\mathbb{F}_e} \times \mathbb{V}_\Upsilon{}^{\mathbb{F}_e} \to \{\mathrm{D}v, n\mathrm{D}v\}.$$

Ayant cela, nous proposons de modéliser le principe moral derrière la théorie du droit naturel de la façon suivante :

Définition 4.13 [Θ_{tdn}]**.** *Soit $d \in \mathbb{D}$ une décision, χ un contexte, dde la doctrine du double effet et la théorie de la valeur Υ_{tdn}.*

— $d \in \mathbb{W}_{tdn}$ *si* $dde(d, \chi, \Upsilon_{tdn}) = \mathrm{D}v$ *;*

— $d \in \mathbb{R}_{tdn}$ *sinon.*

De toutes les théories morales que nous avons modélisées, la théorie du droit naturel est celle qui fait intervenir le plus de processus. Il est intéressant de constater que cette théorie utilise des processus qui se retrouvent aussi bien dans des théories axées sur le devoir qu'axées sur la valeur. Cette richesse est ce qui rend difficile de classer cette théorie soit comme une approche déontologique, soit comme une approche conséquentialiste, dans la typologie classique qui est faite des théories morales. Ce cas est un exemple illustrant la clarification qu'apporte la typologie proposée par TIMMONS [2012]. En effet, la théorie du droit naturel faisant intervenir la théorie de la valeur pour déterminer la théorie du juste, cette théorie morale appartient nécessairement à la famille des théories axées sur la valeur.

Exemple 4.1 [suite]**.** *Si l'aiguilleuse évalue sa décision avec Θ_{tdn}, il faut qu'elle évalue si celle-ci satisfait les trois conditions de la doctrine du double effet.*
Aucune des actions individuelles possibles, $pousser$, $tirer$ ou $diffuser$, n'a de raison d'être intrinsèquement interdite. La première condition est donc satisfaite par toutes les décisions.
La deuxième condition étant la plus compliquée à évaluer, intéressons nous en premier à la troisième qui soulève la question de la proportionnalité. Si nous considérons que tout ce qui arrive est une conséquence de la décision de l'aiguilleuse, toutes les décisions qui contiennent à la fois $\overline{tirer}$ et $\overline{pousser}$ sont injustes car il existe des alternatives qui apportent autant de bien, même plus, sans apporter tant de mal puisque les cinq civils sont sauvés. Le même raisonnement nous permet d'éliminer les décisions contenant à la fois $pousser$ et $diffuser$, car cela aurait comme conséquence que des milliers de personnes seraient exposées à des images traumatisantes. Les décisions restantes sont alors :

$$\left\{tirer, \overline{pousser}, \overline{diffuser}\right\}, \left\{\overline{tirer}, pousser, \overline{diffuser}\right\}.$$

La deuxième condition est plus compliquée à évaluer, plusieurs cas sont possibles. Pour commencer restons dans le cas où l'aiguilleuse ignore que le franc tireur menace la vie de la biologiste. L'action $tirer$ n'empêche pas le franc tireur de tuer la biologiste, mais cette mort qui peut être vue comme une mauvaise conséquence n'est pas la fin de l'action puisque l'aiguilleuse ne le souhaite pas, au contraire. Ce n'est pas non plus un moyen pour une fin puisque, dans l'exemple tel qu'il a été décrit, sa mort n'est pas un moyen pour produire quoique ce soit. La décision $\left\{tirer, \overline{pousser}, \overline{diffuser}\right\}$ satisfait donc la deuxième condition. L'action $pousser$ a comme mauvaise conséquence la mort du franc tireur et comme bonne conséquence de sauver la biologiste et les cinq civils. En supposant que l'aiguilleuse est une personne sans mauvaises

intentions, la mort du franc tireur n'est clairement pas la fin de la décision. Toute la problématique est alors de déterminer s'il s'agit d'un moyen pour atteindre la fin, ou s'il s'agit juste d'un « dommage collatéral ». Selon la définition qui sera donnée à ce qu'est un moyen pour une fin, la décision $\left\{\overline{tirer}, pousser, \overline{diffuser}\right\}$ *pourra ou pas satisfaire cette deuxième condition. Dans le cas où l'aiguilleuse saurait que le franc tireur menace la vie de la biologiste, il semble plus sûr de dire que si la mort du franc tireur n'est pas une fin, il s'agit au moins d'un moyen pour sauver la biologiste. Cette décision ne satisfait pas cette deuxième conditions dans ce cas de figure.*

Avant de clore cette section sur la modélisation des théories morales, faisons un point sur l'architecture modulaire et ses avantages. Premièrement, cette architecture nous semble faciliter la compréhension de la structure générale des différentes théories morales. Alors que dans la description philosophique des théories ce qui ressort le plus est l'idée derrière la théorie, cette structure fait ressortir les différents processus à réaliser pour permettre d'évaluer une décision. Deuxièmement, cette architecture permet de repérer des similarités entre théories morales, comme entre le relativisme moral et l'utilitarisme de la règle. Finalement, la décomposition en processus permet d'avoir une idée générale des différents processus qui sont utilisés dans les différentes théories. Le tableau 4.1 fait un récapitulatif de tous les processus utilisés dans nos modélisations. La combinaison de ce tableau avec l'étude comparative, que nous présentons dans la section suivante, peut être vue comme une cartographie détaillée du domaine.

Processus	Description	Théorie morale l'utilisant
Υ	déterminer la valeur des fluents.	Θ_{ua}, Θ_{ue}, Θ_{cs},Θ_{tdn}
$\pi_{code}(d,\chi,\mathscr{C})$	déterminer d'un code de conduite si une décision doit ou non être réalisée, ou aucune des deux.	Θ_{tcd},Θ_{trm},Θ_{ur},Θ_{tdn}
$\pi_{cocu}(\chi)$	déterminer le code de conduite de la culture à laquelle appartient un agent.	Θ_{trm}
$\pi_{hfs}(d,\chi)$	déterminer si un agissement traite l'humanité comme une fin et ne s'en sert pas comme d'un simple moyen.	Θ_{hfs}
$\pi_{maxi}(d,\chi)$	déterminer la maxime selon laquelle une action est réalisée.	Θ_{lu}
$\pi_{lu}\left(\pi_{maxi}(d,\chi),\chi\right)$	déterminer si une maxime est universalisable.	Θ_{lu}
$\pi_{cons}(d,\chi)$	déterminer les conséquences associées à une décision.	Θ_{ua},Θ_{ue},Θ_{cs},Θ_{tdn}
$\pi_{prob}(d,\chi,f)$	déterminer la probabilité que l'agent attribue à la production effective d'une conséquence suite à la réalisation d'une décision.	Θ_{ue}
$\pi_{coid}(2^{\mathbb{F}_e},\Upsilon)$	déterminer le code de conduite idéal pour une théorie de la valeur donnée.	Θ_{ur}
$\pi_{inte}(d,f,\chi)$	déterminer si une conséquence est intentionnelle.	Θ_{tdn}
$\pi_{prop}\left(\pi_{cons}(d,\chi),\pi_{cons}(d',\chi),\Upsilon\right)$	déterminer s'il y a une raison proportionnellement suffisante pour engendrer le mal.	Θ_{tdn}

TABLEAU 4.1 – Récapitulatif des processus utilisés dans la modélisation des théories morales.

4.3 Étude comparative de l'éthique computationnelle normative

Dans cette section nous faisons une étude comparative des travaux en éthique computationnelle normative s'appuyant sur la structure proposée par le cadre commun. La sélec-

tion des travaux pour cette étude repose sur l'état de l'art proposé par TOLMEIJER et collab. [2021] et sur une extension de cet état de l'art. Notre étude comparative peut être vue comme un approfondissement du travail fait dans les états de l'art existants pour un sous-ensemble de travaux, ceux appartenant au sous-domaine que nous avons appelé l'éthique computationnelle normative. Ce sous-domaine est composé de neuf propositions.

Commençons par détailler les critères choisis qui nous ont permis de former la première partie de notre sélection, celle reposant sur l'état de l'art proposé par TOLMEIJER et collab. [2021]. Ce travail propose trois taxonomies permettant de classer cinquante propositions en éthique computationnelle. La première taxonomie permet de classer ces propositions selon la famille à laquelle appartiennent les raisonnements éthiques formalisés. La deuxième taxonomie permet de classer les propositions selon la façon dont elles formalisent les théories : descendante, ascendante ou hybride. La troisième taxonomie permet de classer les propositions selon le formalisme choisi pour leur représentation et implémentation. Notre étude comparative repose sur une seule taxonomie, celle basée sur la famille à laquelle appartiennent les raisonnements éthiques formalisés. Toutefois, nous ne reprenons pas les catégories des états de l'art existants. En effet, celles-ci ne sont pas intéressantes pour notre étude comparative, elles sont moins précises que ce que nous pouvons proposer grâce à notre cadre.

Pour faire notre sélection par rapport aux cinquante propositions répertoriées par TOLMEIJER et collab. [2021], nous avons formé deux sous ensembles et nous avons appliqué un filtre sur leur intersection. Les différentes étapes de ce processus sont illustrées par le tableau 4.2. Le premier sous ensemble est formé grâce à la deuxième taxonomie. Comme mentionné dans la section 1.5, nous nous intéressons uniquement aux approches descendantes. Nous obtenons ainsi un ensemble composé de trente propositions. Le deuxième sous ensemble est formé grâce à la troisième taxonomie. Nous nous intéressons uniquement aux approches utilisant de « la logique », un « raisonnement probabiliste » ou un mélange des deux et éliminons celles qui sont considérées comme basées sur de « l'apprentissage » ou de « l'optimisation ». Nous obtenons ainsi un deuxième ensemble composé de vingt-trois propositions. L'intersection des deux nous donne un ensemble de dix-sept propositions. Pour obtenir notre sélection finale nous appliquons un dernier filtre. Comme nous l'avons mentionné dans la section 1.5, nous nous intéressons à l'éthique computationnelle normative. Autrement dit, nous nous intéressons aux méthodes qui veulent implémenter des théories morales venant de l'éthique normative. Après lecture des différentes propositions, nous ne gardons que celles qui mentionnent explicitement au moins une fois la théorie morale qu'ils formalisent. Nous obtenons alors la première partie de notre sélection, un ensemble de sept propositions.

La deuxième partie de la sélection est une extension de l'état de l'art existant. La référence la plus récente dans [TOLMEIJER et collab., 2021] est de 2020. Nous appliquons leur méthodologie de « recherche automatisée » sur les travaux plus récents. Nous avons donc recherché dans plusieurs bases de données des entrées qui mentionnent une implémentation et le domaine de l'éthique computationnelle. Les termes précis utilisés pour la recherche sont les suivants :

> implementation AND ('machine ethics' OR 'artificial morality' OR 'machine morality' OR 'computational ethics' OR 'roboethics' OR 'robot ethics' OR 'artificial moral agents')

Cette recherche a été réalisée sur sept bases de données différentes et a couvert une période

	Taxonomie		
	2ème	3ème	Filtre
ANDERSON et ANDERSON [2008]	•		
ANDERSON et collab. [2004]	•		
ARKOUDAS et collab. [2005]	•	•	
ASHLEY et MCLAREN [1994]	•		
ATKINSON et BENCH-CAPON [2008]	•		
BERREBY et collab. [2017]	•	•	•
BONNEMAINS et collab. [2018]	•	•	•
BRINGSJORD et TAYLOR [2012]	•	•	•
CERVANTES et collab. [2016]	•		
CLOOS [2005]	•	•	
COINTE et collab. [2016]	•	•	
DEHGHANI et collab. [2008]	•		
DENNIS et collab. [2016]	•	•	
DENNIS et collab. [2015]	•	•	
FURBACH et collab. [2014]		•	
GANASCIA [2007]	•	•	•
GOVINDARAJULU et BRINGSJORD [2017]	•	•	•
LINDNER et collab. [2017]	•	•	•
MADL et FRANKLIN [2015]		•	
MALLE et collab. [2017]		•	
MCLAREN [2003]	•		
MERMET et SIMON [2016]	•	•	
NETO et collab. [2011]	•	•	
PEREIRA et SAPTAWIJAYA [2009]	•	•	•
PONTIER et HOORN [2012]		•	
REED et collab. [2016]	•		
SHIM et collab. [2017]	•		
THORNTON et collab. [2017]	•		
TUFIS et GANASCIA [2015]		•	
TURILLI [2007]	•	•	
VAN DANG et collab. [2017]	•		
VANDERELST et WINFIELD [2018]	•		
VERHEIJ [2016]	•	•	
WALLACH et collab. [2010]		•	
WIEGEL et BERG [2009]	•	•	
WINFIELD et collab. [2014]	•		

TABLEAU 4.2 – Illustration des différentes étapes du processus de sélection parmi les travaux répertoriés par TOLMEIJER et collab. [2021] pour notre étude comparative.

de cinq ans, entre 2018 et 2023. Nous avons décidé de faire chevaucher notre période de recherche avec celle proposée par TOLMEIJER et collab. [2021] pour vérifier que nous pouvions bien trouver les mêmes travaux, ce qui est bien le cas, mais aussi récupérer des articles qui auraient pu être peu visibles à l'époque par leur récente publication. Voici un récapitulatif de notre recherche sur chacune des sept bases de données :

— Nous avons consulté Web of Science en appliquant un filtre de domaine (« computer science artificial intelligence », « computer science interdisciplinary applications », « computer science theory methods », « computer science information systems », « computer science cybernetics »), recherche qui a renvoyé soixante-treize résultats.

— Nous avons consulté Scopus/ScienceDirect en appliquant un filtre de discipline (« computer science », « engineering »), recherche qui a renvoyé soixante-dix-huit résultats.

— Nous avons consulté ACM Digital Library sans filtre, recherche qui a renvoyé cent-cinquante-six résultats.

— Nous avons consulté Wiley Online Library en appliquant un filtre de discipline (« computer science »), recherche qui a renvoyé huit résultats.

— Nous avons consulté la base de données AAAI Publications, recherche qui a renvoyé cinq résultats.

— Nous avons consulté Springer Link en appliquant un filtre de discipline (« computer science »), de domaine (« artificial intelligence ») et de type d'entrée (« article », « conference paper »), recherche qui a renvoyé deux-cent-cinquante-six résultats.

— Nous avons consulté IEEE Xplore sans filtre, recherche qui a renvoyé cent-trente résultats.

Pour obtenir la deuxième partie de notre sélection, nous avons appliqué les mêmes critères que pour la première partie : des propositions descendantes utilisant la logique et/ou un raisonnement probabiliste et qui veulent implémenter des théories morales venant de l'éthique normative. Nous obtenons alors la deuxième partie de notre sélection, un ensemble de deux propositions [LINDNER et BENTZEN, 2018; SINGH, 2022].

Cette section est divisée en deux parties. La section 4.3.1 présente dans les grandes lignes les neuf propositions de notre étude comparative. Puis, la section 4.3.2 décrit cette étude comparative où une section est dédié à chacune de nos catégories. Pour rappel, nous organisons les propositions selon la théorie morale formalisée. Chaque section commence par discuter du respect ou non de la structure générale de la théorie par les propositions, puis rentre dans les détails des processus composant cette structure.

4.3.1 Brève présentation des travaux choisis pour notre étude comparative

Dans cette section nous présentons dans les grandes lignes les neuf propositions composant notre étude comparative. Nous les présentons dans l'ordre dans lequel elles apparaissent dans l'étude comparative, à l'exception de [LINDNER et collab., 2017] qui est présentée plus tôt pour pouvoir introduire une des extensions qu'ils proposent.

BRINGSJORD et TAYLOR [2012] proposent une formalisation de la théorie du commandement divin Θ_{tcd} dans la logique $\mathscr{LRT}^*$. Il s'agit d'une combinaison de logique du premier ordre, de Predicate Calculi et de la logique modale S5 de LEWIS et LANGFORD [1932]. Ils illustrent leur formalisation avec un exemple de robot à application militaire.

BERREBY et collab. [2017] proposent le cadre modulaire $\mathbb{ACE}$ ($\mathbb{A}$ction-$\mathbb{C}$ausality-$\mathbb{E}$thics) permettant de formaliser plusieurs théories morales dans un langage de description d'action très proche de $\mathscr{E}$ que nous avons présenté dans la section 2.2.7 et implémenté en Answer Set Programming (ASP). Ce cadre modulaire compte trois modules : un module d'action $\mathbb{A}$ qui permet de raisonner sur l'action et le changement; un module causal $\mathbb{C}$ qui enrichit les informations sur la description du monde et son évolution en établissant des relations causales entre les actions et leurs conséquences directes comme indirectes; un module éthique $\mathbb{E}$ dédié à déterminer le statut déontique des actions en prenant en compte les informations du contexte données par $\mathbb{AC}$ et des théories de la valeur données en entrée. Un des avantages principaux de ce cadre est qu'il déplace « de manière globale le fardeau du raisonnement moral du programmeur vers le programme lui même ». Cela est possible en faisant en sorte qu'une partie des informations sur le contexte puisse être déduite plutôt que déclarée par le programmeur grâce aux modules $\mathbb{AC}$. Une partie importante de ces modules est indépendante du problème traité, ils ne doivent pas être changés lorsque le problème change. Pour les quelques parties dépendantes du problème, elles peuvent être exclusivement décrites par des faits ce qui ne demande pas une connaissance poussée de l'ASP. Les théories morales formalisées dans ce cadre sont : une théorie qui pourrait s'apparenter au commandement divin Θ_{tcd}, la théorie morale de Kant Θ_{hfs}, l'utilitarisme de l'acte Θ_{ua}, le conséquentialisme satisfaisant Θ_{cs}, l'utilitarisme de la règle Θ_{ur} et une théorie qui pourrait s'apparenter à la théorie du droit naturel Θ_{tdn}. Ils illustrent leurs formalisations avec un dilemme médical qui peut être retrouvé dans l'annexe A.

BONNEMAINS et collab. [2018] proposent un cadre permettant de formaliser plusieurs théories morales dans un formalisme général proche d'un langage de description d'action. Les théories morales formalisées dans ce cadre sont : une théorie qui pourrait s'apparenter au commandement divin Θ_{tcd}, l'utilitarisme de l'acte Θ_{ua} et une théorie qui pourrait s'apparenter à la théorie du droit naturel Θ_{tdn}. Ils illustrent leurs formalisations avec le dilemme du trolley et un dilemme de drone qui peuvent être retrouvés dans l'annexe A.

GANASCIA [2007] montre que la logique non-monotone se prête à formaliser les normes morales qui demandent de pouvoir gérer des exceptions. Pour cela il propose une formalisation en ASP de trois théories morales : la morale Aristotélicienne, la morale de Kant Θ_{lu} et la « theory of principles » de Benjamin Constant. Il illustre ses formalisations avec le dilemme du mensonge qui peut être retrouvé dans l'annexe A.

LINDNER et collab. [2017] proposent le cadre HERA (Hybrid Ethical Reasoning Agents) permettant de formaliser plusieurs théories morales dans une extension des équations structurelles booléennes qu'ils appellent « boolean causal agency model ». Leur formalisme est composé d'un ensemble de variables d'actions, un ensemble de variables de conséquences, un ensemble d'équations structurelles déterminant la valeur des variables de conséquences, une liste d'ensembles d'intentions (un ensemble pour chaque action), une fonction qui attribue une utilité à chaque action et conséquence et un ensemble d'interprétations de l'ensemble d'actions qui peut être vu comme l'ensemble de décisions $\mathbb{D}$ avec la contrainte qu'une seule action peut être faite à la fois. Les théories morales formalisées dans ce cadre sont : l'utilitarisme de l'acte Θ_{ua} et une théorie qui pourrait s'apparenter à la théorie du droit naturel Θ_{tdn}. Ils illustrent leurs formalisations avec le dilemme du trolley, le dilemme du mensonge et un dilemme de bateau qui peuvent être retrouvés dans l'annexe A.

LINDNER et BENTZEN [2018] proposent une extension au cadre HERA (Hybrid Ethical Reasoning Agents) pour pouvoir y formaliser la théorie morale de Kant Θ_{hfs}. Ils appellent

le formalisme de cette extension « Kantian causal agency models ». Comme pour le formalisme dans [LINDNER et collab., 2017], il est composé d'un ensemble de variables d'actions, un ensemble de variables de conséquences, un ensemble d'équations structurelles déterminant la valeur des variables de conséquences, un ensemble d'interprétations de l'ensemble d'actions qui peut être vu comme l'ensemble de décisions $\mathbb{D}$ avec la contrainte que seule une action peut être faite à la fois. En plus de ces points communs, il est composé d'un ensemble de variables de contexte, une liste d'ensemble de buts (un ensemble pour chaque action), un ensemble de patients moraux ou agents et une fonction ternaire d'affectation indiquant si une action ou une conséquence affecte un patient moral positivement ou négativement. Ils illustrent leurs formalisations avec plusieurs dilemmes utilisés par Kant.

SINGH [2022] propose une formalisation de la théorie morale de Kant Θ_{lu} en logique déontique dyadique. Cette logique remplace l'opérateur modal $\Box$ par un opérateur d'obligation O{A|B} qui représente le fait que A est obligatoire dans le contexte B. Elle implémente cela dans le démonstrateur automatique de théorèmes Isabelle/HOL. Elle illustre sa formalisation avec plusieurs dilemmes utilisés par Kant dont le dilemme du mensonge.

PEREIRA et SAPTAWIJAYA [2009] proposent une formalisation d'une théorie qui pourrait s'apparenter à la théorie du droit naturel Θ_{tdn} dans le système ACORDA reposant sur le formalisme « Prospective Logic Programs ». Ils illustrent leur formalisation avec le dilemme du trolley et ses variantes.

GOVINDARAJULU et BRINGSJORD [2017] montrent que la doctrine du double effet dde peut être formalisée en utilisant le « Deontic Cognitive Event Calculus ($\mathscr{DCEC}$) ». Il s'agit d'une logique modale de premier ordre. Plus exactement, c'est un Calcul des Évènements enrichi avec les opérateurs modaux suivants : **K** pour les connaissances, **B** pour les croyances, **D** pour les désirs, **I** pour les intentions et **O** pour les obligations.

4.3.2 Étude comparative structurée par le cadre commun

Dans cette section nous faisons notre étude comparative qui classe les propositions selon la théorie morale formalisée. Celle-ci dédie une section à chacune de nos catégories. Chacune commence par discuter du respect ou non de la structure générale de la théorie formalisée par les propositions, structure décrite dans les définitions 4.4 à 4.13. Ensuite, dans le reste de la section nous rentrons dans les détails des processus composant cette structure. Cette décomposition est possible grâce à l'architecture modulaire que nous avons adopté pour la modélisation des théories morales. Le fait de pouvoir traiter processus par processus facilite considérablement la réalisation de l'étude comparative. En effet, cette grille de lecture permet de s'abstraire des détails techniques de chaque approche pour essayer d'en extraire la vision spécifique derrière l'implémentation choisie du processus.

4.3.2.1 Théorie du commandement divin

Des trois propositions que nous classons dans cette catégorie [BERREBY et collab., 2017; BONNEMAINS et collab., 2018; BRINGSJORD et TAYLOR, 2012], seule celle de BRINGSJORD et TAYLOR [2012] revendique formaliser la théorie du commandement divin. Les deux autres parlent de codes de conduite. Nous avons décidé de les classer ici pour deux raisons. La première est que, comme discuté précédemment, si nous faisons abstraction de l'origine divine du code de conduite, toute approche qui applique simplement un code de conduite supposé donné, partage le même squelette que la théorie du commandement divin. La

deuxième raison est que, les propositions de BERREBY et collab. [2017]; BONNEMAINS et collab. [2018] ont chacune une façon très différente d'appliquer les codes de conduite. Ne pas intégrer ces propositions nous ferait passer à côté de cette diversité.

Étudions les différentes propositions pour le processus indéfini π_{code}. BRINGSJORD et TAYLOR [2012] s'appuient principalement sur leur logique modale. Le code de conduite $\mathscr{C}$ et le contexte χ sont représentés comme des connaissances dans $\mathscr{LRT}^*$. Une décision d est injuste $\pi_{code}(d,\chi,\mathscr{C}) = n\mathrm{D}o$ si $\mathscr{LRT}^*$ indique que ne pas faire d est obligatoire. Elle est juste $\pi_{code}(d,\chi,\mathscr{C}) = \mathrm{D}o$ si $\mathscr{LRT}^*$ indique que faire d est obligatoire.
BERREBY et collab. [2017] ont plutôt une approche conséquentialiste. Pour eux une décision d est injuste $\pi_{code}(d,\chi,\mathscr{C}) = n\mathrm{D}o$ si l'ensemble de conséquences renvoyé par $\pi_{cons}(d,\chi)$ contient un fluent que le code de conduite $\mathscr{C}$ interdit de produire. Elle est juste si elle n'est pas injuste.
BONNEMAINS et collab. [2018] restent à un niveau d'abstraction plus élevé. Ils utilisent une fonction $\mathrm{D}ecision\mathrm{N}ature$ qui n'est rien d'autre qu'une autre façon de nommer π_{code}. Toutefois, par sa définition, $\mathrm{D}ecision\mathrm{N}ature$ semble indépendante du contexte χ. Cela voudrait dire qu'appliquer un code de conduite revient simplement à classer une fois les décisions. Une décision d est juste quel que soit le contexte, $\forall\chi,\ \pi_{code}(d,\chi,\mathscr{C}) = \mathrm{D}o \vee neither$ si $\mathrm{D}ecision\mathrm{N}ature(d) = \mathrm{D}o \vee neither$. Elle est considérée injuste quel que soit le contexte, $\forall\chi,\ \pi_{code}(d,\chi,\mathscr{C}) = n\mathrm{D}o$ si $\mathrm{D}ecision\mathrm{N}ature(d) = n\mathrm{D}o$.
Nous avons donc ici trois propositions très différentes du processus indéfini π_{code}. Alors que BRINGSJORD et TAYLOR [2012] utilisent leur logique modale, BERREBY et collab. [2017] ont une approche plus conséquentialiste. Si pour ces deux approches le contexte joue un rôle dans le résultat du processus fait par π_{code}, cela n'est pas le cas dans la version de BONNEMAINS et collab. [2018]. Pour finir, il est intéressant de noter que dans la version de BERREBY et collab. [2017], π_{code} fait appel au processus π_{cons}.

4.3.2.2 Théorie morale de Kant

Des quatre propositions que nous classons dans cette catégorie, deux [BERREBY et collab., 2017; LINDNER et BENTZEN, 2018] s'intéressent à Θ_{hfs} donc à la définition 4.6 correspondant à la formulation de l'Impératif Catégorique qui fait apparaître les notions de fin et de moyens, et deux [GANASCIA, 2007; SINGH, 2022] s'intéressent à Θ_{lu} donc à la définition 4.7 correspondant à la formulation de l'Impératif Catégorique qui fait apparaître la notion d'universalisation. Dans les deux cas, la structure est respectée par les deux propositions.

Commençons par étudier les propositions qui traitent Θ_{hfs} et son seul processus indéfini π_{hfs}. BERREBY et collab. [2017] proposent une formalisation où une décision d est injuste $\pi_{hfs}(d,\chi) = 0$ si un individu est traité comme un moyen par cette décision et que la raison pour laquelle il a été traité comme un moyen n'en est pas une fin. Ils considèrent qu'un individu est traité comme un moyen par une décision si dans l'ensemble de conséquences renvoyé par $\pi_{cons}(d,\chi)$ il y a un évènement e et que celui-ci a dans ses effets directs un fluent qui affecte l'individu. Ils considèrent que la raison pour laquelle il a été traité comme un moyen n'est pas une fin, si l'évènement e n'est pas une fin de d, information qui est supposée donnée en entrée. La décision est juste si elle n'est pas injuste.
LINDNER et BENTZEN [2018] proposent une formalisation reposant sur la même idée générale : pour qu'une décision soit juste, soit elle ne traite pas les individus comme des moyens, soit tous les individus traités comme des moyens sont aussi des fins. Toutefois, ces pro-

positions divergent dans la définition de ce qui sera considéré comme moyen et fin. Cette deuxième proposition présente deux définitions pour considérer qu'un individu est traité comme un moyen. Une d'elles est exactement la même que celle proposée par BERREBY et collab. [2017]. La deuxième définition est plus contraignante. Ils considèrent qu'un individu est traité comme un moyen par une décision si dans l'ensemble de conséquences renvoyé par $\pi_{cons}(d,\chi)$ il y a un évènement e qui affecte l'individu et que dans l'ensemble de conséquences renvoyé par $\pi_{cons}(e,\chi)$ il y a une fin de d. Il ne suffit donc pas que la décision d affecte l'individu pour être traité comme moyen, il faut en plus que cela soit une condition pour atteindre un objectif.
LINDNER et BENTZEN [2018] ont également une définition différente et plus contraignante pour considérer que l'individu est traité comme une fin. Ils considèrent que l'individu est une fin de d si au moins un de ses buts affecte l'individu positivement et qu'aucun de ses buts n'affecte l'individu négativement.
Nous avons donc ici deux propositions similaires du processus indéfini π_{hfs}. Celles-ci reposent sur la même idée générale : pour qu'une décision soit juste, soit elle ne traite pas les individus comme des moyens, soit tous les individus traités comme des moyens sont aussi des fins. La différence entre ces propositions se trouve dans la définition de ce qui sera considéré comme moyen et fin. Plusieurs remarques méritent d'être faites. La première est que les deux propositions considèrent que l'information sur les buts d'une décision est donnée en entrée. La deuxième est que pour LINDNER et BENTZEN [2018] la façon dont l'individu est affecté, positivement ou négativement, ne rentre en jeu que dans la définition de quand un individu est traité comme une fin, et non pas dans celle concernant les moyens. Elle n'intervient à aucun moment dans la proposition de BERREBY et collab. [2017]. La troisième remarque est que la définition de fin dans la première proposition utilise l'évènement e qui est utilisé dans la définition de moyens, alors que celle de la seconde proposition ne l'utilise pas. Finalement, la quatrième remarque est que dans les deux propositions présentées, π_{hfs} fait appel au processus π_{cons}.

Étudions maintenant les propositions qui traitent Θ_{lu}. GANASCIA [2007] reste à un niveau d'abstraction élevé. En effet, dans sa formalisation il introduit le prédicat $maxim\_will$ qui indique que chaque agent peut choisir l'ensemble de maximes qu'il souhaite poursuivre du moment où celles-ci respectent l'Impératif Catégorique. Ce prédicat correspond aux deux processus de la théorie $\pi_{lu}\left(\pi_{maxi}(d,\chi),\chi\right)$.
SINGH [2022] propose une formalisation plus détaillée qui traite les processus indéfinis π_{maxi} et π_{lu} individuellement. Commençons par nous intéresser au processus indéfini π_{maxi}. Elle ne dit pas comment former une maxime à partir de la décision d'un agent, mais elle indique bien qu'une maxime est un triplet qui contient une décision, des conditions sur le monde et un objectif, et insiste sur le besoin d'une heuristique pour pouvoir former des maximes. Elle propose une formalisation de trois processus supplémentaires qui doivent faire partie de π_{maxi} : il faut pouvoir vérifier que la maxime est bien formée, qu'elle est désirée et qu'elle est effective. Le premier processus consiste à vérifier que les conditions dans la maxime ne peuvent contenir ni la décision, ni l'objectif. Une maxime n'est pas bien formée si elle est de la forme : « je vais faire d si d afin d'accomplir O » ou « je vais faire d si certaines conditions sur le monde S sont réunies, afin d'accomplir d ». Le deuxième processus consiste à vérifier que dans tous les cas où les conditions sur le monde S sont réunies, alors l'agent va réaliser la décision d. Dans ce cas, la maxime peut être considérée comme désirée par l'agent. Finalement, le troisième processus consiste à vérifier que lorsqu'un agent désire la maxime,

alors le but de la maxime est atteint et si l'agent ne la désire pas, alors il n'est pas atteint. Dans ce cas, la maxime peut être considérée effective.
Passons maintenant à la façon dont SINGH [2022] traite le processus indéfini π_{lu}. Elle propose une formalisation selon laquelle une maxime est dite universalisable si tout le monde désire cette maxime et qu'elle est effective. Pour finir, elle considère qu'une décision d est juste $\pi_{lu}(\pi_{maxi}(d,\chi),\chi) = 1$ si la maxime selon laquelle d est réalisée est universalisable et bien formée.
Nous avons donc ici deux propositions des processus indéfinis π_{maxi} et π_{lu} difficiles à comparer par la différence dans le niveau de détail donné. Aucune des propositions ne rentre dans les détails de comment passer d'une décision à une maxime. Comme nous l'avons mentionné précédemment, ce point suscite de nombreuses discussions.

4.3.2.3 Utilitarisme de l'acte ou plutôt, espéré

Des trois propositions que nous classons dans cette catégorie [BERREBY et collab., 2017; BONNEMAINS et collab., 2018; LINDNER et collab., 2017], deux [BERREBY et collab., 2017; LINDNER et collab., 2017] respectent la structure de Θ_{ua} indiquée dans la définition 4.8. Nous verrons en quoi la proposition de BONNEMAINS et collab. [2018] s'en différencie. Avant de rentrer dans les détails, notez que théoriquement ce que toutes ces théories formalisent correspond plus à de l'utilitarisme espéré que l'utilitarisme de l'acte. Pour rappel, l'utilitarisme de l'acte demande que soient prises en considération toutes les conséquences des décisions, les vrais conséquences qui ont eu lieu, même si celles-ci n'étaient pas attendues. Il s'agit donc d'une théorie qui dans tous les cas ne convient qu'à un raisonnement a posteriori, ce n'est pas une procédure de décision. Dans le cadre de l'éthique computationnelle, il est question de raisonnement a priori. Les conséquences des décisions sur le monde sont simulées. Les conséquences qui peuvent être prises en compte sont donc uniquement les conséquences que l'agent peut envisager, d'où le fait que nous sommes plus dans un utilitarisme espéré. Il peut toutefois être intéressant de parler d'utilitarisme de l'acte pour distinguer le cas où par espéré nous faisons référence à toutes les conséquences qui pouvaient être prévues, du cas où par espéré nous faisons référence uniquement aux conséquences qui étaient désirées. Les propositions de cette catégorie sont dans le premier cas de figure.

Les trois propositions intègrent l'idée qu'une décision d est juste si son utilité est supérieure à celle de toutes les alternatives $\overline{d}$. Ce qui différencie la proposition de BONNEMAINS et collab. [2018] de la structure générale, est la nature de l'utilité et la façon dont elles sont comparées pour déterminer quelle décision a une utilité supérieure. Commençons par étudier le processus indéfini Υ_{ua}.
LINDNER et collab. [2017] restent à un niveau d'abstraction élevé. En effet, dans leur formalisation ils font référence à une fonction d'utilité qu'ils ne détaillent pas, le processus est considéré comme une entrée et aussi bien les fluents que les actions peuvent être valués.
BERREBY et collab. [2017] considèrent également qu'il existe une partie du processus qui est une entrée : ce qui est considéré comme ayant de la valeur intrinsèquement et si les effets directs des actions portent cette valeur ou vont à son encontre. Toutefois, le processus est décomposé en plusieurs étapes et l'utilité intrinsèque n'est pas attribuée aux fluents individuellement, mais directement aux évènements en fonction des effets avec lesquels ils sont formalisés. La première étape est une première valuation où des couples (évènement,valeur) sont classés dans les catégories $\mathbb{G}_\Upsilon$, $\mathbb{B}_\Upsilon$ et $\mathbb{N}_\Upsilon$. Un évènement peut-être considéré comme bon selon une valeur mais neutre ou mauvais pour une autre, un même évène-

ment peut donc se retrouver dans des couples dans des catégories différentes. La deuxième étape consiste à faire une deuxième valuation où chaque couple (évènement,valeur) se voit attribuer un poids. Celui-ci dépend de l'importance de la valeur et du nombre de personnes que l'évènement affecte selon cette valeur; il peut également être pondéré par rapport à l'importance attribuée aux personnes. Ce dernier élément n'est pas utilisé dans l'utilitarisme de l'acte puisque, pour rappel, celui-ci est impartial. Toutefois, cela permet de formaliser des conséquentialismes qui seraient agent-relatifs ou priorisants. La dernière étape consiste à attribuer un poids à chaque évènement en agrégeant le poids de tous les couples dont cet évènement fait partie.
BONNEMAINS et collab. [2018] restent à un niveau d'abstraction élevé et considèrent qu'il s'agit d'une entrée. Toutefois, ce qui fait toute la différence avec les deux autres propositions c'est que la valuation n'attribue a aucun moment une valeur numérique mais se contente d'établir une relation de préférence entre les fluents. Cette relation est considérée comme asymétrique, transitive et non réflexive, il s'agit d'un ordre strict. Ils étendent cette relation aux ensembles de fluents. Ils introduisent ensuite l'idée que certains fluents ne peuvent pas être ordonnés par cette simple relation par leur nature très différente. Par exemple, il n'est pas possible d'ordonner avec cette relation des fluents liés à la vie humaine et des biens matériels. Ils introduisent donc une autre relation de préférence, cette fois-ci entre nature de fluents. Un fluent traitant de vie humaine sera alors toujours supérieur à celui concernant un bien matériel et ce quelle que soit son importance parmi les fluents de même nature. Nous retrouvons là une théorie de la valeur plus proche de celle de MILL [1863] que de BENTHAM [1789]. Pour finir, ils introduisent une fonction qui permet de classer les fluents en deux catégories selon leur valeur, bons ou mauvais.
Nous avons donc ici trois propositions différentes du processus indéfini Υ_{ua}. Alors que pour BERREBY et collab. [2017] et LINDNER et collab. [2017] la sortie du processus est une valeur numérique, pour BONNEMAINS et collab. [2018] la sortie est une relation de préférence entre les fluents. Malgré cette différence principale, les trois propositions partagent le fait que le résultat de ce processus est une entrée du système.

Passons maintenant au processus indéfini π_{cons}. LINDNER et collab. [2017] et BONNEMAINS et collab. [2018] proposent la même formalisation, tous les fluents faisant partie de l'état d'arrivé après que la décision ait eu lieu sont considérés comme des conséquences de la décision.
BERREBY et collab. [2017] proposent une formalisation moins crédule de ce processus. Ils considèrent : qu'une décision qui a lieu dans un état a comme conséquence un fluent si celui-ci fait partie des effets avec lesquels il a été formalisé et que le fluent est vrai à l'état suivant; qu'un fluent qui est vrai dans un état a comme conséquence un évènement si le fluent fait partie de ses préconditions et que l'évènement a lieu dans cet état; qu'une décision qui a lieu dans un état a comme conséquence un évènement qui a lieu dans un autre état, s'il existe un évènement ou fluent intermédiaire qui a comme conséquence l'évènement et comme cause la décision.
Nous avons donc ici deux visions très différentes du processus indéfini π_{cons}. Alors que LINDNER et collab. [2017] et BONNEMAINS et collab. [2018] proposent une version « permissive » qui va attribuer un grand nombre de conséquences à chaque décision, BERREBY et collab. [2017] proposent une version plus sélective. Cette dernière demande de s'intéresser à la causalité effective.

Finissons par nous intéresser à la façon dont est calculée l'utilité et dont les décisions

sont comparées. BERREBY et collab. [2017] et LINDNER et collab. [2017] suivent la structure de la définition 4.8, les valeurs numériques attribuées à chaque conséquence de la décision sont sommées pour donner l'utilité. Chaque décision ayant une utilité totale, leur comparaison est alors triviale.
BONNEMAINS et collab. [2018] ne suivent pas cette même procédure, ils ne déterminent pas une utilité totale. Pour rappel, ils n'attribuent pas de valeur numérique aux fluents ou aux évènements, mais établissent un ordre entre les fluents. Ils considèrent qu'une décision est supérieure à toutes les autres, et donc juste d'après l'aspect maximisant de Θ_{ua}, si les conséquences bonnes de la décision sont supérieures aux conséquences bonnes de chaque alternative et que les conséquences mauvaises de la décision sont inférieures aux conséquences mauvaises de chaque alternative. Cette proposition peut être vue comme une version plus contraignante de l'utilitarisme de l'acte, il se peut qu'aucune décision ne remplisse ces conditions et donc qu'aucune décision ne soit considérée comme juste.

4.3.2.4 Conséquentialisme satisfaisant

Parmi notre sélection, seul BERREBY et collab. [2017] cherchent à formaliser la théorie morale Θ_{cs} de la définition 4.10. Cette proposition peut être vue soit comme une version réduite du conséquentialisme satisfaisant, soit comme une version plus stricte de celui-ci. En effet, nous y retrouvons le cas où une décision d est juste si son utilité est supérieure à un seuil fixé, $\sum_{f\in\pi_{cons}(d,\chi)} \Upsilon_{ua}(f) \geq \tau$. Toutefois, il manque le cas où l'utilité d'aucun décision ne satisfait le seuil et donc la décision d peut être juste si son utilité est supérieure à l'utilité de toutes les décisions alternatives, $\left(\sum_{f\in\pi_{cons}(d,\chi)} \Upsilon_{ua}(f) < \tau\right) \wedge (d \in \mathbb{R}_{ua})$. Il s'agit du cas où l'aspect maximisant de l'utilitarisme de l'acte est réintroduit ce qui permet d'obtenir au moins une décision juste dans tous les cas. Dans cette proposition il se peut donc qu'aucune décision ne soit considérée comme juste.

4.3.2.5 Utilitarisme de la règle

Parmi notre sélection, seul BERREBY et collab. [2017] cherchent à formaliser la théorie morale Θ_{ur} de la définition 4.11. Cette proposition ne correspond pas à la structure proposée dans cette définition, elle n'intègre pas la construction d'un code de conduite idéal, mais s'arrête à l'évaluation de règles individuelles. Pour BERREBY et collab. [2017], la décision d est injuste si celle-ci peut être vue comme l'application d'une règle morale et que l'utilité de cette règle est inférieure à un seuil fixé à zéro. L'utilité de la règle est calculée en simulant que tous les individus agissent selon cette règle et en sommant l'utilité des décisions qui en découlent. D'une certaine façon cette version remplace l'aspect maximisant qui peut être retrouvé dans l'utilitarisme de la règle classique et formalise une version satisfaisante. Pour être plus précis, nous pourrions dire que ce que BERREBY et collab. [2017] modélisent est un conséquentialisme hédoniste, agent-neutre, non priorisant, indirect (au lieu de direct), effectif, universaliste et satisfaisant (au lieu de maximisant).

4.3.2.6 Théorie du droit naturel ou plutôt, doctrine du double effet

Des cinq propositions que nous classons dans cette catégorie [BERREBY et collab., 2017; BONNEMAINS et collab., 2018; GOVINDARAJULU et BRINGSJORD, 2017; LINDNER et collab., 2017; PEREIRA et SAPTAWIJAYA, 2009], aucune ne revendique formaliser la théorie du droit

naturel. En effet, toutes les théories parlent de doctrine du double effet mais aucune de théorie du droit naturel. Nous avons décidé de les classer ici car, comme le montre la définition 4.13, cette théorie repose exclusivement sur le processus dde consistant à appliquer la doctrine du double effet. Toutefois, il est important de préciser que si la doctrine du double effet semble pouvoir être un principe moral à elle seule, la déconnecter de son origine n'est pas sans conséquences. En effet, comme nous l'avons vu dans la section 1.3.2, l'absolutisme morale que nous retrouvons dans la première condition de la définition 4.12 est étroitement lié à la théorie du droit naturel. Nous verrons par la suite comment chaque proposition a fait face à cette problématique. Des cinq propositions, quatre contiennent tous les éléments composant la doctrine du double effet, seuls PEREIRA et SAPTAWIJAYA [2009] proposent une version réduite sans absolutisme moral.

Commençons par étudier comment les quatre autres propositions formalisent cette première condition liée à l'absolutisme moral. BERREBY et collab. [2017] et LINDNER et collab. [2017] proposent une approche plus axée sur la théorie de la valeur. Pour BERREBY et collab. [2017] la décision d échoue cette première condition $\pi_{code}(d,\chi,\mathscr{C}) = n\mathrm{D}o$ si la première valuation, où des couples (évènement,valeur) sont classés dans les catégories $\mathbb{G}_\Upsilon$, $\mathbb{B}_\Upsilon$ et $\mathbb{N}_\Upsilon$, classe un des couples dont la décision fait partie comme mauvais. Dans le cas contraire elle satisfait la condition. Pour LINDNER et collab. [2017] la décision d échoue cette première condition $\pi_{code}(d,\chi,\mathscr{C}) = n\mathrm{D}o$ si l'utilité de d est strictement négative. Dans le cas contraire elle satisfait la condition.
GOVINDARAJULU et BRINGSJORD [2017] et BONNEMAINS et collab. [2018] proposent une approche plus axée sur la théorie du juste. GOVINDARAJULU et BRINGSJORD [2017] s'appuient principalement sur leur logique modale. Ils considèrent que la décision d satisfait cette première condition $\pi_{code}(d,\chi,\mathscr{C}) \neq n\mathrm{D}o$ s'il est établi qu'il n'est pas obligatoire que l'action n'ait pas lieu. Dans le cas contraire elle échoue la condition. BONNEMAINS et collab. [2018] restent à nouveau à un niveau d'abstraction plus élevé et, comme pour la théorie du commandement divin, utilisent la fonction $\mathrm{D}ecision\mathrm{N}ature$. La décision d satisfait cette première condition quel que soit le contexte $\forall\chi$, $\pi_{code}(d,\chi,\mathscr{C}) = \mathrm{D}o \vee neither$ si nous avons $\mathrm{D}ecision\mathrm{N}ature(d) = \mathrm{D}o \vee neither$.
Nous avons donc ici des propositions très différentes du processus indéfini π_{code}. D'un côté, BERREBY et collab. [2017] et LINDNER et collab. [2017] proposent une approche plus axée sur la théorie de la valeur, où la version proposée ici par BERREBY et collab. [2017] est différente de celle dans leur proposition pour la théorie du commandement divin. Toutefois, il est intéressant de noter que cette version de π_{code} fait également appel au processus π_{cons}. De l'autre côté, nous avons la proposition de BONNEMAINS et collab. [2018] qui est la même que pour la théorie du commandement divin et la proposition de GOVINDARAJULU et BRINGSJORD [2017] qui s'appuie principalement sur le langage modal utilisé, comme BRINGSJORD et TAYLOR [2012] dans la théorie du commandement divin.

Passons maintenant à la deuxième condition, celle liée à l'intentionnalité et donc au processus π_{inte}. Nous pouvons distinguer trois visions différentes. PEREIRA et SAPTAWIJAYA [2009] proposent une formalisation qui repose principalement sur une entrée indiquant quels sont les objectifs poursuivis et donc si une conséquence est intentionnelle ou non.
BERREBY et collab. [2017] et BONNEMAINS et collab. [2018] supposent que l'information d'intentionnalité n'est pas disponible et vont plutôt essayer de la déduire à partir de relations causales. Pour BERREBY et collab. [2017] la décision d échoue cette deuxième condition si elle a pour conséquence un évènement considéré comme mauvais par la première valua-

tion des couples (évènement,valeur) et que cet évènement a pour conséquence un évènement considéré comme bon par la première valuation. Le mal engendré est alors considéré comme étant un moyen pour produire le bien. Dans le cas contraire elle satisfait la condition. BONNEMAINS et collab. [2018] adoptent la même idée, sauf que la relation de causalité est établie différemment, ils font appel à la nécessité forte. L'occurrence de l'évènement mauvais doit entraîner l'occurrence de l'évènement bon dans tous les futurs possibles. Pour cela ils utilisent l'opérateur F de la logique modale temporelle linéaire. Seulement à cette condition le mal engendré est considéré comme étant un moyen pour produire le bien.
GOVINDARAJULU et BRINGSJORD [2017] et LINDNER et collab. [2017] formalisent la même vision de cette deuxième condition. Il s'agit des formalisations de l'intentionnalité les plus complexes, elles regroupent la vision de PEREIRA et SAPTAWIJAYA [2009] et de BERREBY et collab. [2017] et BONNEMAINS et collab. [2018]. D'un côté ils prennent en compte l'information sur les intentions de l'agent. LINDNER et collab. [2017] supposent que cette information est donnée en entrée, alors que GOVINDARAJULU et BRINGSJORD [2017] revendiquent pouvoir la déduire des connaissances de l'agent, de ses croyances et de ses obligations. Ces deux propositions considèrent que la décision d satisfait une première partie de la deuxième condition si l'agent a l'intention de produire au moins une des bonnes conséquences de la décision et aucune des mauvaises. D'un autre côté ils prennent en compte l'information causale. Ils considèrent que la décision d satisfait la partie restante de la deuxième condition si aucune des bonnes conséquences de la décision n'est une cause d'au moins une des mauvaises conséquences de la décision. Ils adoptent une vision causale plus proche de celle de BONNEMAINS et collab. [2018] que de BERREBY et collab. [2017] puisqu'ils utilisent le but-for test pour déterminer les relations causales, et font donc intervenir la nécessité forte.
Nous avons donc ici trois visions différentes du processus indéfini π_{inte}. La première est celle de PEREIRA et SAPTAWIJAYA [2009] qui considèrent que l'intentionnalité est une information qui est donnée en entrée. La deuxième est celle de BERREBY et collab. [2017] et BONNEMAINS et collab. [2018] qui ne partagent pas cette même hypothèse. Pour eux, l'information d'intentionnalité n'est pas disponible. Ils vont donc essayer de la déduire avec un raisonnement causal. Il est intéressant de noter que cette version de π_{inte} fait donc appel au processus π_{cons}. La troisième vision est celle de GOVINDARAJULU et BRINGSJORD [2017] et LINDNER et collab. [2017] qui regroupent les deux premières visions pour faire un version plus complexe : il faut prendre en compte certaines informations sur l'intentionnalité données en entrée et utiliser une raisonnement causal pour en déduire d'autres.

Pour finir, intéressons nous à la troisième condition, celle liée à la proportionnalité. Trois visions peuvent encore être identifiées. PEREIRA et SAPTAWIJAYA [2009] sont les seuls à proposer une formalisation qui considère les alternatives. Ils considèrent que la décision d satisfait cette troisième condition si la décision d est celle avec l'utilité totale la plus élevée. Nous retrouvons là une vision maximisante. Toutes les autres propositions comparent les conséquences négatives et les conséquences positives de la décision évaluée uniquement.
BERREBY et collab. [2017], GOVINDARAJULU et BRINGSJORD [2017] et LINDNER et collab. [2017] calculent l'utilité totale de la décision et considèrent que la décision d satisfait cette troisième condition si la décision d a une utilité totale supérieure à un seuil. Nous retrouvons là une vision proche à celle du conséquentialisme satisfaisant. BERREBY et collab. [2017] et LINDNER et collab. [2017] fixent ce seuil à zéro, GOVINDARAJULU et BRINGSJORD [2017] ne le précisent pas.
BONNEMAINS et collab. [2018] comparent les conséquences négatives et les conséquences

positives de la décision évaluée uniquement, mais ne passent pas par l'utilité totale. Pour rappel, ils n'attribuent pas de valeur numérique aux fluents ou aux évènements, mais établissent un ordre entre les fluents. Ils considèrent que la décision d satisfait cette troisième condition si les conséquences négatives de la décision d ont une importance faible ou au plus égale par rapport à ses conséquences négatives. Pour cette évaluation ils introduisent un opérateur de proportionnalité entre fluents qu'ils étendent aux ensembles de fluents. Cet opérateur est réflexif et transitif, mais n'est ni symétrique, ni asymétrique, cela dépend des cas.

Concluons cette étude comparative par quelques perspectives. La sélection que nous avons obtenue n'est qu'une première étape, une étude comparative beaucoup plus large peut être réalisée en assouplissant certaines contraintes. En l'occurrence, il serait possible de prendre en compte des propositions qui ne font pas explicitement mention à une théorie morale. PAGNUCCO et collab. [2021] et CARDOSO et collab. [2022] proposent des structures générales permettant de formaliser des théories morales axées sur le devoir et sur la valeur, sans en mentionner une explicitement. Une extension possible de cette étude comparative serait de ne pas classer les approches par théorie morale implémentée, mais d'aller à un niveau de granularité plus fin et s'intéresser aux processus indéfinis. Au lieu de parler de théorie du droit naturel, nous aurions une catégorie explorant les différentes façons dont les travaux formalisent un des processus la composant, même si cela est en dehors d'une théorie morale.
Il serait également possible de prendre en compte des propositions qui sont des extensions de travaux précédents qui s'intéressent à des aspects techniques précis et qui par conséquent laissent de côté l'idée de traduire des théories morales spécifiques. Par exemple, LINDNER et collab. [2020] explorent les façons dont leurs précédents travaux peuvent s'appliquer à des plans et non pas à des actions individuelles. DENNIS et collab. [2021] étudient comment les changements dans le contexte ont une influence sur les résultats de la théorie morale une fois formalisée.
Finalement, il serait possible de considérer les propositions hybrides dont la partie descendante est dominante dans l'étape de formalisation de la théorie morale et qui utilisent la partie ascendante pour l'obtention d'entrées du processus Θ comme la représentation du contexte χ ou une partie de la théorie de la valeur Υ. Les travaux de KIM et collab. [2021] font partie de ce type de propositions.

4.4 Causalité, pièce fondamentale à l'édifice

Dans cette section nous défendons que traiter la causalité effective dans toute sa complexité est nécessaire à la formalisation de la plupart des théories morales, et donc à l'éthique computationnelle. En effet, une partie des limites actuelles des approches en éthique computationnelle sont dues à l'absence systématique, d'un mécanisme permettant d'établir des relations causales complexes.

Pour commencer, faisons abstraction de l'aspect éthique. Tout système de prise de décision fondé sur une représentation du monde doit inévitablement traiter plus ou moins en profondeur la question de la causalité. En effet, la compréhension rationnelle de l'évolution du monde physique est intrinsèquement liée à l'idée de causalité. Toutefois, comme nous l'avons vu dans le chapitre 3, traiter la causalité dans toute sa complexité n'est pas simple, son intégration dans la planification l'est encore moins. Dans le chapitre 1 nous avons vu

que l'éthique était sensible au contexte. Autrement dit, le statut déontique d'une action dépend en partie de faits non moraux reliés au contexte. Formaliser le raisonnement éthique demande donc d'avoir une représentation du monde dans laquelle le raisonnement éthique s'inscrit. La question qui se pose alors est : avec quel niveau de profondeur la question de la causalité doit-elle être traitée en éthique computationnelle ?

La causalité effective est une brique nécessaire à la formalisation de la plupart des théories morales, en dehors de l'aspect purement représentation du monde. L'importance de la causalité paraît évidente lorsqu'il est question de modéliser des théories morales conséquentialistes, c'est-à-dire qui adhèrent à l'idée que la valeur des conséquences associées à une action est la seule chose à prendre en compte pour déterminer si une action est juste. En l'occurrence, les théories modélisée dans les sections 4.2.2.1, 4.2.2.2, 4.2.2.3 et 4.2.2.4 sont toutes conséquentialistes. Paradoxalement, comme le montre l'étude comparative, même dans la modélisation de ces théories, la causalité se voit rarement accorder la place qu'elle mérite. Comme nous le verrons par la suite, cela se traduit par la proposition d'un processus π_{cons} ne pouvant gérer que les cas les plus simples. Comme le montre également l'étude comparative, la causalité peut également être nécessaire lorsqu'il s'agit de déterminer si une action est un moyen pour une fin dans la deuxième formulation de l'Impératif Catégorique modélisée par Θ_{hfs} dans la section 4.2.1.3, ou pour déterminer si une conséquence est intentionnelle dans la doctrine du double effet essentielle pour la théorie du droit naturel modélisée dans la section 4.2.2.5. Bien que le lien semble moins intuitif, la causalité peut intervenir dans d'autres théories axées sur le devoir qui utilisent des codes de conduite, comme celles des sections 4.2.1.1 et 4.2.1.2. Comme il a été mentionné dans la section 4.2.1, au sein d'un code de conduite il est possible de trouver des normes morales de tout type. En particulier, il est tout à fait envisageable d'avoir des normes qui demandent de considérer des conséquences, comme : « il est interdit d'utiliser des substances du moment où elles causent des dommages environnementaux à l'eau ». Même si la causalité ne joue pas un rôle central dans toutes les théories morales, elle peut être amenée à intervenir dans toutes. Cette omniprésence lui confère une importance particulière.

Traiter la causalité effective dans toute sa complexité est nécessaire à la formalisation de certaines théories morales. L'absence d'un processus permettant cela a deux conséquences principales sur les propositions faites jusqu'à présent en éthique computationnelle : (i) la simplification excessive de la représentation des problèmes et (ii) le fait de systématiquement laisser de côté des problèmes de surdétermination comme les exemples 3.3 et 3.4.

La simplification excessive de la représentation des problèmes tient à l'impossibilité de créer des liens de causalité complexes. En l'absence de telles relations, la seule façon de relier une action à une conséquence est soit de la représenter comme un effet intrinsèque de l'action, soit d'évaluer un état initial et un état final et d'attribuer tous les changements à l'action. Si nous considérons le problème de l'exemple 4.1, la première stratégie de simplification correspond à représenter l'action $pousser$ comme ayant pour effet intrinsèque de tuer le franc tireur et toute décision ayant à la fois $\overline{pousser}$ et $\overline{tirer}$ aura comme effet intrinsèque la mort des cinq individus sur la voie principale. C'est ce que nous retrouvons dans la formalisation de la dde de Pereira et Saptawijaya [2009]. La seconde stratégie de simplification prend l'état initial où tous les individus sont vivants, le compare à l'état final où certains individus sont morts, et considère que tous les changements sont des effets de la décision réalisée entre les deux. C'est ce que nous retrouvons dans la formalisation de Θ_{ua} de Lindner et collab. [2017] et Bonnemains et collab. [2018] ou dans la modélisation générale du raisonnement conséquentialiste de Limarga et collab. [2020]. Cela revient à considérer que l'action $pousser$ a comme conséquence la mort du franc tireur, l'action $tirer$ a comme conséquence la mort de la biologiste et toute décision ayant à la fois $\overline{pousser}$ et $\overline{tirer}$ a comme conséquence la mort des cinq individus sur la voie principale. Ces raccourcis sont apparus comme des solutions viables car les problèmes étudiés sont pour la plupart

des expériences de pensée. Un outil philosophique très utile pour nous aider à comprendre l'asymétrie de notre jugement dans certaines situations, entre autres. Par exemple, pourquoi il serait acceptable de sacrifier une personne pour en sauver cinq dans un contexte donné, mais pas dans un autre très similaire [NYHOLM et SMIDS, 2016] ? Ainsi, dans les problèmes étudiés, le nombre de facteurs pouvant être pris en considération est très faible, les résultats sont certains et le nombre d'actions possibles est très limité, un contexte éloigné de celui du monde réel. Cette apparente simplicité a donc masqué la complexité d'un des défis du domaine : saisir la subtilité de chaque problème dans sa représentation. Ces raccourcis sont d'autant plus inappropriés qu'ils peuvent avoir une influence négative sur l'évaluation éthique du problème.

Conformément à la première stratégie de simplification, l'action $pousser$ a pour effet intrinsèque de tuer le franc tireur. Si nous évaluions ces décisions avec la théorie du relativisme moral Θ_{trm} reposant sur un code de conduite indiquant qu'aucune action intrinsèquement mauvaise ne peut être réalisée, et que par « intrinsèquement mauvaise » ce code entend toute action dont un effet direct est de tuer un être sentient, alors l'action $pousser$ serait considérée injuste. L'action $pousser$ serait en quelque sorte assimilée au fait de tirer sur quelqu'un avec une arme. Le lecteur comprendra rapidement que cela peut être problématique et qu'il s'agit simplement d'un effet de bord de la représentation plus que d'un choix ou une réalité.

Conformément à la deuxième stratégie de simplification, toutes les décisions ont pour conséquence la mort d'au moins un individu. Si nous évaluions ces décisions avec la théorie du relativisme moral Θ_{trm} reposant sur un code de conduite indiquant qu'aucune décision ôtant la vie à un individu ne peut être réalisée, alors aucune décision ne serait juste. Pourtant, il est possible d'imaginer des nuances et d'interroger la capacité de cette deuxième stratégie à gérer correctement l'imputabilité. Est-ce que par le fait que tirer le levier dévie le train sur la voie vide et donc n'empêche pas le franc tireur de tuer la biologiste nous voulons considérer que cette action a pour effet d'ôter la vie à une personne ? L'attribution de ce décès au fait que l'agent ait dévié le train peut s'avérer problématique. C'est pourtant ce qui se produit inévitablement lorsque la causalité est réduite à une simple comparaison d'états. Voulons nous considérer cette action de la même façon que ne rien faire et laisser les cinq civils mourir ou que dévier le train pour tuer le franc tireur ? Il s'agit d'une question complexe qui relève de la causalité négative dont nous avons soulevé les difficultés dans la section 3.2.2. En ayant recours à cette simplification, nous n'avons tout simplement pas la possibilité de représenter toutes les nuances possibles. Ces cas deviennent inévitables lorsque d'autres agents peuvent agir sur le monde et qu'il est possible d'avoir des évènements qui se produisent simultanément, des situations qui sont loin d'être exceptionnelles et que l'éthique computationnelle ne peut pas simplement ignorer.

La solution aux effets indésirables de ces deux stratégies de simplification consiste à représenter la dynamique du problème. Par exemple, il serait plus approprié de représenter les actions $tirer$ et $pousser$ comme ayant un effet intrinsèque sur le mouvement du trolley et non sur la vie des individus. Dans ce cas de figure, les actions de l'agent ne sont plus directement liées à la mort des individus, d'où le besoin essentiel d'être capables d'établir des relations causales complexes. Autrement dit, la solution au problème de ramification exige en éthique computationnelle un raisonnement causal plus complexe par le besoin d'un lien entre l'agent et la conséquence pour établir une quelconque imputabilité [BEEBEE et collab., 2009] : « no blame or praise may be assigned without some account of causal relationship

between an agent and an outcome ».

Le fait de systématiquement laisser de côté des problèmes de surdétermination tient aux difficultés qui en découlent et qui ne peuvent être surmontées qu'en traitant la causalité effective dans toute sa complexité. Comme nous l'avons vu dans le chapitre 3, de nombreuses situations dont il est important de connaître les causes sont presque inévitablement des cas de surdétermination. Nous avions pris en exemple le réchauffement climatique, la pollution d'un site protégé, la délocalisation d'un site industriel, l'inégalité de représentation en politique entre les femmes et les hommes, l'existence de déserts médicaux, une perte économique pour une filière agricole ou le suicide d'une personne. Dans la mesure où la surdétermination est un phénomène auquel peut être confronté un agent dans un contexte réel, l'incapacité de toutes les approches en éthique computationnelle à traiter ces cas est une limitation importante. Les quelques propositions qui mettent en place un mécanisme pour déduire des relations causales, en formalisant un but-for test par exemple, rendent impossible la représentation de certains problèmes de surdétermination dans leur formalisme. Cela est fait en ne permettant pas d'avoir des préconditions disjonctives ou en ne permettant pas de représenter l'occurrence simultanée de deux évènements. Cela diminue encore le nombre de cas courants auxquels un agent peut être confronté qui peuvent être traités dans le domaine. Ainsi, pour atteindre l'objectif souhaité, il est non seulement nécessaire que les approches du domaine puissent établir des relations causales, mais aussi qu'elles puissent en établir des suffisamment complexes pour traiter les cas de surdétermination causale.

4.5 Conclusion

Dans ce chapitre nous avons présenté deux contributions au domaine de l'éthique computationnelle. D'un côté nous avons proposé un cadre commun pour la formalisation du raisonnement éthique qui propose une modélisation des concepts essentiels et des principales théories morales définis en éthique normative. D'un autre côté, en nous appuyant sur les avantages de ce cadre commun, nous avons proposé une étude comparative plus détaillée que les états de l'art existants sur les propositions faites en éthique computationnelle normative.

Le cadre que nous avons proposé a été conçu de sorte à contenir tous les concepts essentiels à l'éthique normative tout en étant suffisamment général pour englober les formalisations déjà proposées dans le domaine.
Pour commencer, nous avons présenté ses entrées et sorties. Nous avons vu que les entrées générales à toutes les théories sont le contexte et un ensemble de décisions possibles. Pour les théories axées sur la valeur nous avons défini une entrée supplémentaire, la théorie de la valeur. Puis, nous avons choisi pour sortie de notre cadre l'évaluation la plus courante en éthique normative, le statut déontique de la décision.
Nous avons ensuite proposé une modélisation de plusieurs théories morales étant aussi bien des théories axées sur le devoir que sur la valeur. Pour rappel, celles-ci ont été choisies de façon à donner l'aperçu le plus large possible des structures existantes. En reprenant ces théories nous espérons avoir proposé une modélisation pour la plupart des structures existantes. Nos modélisations ont été inscrites dans le cadre commun. Elles ont toutes été faites en utilisant une architecture modulaire qui permet d'identifier clairement tous les processus y intervenant. Nous avons distingué d'un côté le squelette clairement défini des théo-

ries, de l'autre les processus qui ne le sont pas. De cette façon, pour s'assurer d'être fidèle à une théorie donnée, toute formalisation peut reprendre la modélisation du squelette que nous proposons, puis proposer une version des processus indéfinis. Identifier clairement les choix réalisés permet ainsi de comparer différentes implémentations d'une théorie plus facilement. C'est ce que nous avons ensuite montré avec notre étude comparative.

L'étude comparative des travaux en éthique computationnelle normative que nous avons proposé peut être vue comme un approfondissement du travail fait dans les états de l'art existants pour un sous-ensemble de travaux, ceux appartenant au sous-domaine que nous avons appelé l'éthique computationnelle normative. Ce sous-domaine est composé de neuf propositions. Nous avons commencé par détailler la procédure de sélection que nous avons suivi. Celle-ci s'appuie pour une partie sur l'état de l'art proposé par TOLMEIJER et collab. [2021] et pour une autre sur une extension de cet état de l'art. Puis, nous avons présenté dans les grandes lignes les neuf propositions composant notre étude comparative. Finalement, nous sommes rentrés dans le vif du sujet en comparant les différentes façons dont les propositions formalisent les différentes théories morales. Cette étude a été structurée en suivant la structure proposée par le cadre commun. Pour commencer, les propositions ont été classées selon la théorie morale formalisée. Une section a été dédiée à l'étude de la formalisation de chacune des théories morales. Nous avons en premier discuté du respect ou non de la structure générale de la théorie par les propositions, structure décrite dans les définitions 4.4 à 4.13. Puis, nous sommes rentrés dans les détails des processus composant cette structure en les étudiant un par un. Un paragraphe a été dédié à chaque processus.
Nous avons observé grâce à cette étude que, dans le meilleur des cas ces formalisations étaient des instantiations de notre cadre, et sinon ce dernier facilite la comparaison en offrant une structure générale de la théorie morale qui sert de grille d'analyse.

Pour finir, dans la dernière section de ce chapitre nous avons défendu que traiter la causalité effective dans toute sa complexité est nécessaire à la formalisation de la plupart des théories morales, et donc à l'éthique computationnelle. Nous avons donné quelques exemples de comment l'absence d'un mécanisme permettant d'établir des relations causales complexes est à l'origine d'une partie des limites actuelles des propositions dans le domaine. En l'occurrence, nous avons montré que cette absence entraîne une simplification excessive de la représentation des problèmes et empêche de traiter tous les problèmes de surdétermination pourtant si courants.
L'importance de la causalité dans le domaine est ce qui justifie le choix que nous avons fait de contribuer au domaine en proposant un mécanisme adapté aux enjeux éthiques qui permette d'établir des relations causales complexes. Ayant montré que la formalisation de la plupart des théories peut faire intervenir un tel mécanisme, il est primordial que celui-ci soit dénué de tout aspect normatif qui pourrait le rendre impropre a être utilisé dans une théorie morale.

Chapitre 5

Contribution : modélisation de la surdétermination en causalité effective

« *Roughly put, the typical setup is to go over some examples for which existing definitions give counterintuitive answers, and then to construct a new definition that does not do so. It is unrealistic to expect that this [...] strategy in and on itself can deliver a satisfactory account of causation, because there are too many examples and even more intuitions.* »

BECKERS [2021a]

Sommaire

Cette thèse se veut une contribution au domaine de l'éthique computationnelle. Dans la section 4.4 il a été montré que la causalité était une pièce fondamentale lorsqu'il est question de formaliser une grande partie des théories morales vues dans le chapitre 1. C'est par le biais de la causalité, principalement, que nous avons choisi d'atteindre cet objectif.

Ce chapitre se veut une clarification du concept de surdétermination en causalité effective. Comme présenté dans la section 3.1.2.1, il s'agit de cas où plus d'une cause pourrait produire le résultat à elle seule. Ces cas sont particulièrement importants pour deux raisons. La première est que de nombreuses situations dont il est important de connaître les causes sont presque inévitablement des cas de surdétermination. Mentionnons par exemple le réchauffement climatique, la pollution d'un site protégé, la délocalisation d'un site industriel, l'inégalité de représentation en politique entre les femmes et les hommes, l'existence de déserts médicaux, une perte économique pour une filière agricole ou le suicide d'une personne.
La deuxième raison est qu'une partie importante des débats encore ouverts sur la notion de causalité ont à voir de près ou de loin avec la surdétermination, et cela que ce soit en philosophie, psychologie, droit, mathématiques ou informatique. Comme l'indique la citation de BECKERS [2021a] en début de chapitre, lorsqu'une nouvelle définition de causalité est proposée, elle est la plupart du temps validée en étant confrontée à des exemples complexes. Ces exemples sont tous des exemples de surdétermination.
Pour ces deux raisons, il est possible de dire que s'intéresser à la causalité dans toute sa complexité passe nécessairement par s'intéresser à la surdétermination.

La clarification que nous proposons consiste en deux étapes : une définition formelle de ce qu'est la surdétermination et une typologie formelle des différents cas de surdétermination. Pour ce faire, nous utilisons un système de transition d'états étiqueté très général. En effet, nous avons vu dans la section 3.2.1.2 qu'une partie des débats encore ouverts peut venir du manque d'expressivité des formalismes classiquement utilisés dans le domaine de la causalité et le manque de définitions formelles lorsqu'il est question de surdétermination. Plus précisément, nous pouvons dire que ces débats en particulier reposent sur ce qui a été défini comme des désaccords causaux non fondamentaux. À travers la typologie que nous proposons, nous espérons que ce type de désaccords n'ait plus lieu d'être.
L'utilisation d'un STEE très général a deux avantages. D'un côté il est très expressif ce qui permet de saisir toutes les subtilités des différents cas de surdétermination. D'un autre, sa généralité fait que les résultats présentés ici peuvent être utilisés par un plus grand nombre d'approches existantes. En effet, du moment où la représentation du monde sur laquelle une approche causale est faite peut être assimilée au STEE utilisé ici, la typologie proposée pourra y être appliquée.

Ce chapitre est divisé en quatre sections. La section 5.1 introduit la représentation de l'action, du changement et de la causalité dans laquelle notre étude de la surdétermination est plongée. Les deux sections suivantes contiennent les contributions formelles principales que nous faisons. La section 5.2 présente la définition de surdétermination et la section 5.2.2 la typologie des cas de surdétermination. Finalement, la section 5.3 discute des résultats qui peuvent être tirés de ces contributions. Les travaux présentés dans ce chapitre font l'objet d'un article soumis en 2024 à la « International Joint Conference on Artificial Intelligence » (IJCAI 2024).

5.1 Modélisation de l'action, du changement et de la causalité pour l'étude de la surdétermination

Dans cette section nous présentons la modélisation du monde dans laquelle s'inscrit cette contribution de clarification. Pour que cette contribution puisse bénéficier au plus grand nombre, nous nous plaçons dans un STEE général et évitons de choisir un langage de description d'action en particulier. Nous le notons $\mathscr{S}_s$, où l'indice s fait référence à la surdétermination. Dans le même esprit, nous introduisons deux types de relations causales entre objets du $\mathscr{S}_s$ sans en donner une définition spécifique. L'objectif recherché est que toute représentation de l'action, du changement et de la causalité suffisamment proche puisse facilement trouver une équivalence avec notre modélisation.

Exemple 5.1 [peloton d'exécution]**.** *Reprenons l'exemple de l'opéra Tosca de Puccini. Dans le troisième acte le peintre et amant de Tosca, Mario Cavaradossi, est exécuté, par ordre de Scarpia, en raison de sa sympathie envers les idées républicaines. Dans la mise en scène de Pierre Audi pour l'Opéra National de Paris, cette exécution est faite au moyen d'un peloton d'exécution, à savoir un groupe de soldats auxquels il est simultanément ordonné de tirer sur le condamné. L'utilisation d'une telle procédure a deux buts : prévenir n'importe lequel des soldats d'empêcher l'exécution et prévenir l'identification du soldat ayant porté le coup fatal. Ce deuxième but fait de cette méthode un cas type de situation où la surdétermination est utilisée pour diluer la responsabilité des agents.*
La figure 5.1 illustre un circuit électrique pouvant correspondre à un peloton d'exécution où le condamné est soumis à la chaise électrique au lieu d'être fusillé. Le circuit est composé d'une batterie, d'un individu attaché et connecté à des électrodes, et de deux interrupteurs montés en parallèle. Nous supposons que quatre agents sont impliqués dans la situation : celui attaché et trois autres, chacun pouvant contrôler les composants du circuit mais ignorant ce que font les autres.

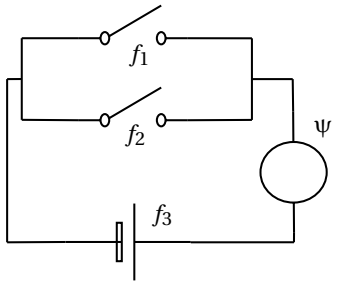


FIGURE 5.1 – Circuit électrique pouvant correspondre à un peloton d'exécution où le condamné est soumis à la chaise électrique au lieu d'être fusillé. Le circuit est composé d'une batterie (f_3), d'un individu attaché et connecté à des électrodes (ψ), et de deux interrupteurs montés en parallèle (f_1, f_2).

Tous les types de surdétermination peuvent être illustrés par l'exemple 5.1, il suffit de combiner la façon dont chaque agent peut agir. Cet exemple est donc utilisé tout au long du chapitre. Des exemples couramment utilisés dans le domaine de la causalité effective sont également utilisés en complément de celui-ci.

Pour traiter n'importe lequel de ces exemples il est nécessaire de les modéliser. C'est ce que nous faisons au moyen du STEE général $\mathscr{S}_s$ que nous introduisons dans la section 5.1.1.

Puis, dans la section 5.1.2 nous introduisons le type de relations causales entre éléments du $\mathscr{S}_s$ nécessaires à l'étude de la surdétermination.

5.1.1 Système de transition d'états étiqueté pour la surdétermination $\mathscr{S}_s$

Le STEE que nous présentons dans cette section adopte une structure classique telle qu'introduite dans la section 2.1. D'un côté, l'état du monde peut être défini comme une collection de variables le décrivant. De l'autre, les transitions entre états du monde est le résultat de l'occurrence d'un ensemble d'évènements. L'évolution du monde peut être vue comme une suite d'états qui s'enchaînent les uns après les autres, au fur et à mesure que des évènements se produisent. Nous notons $\mathbb{F}$ l'ensemble de variables décrivant l'état du monde. Ces variables représentent classiquement les propriétés du monde pouvant varier dans le temps, elles sont appelées des fluents. Puis, nous notons $\mathbb{E}$ l'ensemble de variables décrivant les transitions. Ces variables sont appelées des *évènements*.
Par rapport à la représentation de l'action et du changement extrêmement générale de la section 4.1 notée $\mathscr{S}_e$, nous sommes dans un cas un peu plus spécifique. Dans l'ensemble $\mathbb{F}$ nous faisons le choix de ne garder que des fluents ontiques et ne pas inclure de fluents épistémiques. En effet, contrairement au cas de l'éthique computationnelle, cette distinction n'a aucun impacte sur les aspects causaux. La séparation redevient pertinente si l'intérêt porte sur la responsabilité, mais comme il a été expliqué dans la section 3.1.2.2, la causalité n'est qu'une étape pour déterminer la responsabilité et une claire séparation des deux ne peut qu'être bénéfique. Quant à l'ensemble $\mathbb{E}$, celui-ci n'apparaissait pas dans $\mathscr{S}_e$ où l'intérêt était focalisé sur les décisions des agents. Ici, nous faisons un premier pas vers une représentation plus fine du monde et de sa dynamique. C'est pourquoi nous avons besoin de représenter aussi bien les actions des agents, que d'autres évènements non dépendants des agents qui peuvent aussi avoir une influence sur la causalité, sans pour autant avoir besoin de les distinguer à cette étape.

Un *état* dans $\mathscr{S}_s$, noté $S \subseteq \mathbb{F}$, est défini comme un ensemble de fluents. Nous notons $\mathscr{F}$ les formules d'état, qui pour rappel sont des formules de fluents construites en utilisant les opérateurs logiques classiques. Étant donné une formule $\psi \in \mathscr{F}$, la relation $S \models \psi$ est définie classiquement tel que $S \models f \Leftrightarrow f \in S$ et $S \models \neg f \Leftrightarrow f \notin S$ dans le cas où ψ est un fluent. Cette relation est définie classiquement pour les autres formes de ψ.

Exemple 5.1 [suite]**.** *Les fluents $f_1, f_2, f_3 \in \mathbb{F}^3$ correspondent chacun à l'état fermé d'un des éléments du circuit. Ainsi, lorsque le fluent est vrai, cela indique que le courant peut passer pour les interrupteurs et que du courant est généré pour la batterie.*

Définition 5.1 [Système de transition d'états étiqueté $\mathscr{S}_s$]**.** *Le système de transition d'états étiqueté $\mathscr{S}_s$ est un triplet $\langle \mathbb{S}, V, \tau \rangle$ composé de :*

1. *un ensemble d'états $\mathbb{S}$;*
2. *une fonction $V : Lit_{\mathbb{F}} \times \mathbb{S} \rightarrow \{false, true\}$;*
3. *un ensemble de relations étiquetées de transition entre états $\tau \subseteq \mathbb{S} \times 2^{\mathbb{E}} \times \mathbb{S}$.*

Comme montré dans la section 3.1.2.1, lorsqu'il est question de surdétermination, il est inévitablement question de préconditions disjonctives et très souvent de cooccurrence d'évènements. Dès lors qu'une proposition causale est faite, si elle souhaite traiter les cas de surdétermination, elle doit s'appuyer sur une représentation du monde contenant les deux

éléments. $\mathscr{S}_s$ permet les préconditions disjonctives par la forme des formules d'état $\mathscr{F}$ et la cooccurrence d'évènements par les conditions imposées aux relations dans τ.

Un évènement $e \in \mathbb{E}$ est une formule atomique. En causalité, il est important de comprendre pourquoi un évènement e peut se produire. C'est pourquoi il est utile de rendre explicites les *préconditions* des évènements, i.e. les conditions qui doivent être satisfaites par l'état S pour que l'évènement puisse se produire. La fonctions qui associe à chaque évènement ses préconditions est définie comme $pre : \mathbb{E} \to \mathscr{F}$.

Exemple 5.1 [suite]**.** *La précondition pour que l'individu connecté aux électrodes soit électrocuté, évènement noté e_ψ, est $\psi = (f_1 \wedge f_3) \vee (f_2 \wedge f_3)$, où $\psi \in \mathscr{F}$. Si cet évènement se produit, l'individu est considéré comme mort $d \in \mathbb{F}$. Les évènements $e_1, e_2 \in \mathbb{E}^2$ mettent chacun un des interrupteurs dans son état fermé. L'évènement $e_3 \in \mathbb{E}$ fait de même pour la batterie. L'information donnée jusqu'ici fait partie du contexte κ_s de l'exemple. Elle peut se formaliser ainsi :*

$$\forall i \in \{1,2,3\},\ pre(e_i) = \top ;$$
$$pre(e_\psi) = \psi.$$

Du fait que nous proposons un STEE pour la causalité effective, nous nous intéressons à des séquences particulières d'évènements et d'états, et non pas au STEE dans son ensemble. Ces séquences peuvent être vues comme des chemins dans le STEE et un temps $t \in \mathbb{T}$, où $\mathbb{T} = \{0, \ldots, N\}$, peut être associé à chaque élément de la séquence. Nous notons S_0 l'état initial. De plus, étant donné que nous nous intéressons à la surdétermination qui met en défaut la notion de nécessité de la cause pour la conséquence, il est nécessaire que nous puissions raisonner de façon contrefactuelle. Ainsi, pour être capables d'identifier un chemin tout en permettant un raisonnement contrefactuel, nous ne définissons pas une politique pour des états, mais pour des couples état et temps.

Définition 5.2 [Politique π_s]**.** *Dans $\mathscr{S}_s$, la politique π_s est une fonction qui associe à un couple état et temps l'ensemble d'évènements qui sont supposés se produire dans cet état à ce moment là. Formellement cela s'écrit :*

$$\pi_s : 2^{\mathbb{F}} \times \mathbb{T} \to 2^{\mathbb{E}}.$$

Définition 5.3 [*Cadre causal* χ]**.** *Dans $\mathscr{S}_s$, le cadre causal χ est le couple (κ_s, π_s), où π_s une politique et κ_s le* Contexte *étant le quintuplet $(\mathbb{E}, \mathbb{F}, pre, S_0, \mathbb{T})$.*

Les chemin dans un STEE peuvent être vus comme des traces où $S^\chi(t)$ sont les états traversés et $E^\chi(t)$ les transitions qui nous y ont emmenés. Étant dans un cadre purement déterministe, à chaque χ correspond une unique trace qui vérifie :

— $S^\chi(0) = S_0$;

— $\forall t \in \mathbb{T},\ E^\chi(t) = \pi_s(S^\chi(t), t)$;

— $\forall t \in \mathbb{T},\ (S^\chi(t), E^\chi(t), S^\chi(t+1)) \in \tau$.

Exemple 5.1 [suite]**.** *Voici une trace possible pour le χ décrit. Celle-ci correspond à un cas de surdétermination :*

$$S^\chi(0) = \{\neg f_1, \neg f_2, \neg f_3, \neg d\}, E^\chi(0) = \{e_3\} ;$$
$$S^\chi(1) = \{\neg f_1, \neg f_2, f_3, \neg d\}, \quad E^\chi(1) = \{e_1, e_2\} ;$$
$$S^\chi(2) = \{f_1, f_2, f_3, \neg d\}, \quad E^\chi(2) = \{e_\psi\} ;$$
$$S^\chi(3) = \{f_1, f_2, f_3, d\}.$$

5.1.2 Causalité effective

Ayant introduit le $\mathscr{S}_s$ permettant de modéliser l'exemple 5.1, nous présentons dans cette section deux types de relations causales nécessaires pour y raisonner sur la surdétermination. Comme pour le $\mathscr{S}_e$, nous essayons de rester les plus généraux possibles. C'est pourquoi nous ne donnerons pas une définition spécifique de ce qu'est la causalité effective, nous nous contenterons de donner le type de relations causales dont nous avons besoin. De cette façon, la typologie que nous proposons peut être utilisée par n'importe quelle approche causale dont le formalisme repose sur un STEE, et cela quelles que soient les définitions de causalité spécifiques que ces approches adoptent.

De façon générale, une relation causale peut être décrite comme une relation binaire qui lie une cause à une conséquence. Dans un STEE, il est aussi bien possible de vouloir savoir qu'elles sont les causes de $\psi \in \mathscr{F}$ étant vraie dans un état, ou les causes d'une occurrence d'évènement $e \in \mathbb{E}$. En conséquence, deux types de relations causales différentes sont nécessaires. Leur différence est illustrée sur la figure 5.2. Une formule vraie dans un état correspond à $S^\chi(t) \vDash \psi$ et une occurrence d'évènement à $e \in E^\chi(t)$. Pour rester concis, nous notons la première (ψ, t) et la seconde (e, t).

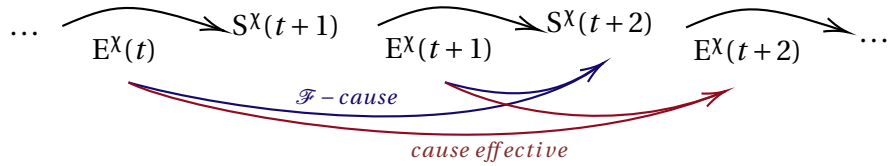


FIGURE 5.2 – Illustration de l'extrait d'une trace avec les deux types de relations causales, à savoir des $\mathscr{F}$-causes et des causes effectives.

Supposée comme étant une entrée propre à chaque approche, nous définissons de façon abstraite la relation causale de base que nous appelons des $\mathscr{F}$-causes.

Définition 5.4 [$\mathscr{F}-causes$]**.** *Étant donné une occurrence d'évènement* $(e, t) \in \mathbb{E} \times \mathbb{T}$*, une formule étant vraie dans un état* $(\psi, t_\psi) \in \mathscr{F} \times \mathbb{T}$ *et* $t < t_\psi$*, une* $\mathscr{F}$*-cause est une relation causale où* (e, t) *est la cause et* (ψ, t_ψ) *est la conséquence.*

Chacune des approches causales dont le formalisme repose sur un STEE peut spécifier sa propre définition pour les $\mathscr{F}$-causes [BATUSOV et SOUTCHANSKI, 2018; BERREBY et collab., 2018; LEBLANC et collab., 2019; SARMIENTO et collab., 2022]. À partir de cette définition de base, il est possible de définir la relation causale la plus commune parmi toutes les approches de causalité, les causes effectives.

Définition 5.5 [Causes effectives]**.** *Étant donné deux occurrences d'évènements* $(e, t) \in \mathbb{E} \times \mathbb{T}$ *et* $(e', t') \in \mathbb{E} \times \mathbb{T}$*, où* $t < t'$*,* (e, t) *est une cause effective de* (e', t') *si* (e, t) *est une* $\mathscr{F}$*-cause de* $(pre(e'), t')$*.*

Dans la section 3.1.2.1, nous avons vu que les travaux existants font tous référence d'une façon ou d'une autre au concept de « chemin causal » lorsqu'ils parlent des différents cas de surdétermination. Ce concept paraît donc être central pour définir ces cas. Toutefois, ce concept est surtout présent implicitement dans le raisonnement, il est rarement explicité et encore moins défini formellement, hormis quelques exceptions où il n'est pas question de

surdétermination [BAUMGARTNER, 2013; BECKERS, 2021b]. En vue de clarifier les différents cas de surdétermination, nous définissons ce qu'est un chemin causal dans la modélisation du monde introduite en section 5.1.2.

Définition 5.6 [Chemin causal ω]**.** *Étant donné un cadre causal* χ *et un évènement* $e_\psi \in E^\chi(t_\psi)$*, avec* $\psi = pre(e_\psi)$*, la séquence d'occurrences d'évènements* $\omega = (e_n, t_n), \ldots, (e_1, t_1)$ *est un chemin causal reliant* (e_n, t_n) *à* (e_ψ, t_ψ) *ssi :*

- (e_n, t_n) *est une* $\mathscr{F}$*-cause de* $(pre(e_{n-1}), t_{n-1})$ *;*
- $\ldots$*;*
- (e_2, t_2) *est une* $\mathscr{F}$*-cause de* $(pre(e_1), t_1)$ *;*
- (e_1, t_1) *est une* $\mathscr{F}$*-cause de* (ψ, t_ψ)*.*

Un chemin causal reliant (e_n, t_n) *à* (e_ψ, t_ψ) *est dit complet s'il ne peut être obtenu en retirant des occurrences à une séquence d'occurrences d'évènements étant également un chemin causal reliant* (e_n, t_n) *à* (e_ψ, t_ψ)*.*

Dit autrement, un chemin causal entre une cause et une conséquence est une séquence d'occurrences d'évènements où chaque évènement est une $\mathscr{F}$-cause des préconditions du suivant, si bien qu'il contribue au déclenchement du suivant.

Proposition 5.1 [Existence d'un chemin causal complet]**.** *Chaque* $\mathscr{F}$*-cause a au moins un chemin causal complet lui correspondant.*

Démonstration. L'existence d'un chemin causal découle directement de la définition 5.6 avec $n = 1$. Alors, soit ce chemin causal est complet, soit il est une sous-séquence d'un chemin causal complet. □

Si la définition adoptée pour les $\mathscr{F}$-causes est transitive, la réciproque de la proposition 5.1 est vraie. Cela veut dire entre autres que, s'il existe un chemin causal ω reliant (e_n, t_n) à (e_ψ, t_ψ), (e_n, t_n) est une cause effective de (e_ψ, t_ψ). Dans la suite de ce chapitre, quand nous ferrons référence à un chemin causal, nous serrons toujours en train de faire référence au chemin causal complet.

5.2 La surdétermination en causalité

Avant de présenter notre proposition de formalisation, nous passons en revue les différentes catégories de la typologie existante. Un exemple de définition pour chacune des catégories a été donné dans la section 3.1.2.1. Pour rappel, à notre connaissance, ces catégories n'ont été définies dans la littérature que de manière générale en utilisant le langage naturel, ce qui peut donner lieu à diverses interprétations et donc à des confusions. Nous présentons maintenant des exemples de traces correspondant à chaque type de cas de surdétermination. Pour cela nous nous appuyons sur les exemples 5.2 à 5.5, et sur les figures 5.3 à 5.6. Quelques détails et discussions sont omis consciemment, car ils font l'objet d'une discussion plus approfondie dans la section 5.3.1.

Exemple 5.1 [suite]**.** *À partir de maintenant, nous considérons une version de l'exemple 5.1 plus complexe. Les agents ne contrôlent pas les interrupteurs directement, il y a une multiplicité de mécanismes constitués de poulies, cordes et engrenages entre leurs actions et la modification effective de l'état des interrupteurs. Les actions faites par deux des agents seront à présent notées* e^1_m *et* e^2_n*. Pourvu que les conditions de la définition 5.6 soient remplies, à chacune*

de ces actions peuvent être associés respectivement les chemins causaux $\omega^1 = (e^1_m, t^1_m), \ldots, (e^1_1, t^1_1)$ *et* $\omega^2 = (e^2_n, t^2_n), \ldots, (e^2_1, t^2_1)$. *Les figures 5.3 à 5.6 illustrent les différents types de cas de surdétermination. Notez que le fait que les évènements* e^1_m *et* e^2_n *se produisent dans tous les cas simultanément n'est qu'un choix pour simplifier l'illustration et non pas une condition pour parler de surdétermination. Comme nous le verrons par la suite, le début du chemin causal n'est pas central dans la classification des différents cas de surdétermination.*

Exemple 5.2 [symétrique/duplicative]**.** *La figure 5.3 illustre les traces de ce qui est considéré comme un cas classique de surdétermination symétrique/duplicative. Dans celle-ci* ω^1 *et* ω^2 *contribuent conjointement à l'occurrence de* e_Ψ *à travers* f_2. *Cela correspond à l'exemple de la littérature où deux assassins versent une dose létale de poison dans la boisson de la victime [*HITCHCOCK, *2007].*

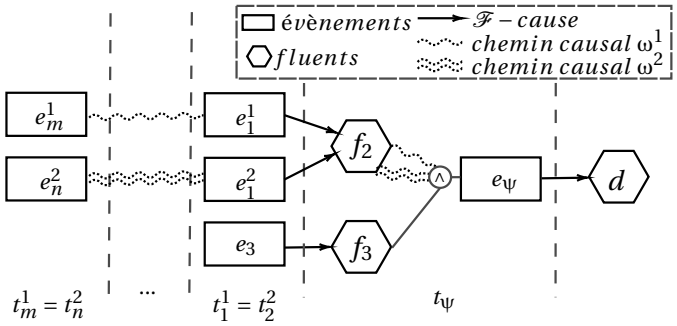


FIGURE 5.3 – Illustration d'un cas de surdétermination symétrique/duplicative.

Exemple 5.3 [imitative]**.** *La figure 5.4 illustre les traces de ce qui est considéré comme un cas classique de surdétermination imitative. Celle-ci ressemble fortement à la figure 5.3. Toutefois, l'occurrence de* e^1_1 *et* e^2_1 *ici n'est pas simultanée et donc* e^2_1 *a comme effet un fluent qui était déjà vrai. Cela correspond à l'exemple de la littérature où un bateau sur un fleuve est obligé de s'arrêter car la rivière est bloquée. Non seulement le pont* A *s'est effondré dans la rivière à quelques mètres du bateau, le pont* B *s'est également effondré un peu plus loin dans la rivière bloquant également le passage [*WRIGHT, *2011].*

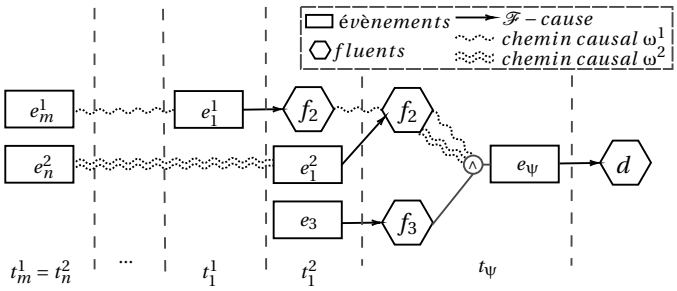


FIGURE 5.4 – Illustration d'un cas de surdétermination imitative.

Exemple 5.4 [préemptive précoce]**.** *La figure 5.5 illustre les traces de ce qui est considéré comme un cas classique de surdétermination préemptive précoce. Dans celle-ci* ω^1 *interrompt* ω^2 *avant que son effet ne puisse se produire. Cette interruption peut être faite par n'importe quel élément de* ω^1 *qui rend faux un élément nécessaire au déclenchement de* ω^2*. Cela correspond à l'exemple de la littérature où un assassin* A *empoisonne la gourde d'un voyageur dans le désert, gourde étant sa seule source d'eau. Puis un assassin* B *vide la gourde avant même que le voyageur puisse y boire. Le voyageur est retrouvé mort déshydraté quelques jours plus tard [*BOCHMAN*, 2018b].*

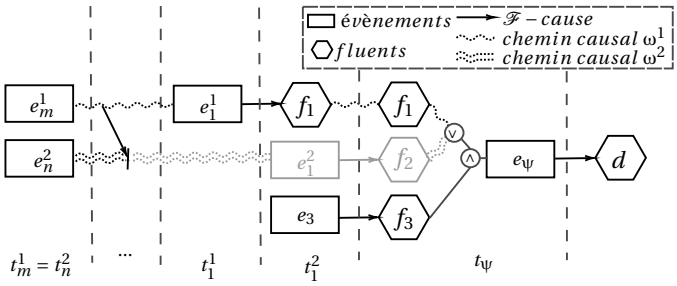


FIGURE 5.5 – Illustration d'un cas de surdétermination préemptive précoce.

Exemple 5.5 [préemptive tardive]**.** *La figure 5.6 illustre les traces de ce qui est considéré comme un cas classique de surdétermination préemptive tardive. Dans celle-ci* ω^2 *est interrompue par la conséquence du chemin causal lui même, i.e. l'évènement qui a été causé par un cas de surdétermination. Ce cas pourrait se produire si nous modifions légèrement la condition pour qu'il y ait électrocution :* $\psi = (f_1 \wedge f_3 \wedge \neg d) \vee (f_2 \wedge f_3 \wedge \neg d)$*. En effet, comme l'occurrence* (e_ψ, t_ψ) *cause* d *vrai au temps* $t_\psi + 1$*, la précondition* ψ *ne peut plus être rendue vraie au temps* $t_\psi + 1$*, et donc* ω^2 *est interrompu. Cela correspond à l'exemple de la littérature où deux feux de forêt sont déclenchés et chacun peut à lui seul détruire une maison. Il se trouve que l'un des deux feux atteint la maison en premier et brûle la maison dans son entièreté juste avant que le deuxième feu n'atteigne la maison [*BATUSOV *et* SOUTCHANSKI*, 2018].*

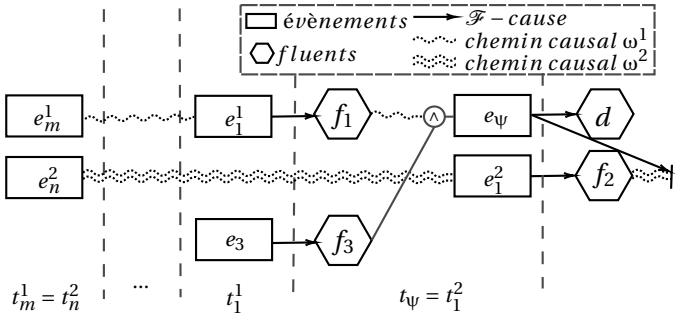


FIGURE 5.6 – Illustration d'un cas de surdétermination préemptive tardive.

La section 5.2.1 présente notre définition formelle de la surdétermination dans $\mathscr{S}_s$. Toutes les catégories de la typologie existante décrites ci-dessus peuvent être considérées comme des cas particuliers de la définition 5.7. Puis, la section 5.2.2 présente notre proposition de typologie formelle des cas de surdétermination.

5.2.1 Définition formelle de la surdétermination

Dans cette section nous proposons une définition formelle de ce qu'est la surdétermination, un concept rarement défini, mais qui à nos yeux est indispensable pour pouvoir comprendre puis traiter les différents cas de surdétermination.

Le formalisme décrit dans la section 5.1.1 fait une distinction forte entre les états du monde et les évènements. Lorsque nous parlons de surdétermination, cela pourrait être fait aussi bien par rapport aux $\mathscr{F}$-causes, qu'aux causes effectives. Le premier cas revient à s'intéresser au fait que l'individu soit considéré comme mort, alors que le deuxième revient plutôt à s'intéresser à l'occurrence de l'électrocution. Comme montré dans le chapitre 2 et la section 3.2.1.2, la distinction entre fluents et évènements semble être pertinente pour représenter le monde, mais aussi lorsqu'il est question de causalité effective. Toutefois, tous les formalismes ne font pas la différence. En l'occurrence, dans les approches basées sur les équations structurelles, toutes les variables sont considérées comme des évènements. Afin de permettre la comparaison avec le plus grand nombre d'approches possibles, nous faisons le choix de garder le type de relation causale commun à toutes pour parler de surdétermination : les causes effectives. Si nous reprenons l'exemple de Suzy et Billy, cela veut dire que nous regardons les causes de l'évènement $bouteille\_se\_brise$ se produisant, plutôt que les causes du fluent $bouteille\_brisee$ étant vrai.

Comme mentionné dans la section 5.1, pour parler de surdétermination il faut pouvoir raisonner de façon contrefactuelle. Pour ce faire, il est nécessaire d'imaginer des traces différents dans $\mathscr{S}_s$. Ayant mentionné qu'à un cadre causal $\chi = (\kappa_s, \pi_s)$ correspondait une trace unique, imaginer d'autres traces revient à modifier soit le contexte κ_s, soit la politique π_s. Nous avons vu dans le chapitre 3 que les raisonnements contrefactuels classiques en causalité revenaient à imaginer ce qui se serait passé si des évènements avaient ou n'avaient pas eu lieu, et non pas à imaginer ce qui se serait passé si les évènements avaient d'autres effets ou préconditions. Partant de ce constat, nous imaginons des traces alternatives en modifiant la politique π_s plutôt que κ_s. Ces politiques alternatives peuvent être vues comme des politiques contrefactuelles. Pour les construire nous définissons l'opération :

$$\pi_s \setminus \mathrm{E} \overset{\text{def}}{=} \forall \mathrm{S}, \forall t,\ \pi_s(\mathrm{S}, t) = \pi_s(\mathrm{S}, t) \setminus \mathrm{E}.$$

Définition 5.7 [Surdétermination]**.** *Soit* $\chi = (\kappa_s, \pi_s)$ *le cadre causal et* $(a^1, t^1), (a^2, t^2), (e_\psi, t_\psi)$ *trois occurrences d'évènements. Étant donné trois cadres causaux contrefactuels :*

$$\chi_{\mathrm{I}}^1 = (\kappa_s, \pi_s \setminus \{a^2\}), \qquad \chi_{\mathrm{I}}^2 = (\kappa_s, \pi_s \setminus \{a^1\}), \qquad \chi_- = (\kappa_s, \pi_s \setminus \{a^1, a^2\})$$

où χ_{I}^1 *et* χ_{I}^2 *sont dits des cadres causaux Individuels, nous sommes dans un cas de surdétermination entre* a^1 *et* a^2 *dans le cadre causal* χ *si sont vérifiées :*

- $e_\psi \in \mathrm{E}^{\chi}(t_\psi)$;
- $e_\psi \notin \mathrm{E}^{\chi_-}(t_\psi)$;
- $e_\psi \in \mathrm{E}^{\chi_{\mathrm{I}}^1}(t_\psi)$;
- $e_\psi \in \mathrm{E}^{\chi_{\mathrm{I}}^2}(t_\psi)$.

Dans la définition 5.7 la surdétermination est définie dans le cas où au plus deux éléments sont en surdétermination. Dans les cas de surdétermination impliquant N évènements, il est possible d'élaborer une politique dans laquelle les N évènements peuvent être traités par paire. Pour cela il serait nécessaire de définir qu'il existe une surdétermination

indirecte entre a^1 et a^2 pour $\chi = (\kappa_s, \pi_s)$ s'il existe $E \subset \mathbb{E}$ tel que a^1 et a^2 sont en surdétermination dans $\chi = (\kappa_s, \pi_s \setminus E)$. Comme la plupart des exemples de la littérature concernent la surdétermination entre deux évènements, nous considérons qu'il s'agit d'une extension de ce travail plutôt que d'un élément essentiel à traiter ici.

5.2.2 Typologie formelle des cas de surdétermination

Ce que nous appelons surdétermination étant maintenant défini formellement, nous passons à notre proposition de typologie des cas de surdétermination. Le tableau 5.1 indique le type de cas de surdétermination qui correspond à certaines conditions causales, données par l'entête des colonnes, auxquelles nous ajoutons certaines conditions temporelles, données par l'entête des lignes. Il s'agit là d'une typologie qui permet de classifier clairement les cas de surdétermination de la littérature dans six catégories distinctes. Cette typologie est le fruit d'une étude exhaustive de ces cas. Cette typologie pouvant être vue comme de la clarification de concepts, nous pensons important de détailler le processus par lequel nous y sommes parvenus. C'est ce que nous faisons tout au long de cette section.

Algèbre des intervalles d'Allen		$\Omega^1 = \{\omega^1\}, \Omega^2 = \emptyset$		$\Omega^1 = \{\omega^1\}, \Omega^2 = \{\omega^2\}$	
		$(e_i^1, t_i^1) \rightarrow \left(\neg pre(e_j^2), t_j^2\right)$	$(e_\psi, t_\psi) \rightarrow \left(\neg pre(e_j^2), t_j^2\right)$	$W^1 \neq W^2$	$W^1 = W^2$
ω^1 égal à ω^2		Préemptive précoce	*	Duplicative synchrone	Symétrique
ω^1 termine ω^2		Préemptive précoce	*	Duplicative synchrone	Symétrique
ω^1 terminé par ω^2		Préemptive précoce	*	Duplicative synchrone	Symétrique
ω^1 chevauche ω^2		Préemptive précoce	Préemptive tardive/Durative	Duplicative asynchrone	Imitative
ω^1 chevauché par ω^2		Préemptive précoce	*	**	**
ω^1 démarre ω^2		Préemptive précoce	Préemptive tardive/Durative	Duplicative asynchrone	Imitative
ω^1 démarré par ω^2		Préemptive précoce	*	**	**
ω^1 se déroule pendant ω^2		Préemptive précoce	Préemptive tardive/Durative	Duplicative asynchrone	Imitative
ω^1 englobe le déroulé de ω^2		Préemptive précoce	*	**	**
ω^1 rencontre ω^2		Préemptive précoce	Préemptive tardive	Duplicative asynchrone	Imitative
ω^1 est rencontré par ω^2		*	*	**	**
ω^1 se déroule avant ω^2		Préemptive précoce	Préemptive tardive	Duplicative asynchrone	Imitative
ω^1 se déroule après ω^2		*	*	**	**

TABLEAU 5.1 – Typologie formelle des cas de surdétermination prenant en compte toutes les relations temporelles possibles entre deux chemins causaux. (*) Incohérence entre la relation causale à l'origine de l'interruption de ω^2 et la relation temporelle entre les intervalles. (**) Incohérence entre l'hypothèse que ω^1 est toujours le premier chemin causal à aboutir et la relation temporelle entre les intervalles.

Nous nous plaçons dans le cadre causal χ et nous nous considérons dans un cas de surdétermination. D'après la définition 5.7, nous avons donc un évènement $e_\psi \in E^\chi(t_\psi)$ avec $pre(e_\psi) = \psi$, deux occurrences d'évènements $(e_m^1, t_m^1), (e_n^2, t_n^2)$ et trois cadres causaux contrefactuels $\chi_I^1 = (\kappa_s, \pi_s \setminus \{e_n^2\})$, $\chi_I^2 = (\kappa_s, \pi_s \setminus \{e_m^1\})$ et $\chi_- = (\kappa_s, \pi_s \setminus \{e_m^1, e_n^2\})$.
Étant donné un couple d'indices $(i,k) \in \{(1,m),(2,n)\}$, l'ensemble de tous les chemins causaux qui relient (e_k^i, t_k^i) à (e_ψ, t_ψ) dans χ_I^i est noté Ω_I^i. Si $i = 1$, nous sommes dans le cadre χ_I^1 où nous supposons que (e_n^2, t_n^2) ne s'est pas produit puisqu'il est retiré de la politique, et donc nous étudions les chemins causaux partant de (e_m^1, t_m^1) individuellement. Si $i = 2$, nous sommes dans le cadre χ_I^2 où nous supposons que (e_m^1, t_m^1) ne s'est pas produit puisqu'il est retiré de la politique, et donc nous étudions les chemins causaux partant de (e_n^2, t_n^2) individuellement.
De la même façon, l'ensemble de tous les chemins causaux qui relient (e_k^i, t_k^i) à (e_ψ, t_ψ) dans χ est noté Ω^i. Dans ce cas là, nous sommes dans le cadre χ où nous avons aussi bien (e_m^1, t_m^1), que (e_n^2, t_n^2).

En s'appuyant sur la définition 5.7, nous pouvons déduire que $\Omega_I^1 \neq \varnothing$ et $\Omega_I^2 \neq \varnothing$. Nous savons que dans le cadre causal χ_-, $e_\psi \notin E^{\chi_-}(t_\psi)$, tandis que dans χ_I^i, $e_\psi \in E^{\chi_I^i}(t_\psi)$. La seule différence entre χ_- et χ_I^i étant la présence dans le deuxième de (e_k^i, t_k^i), il est possible de déduire qu'il existe nécessairement un chemin causal reliant (e_k^i, t_k^i) et (e_ψ, t_ψ) dans χ_I^i. En définitive, $\Omega_I^i \neq \varnothing$.
Selon un raisonnement analogue nous pouvons déduire que $\Omega^1 \cup \Omega^2 \neq \varnothing$. Nous savons que dans le cadre causal χ_-, $e_\psi \notin E^{\chi_-}(t_\psi)$, tandis que dans χ, $e_\psi \in E^{\chi}(t_\psi)$. La seule différence entre χ_- et χ étant la présence dans le deuxième de (e_m^1, t_m^1) et (e_n^2, t_n^2), il est possible de déduire qu'il existe nécessairement un chemin causal reliant au moins une des deux occurrences à (e_ψ, t_ψ) dans χ. En définitive, $\Omega^1 \cup \Omega^2 \neq \varnothing$.
Sans perte de généralité, nous pouvons assumer que $\Omega^1 \neq \varnothing$, i.e. que dans χ il existe toujours un chemin causal reliant (e_m^1, t_m^1) à (e_ψ, t_ψ) ; si bien que l'ensemble Ω^2 peut très bien être vide ou pas, dans les deux cas $\Omega^1 \cup \Omega^2 \neq \varnothing$.

L'existence de multiples chemins causaux pouvant relier deux mêmes occurrences d'évènements, soit les cas où $\left|\Omega_I^i\right| > 1$, soulève des nouveaux cas intéressants, mais au delà du cadre de la présente contribution. En premier lieu, certains chemins causaux présents dans Ω_I^i peuvent ne plus l'être dans Ω^i. L'ensemble $\Omega_I^i \setminus \Omega^i$ contient ces chemins « interrompus ». En second lieu, d'autres chemins causaux présents dans Ω_I^i peuvent tout simplement être retrouvés dans Ω^i. L'ensemble $\Omega_I^i \cap \Omega^i$ contient ces chemins « conservés ». En dernier lieu, des chemins absents dans Ω_I^i peuvent apparaître dans Ω^i de l'interaction entre chemins dans Ω_I^1 et Ω_I^2. L'ensemble $\Omega^i \setminus \Omega_I^i$ contient ces chemins « créés ». À ces trois cas s'ajoute une question complexe. Imaginons que $|\Omega^1| = 3$ et $|\Omega^2| = 1$. Faut-il considérer que des cas de surdétermination de différents types coexistent, un cas de préemption tardive peut-il coexister avec un cas symétrique/duplicatif, ou y a-t-il des règles de subsomption ? À notre connaissance, ces cas n'ont jamais été envisagés dans le domaine de la causalité effective. C'est grâce à l'analyse formelle du problème de surdétermination que leur existence nous est apparue. Notre but étant d'aider à clarifier les débats existants, ces cas dépassent le cadre de la présente contribution. Toutefois, identifier leur existence est une première contribution nécessaire à leur résolution.
Par souci de concision et de clarté, nous restreignons donc le cadre de notre analyse en faisant deux hypothèses : il existe un unique chemin causal reliant deux même occurrences d'évènements, i.e. $|\Omega_I^1| = |\Omega_I^2| = 1$, et aucun chemin causal n'est créé de l'interaction entre chemins dans Ω_I^1 et Ω_I^2, i.e. $\Omega^i \setminus \Omega_I^i = \varnothing$. Ces hypothèses posées, nous pouvons en déduire que $\Omega_I^1 = \Omega^1 = \{\omega^1\}$, $\Omega_I^2 = \{\omega^2\}$ et soit $\Omega^2 = \{\omega^2\}$, ou $\Omega^2 = \varnothing$. Si $\Omega^2 = \{\omega^2\}$, cela veut dire que $\forall (e^2, t^2) \in \omega^2$, $e^2 \in E^{\chi}(t^2)$, alors que si $\Omega^2 = \varnothing$, cela veut dire que ω^2 n'est pas un chemin causal dans χ de fait que $\exists (e^2, t^2) \in \omega^2$, $e^2 \notin E^{\chi}(t^2)$. Cette distinction entre les deux cas possibles pour Ω^2 est cruciale car elle permet de différencier les cas de surdétermination préemptive des autres. La typologie que nous présentons pouvant être organisée comme un tableau à double entrée, chacun des deux cas possibles pour Ω^2 correspond dans notre typologie à une première différenciation en colonnes. La figure 5.7 illustre la configuration à laquelle nous sommes arrivés et qui est celle considérée par la typologie que nous proposons.

Une étude exhaustive des relations temporelles entre chemins causaux est indispensable. En effet, dans l'étude des exemples utilisés dans la littérature, il apparaît que des termes faisant référence aux relations temporelles entre les chemins causaux sont omniprésents. De plus, s'il existe bien une différence entre les cas de préemption tardive duratifs

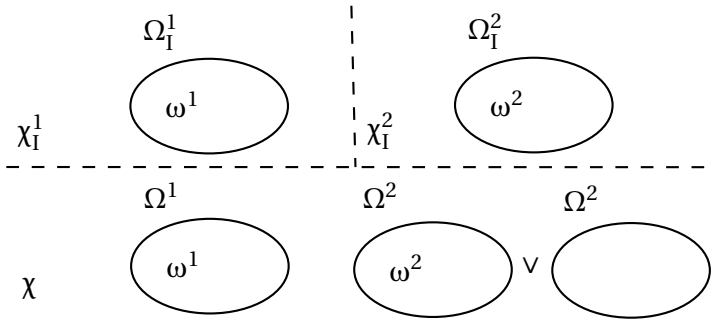


FIGURE 5.7 – Illustration des différents ensembles de chemins causaux dans les cadres causaux χ_I^i et χ.

et non duratifs décrits en section 3.1.2.1, celle-ci est nécessairement liée à un aspect temporel, en plus des aspects propres à la formalisation du temps continu.
Pour réaliser l'étude exhaustive des relations temporelles entre chemins causaux nécessaires, nous utilisons l'algèbre des intervalles d'ALLEN [1983]. Les chemins causaux ω^i sont représentés comme des intervalles temporels qui commencent à t_k^i et qui finissent à t_1^i, temps correspondant respectivement à la première occurrence (e_k^i, t_k^i) et à la dernière occurrence (e_1^i, t_1^i) du chemin causal. Quelles que soient les combinaisons de Ω^1 et Ω^2 testées, nous les confrontons toutes aux treize relations possibles entre deux intervalles, comme définies par l'algèbre des intervalles d'Allen, et regardons si cela à une quelconque influence sur le type de cas de surdétermination. Dans la représentation de notre typologie dans le tableau 5.1, chacune des treize relations d'Allen correspond à une ligne du tableau.
Notez que dans l'algèbre des intervalles d'Allen, chacune des treize relations a un nom particulier. Nous utilisons ces noms pour les identifier. Toutefois, la sémantique de ces noms ne doit pas être prise en compte dans la compréhension de la situation analysée par la typologie. En l'occurrence, ω^1 et ω^2 étant des intervalles, une des relations est identifiée par Allen comme « ω^1 termine ω^2 ». Sur cette ligne du tableau, il faut uniquement considérer la relation temporelle entre les intervalles, relation qui est illustrée à la suite du nom. Il ne faut surtout pas comprendre cette ligne comme un cas où ω^1 aurait une quelconque influence sur ω^2. Cela est peut être le cas dans une des cases de la ligne, mais cette information est donnée par l'entête de la colonne. Pour résumer, les relations temporelles sont gérées par les entêtes de lignes et les relations causales par les entêtes de colonnes.

La première grande colonne du tableau 5.1 considère le cas où $\Omega^2 = \varnothing$. Dans cette colonne nous pouvons déduire que l'occurrence (e_m^1, t_m^1) a quelque chose à voir avec l'interruption de ω^2. Nous savons que dans le cadre causal χ_I^2, ω^2 est bien un chemin causal, tandis que dans χ il ne l'est plus. La seule différence entre χ_I^2 et χ étant la présence dans le deuxième de (e_m^1, t_m^1), il est possible de déduire que cette occurrence est liée avec l'interruption de ω^2. Interrompre un chemin causal revient à empêcher le déclenchement d'une des occurrences composant ce chemin. D'après l'étude des exemples utilisés dans la littérature, il faut alors se demander si ce qui a interrompu ω^2 est l'occurrence de (e_ψ, t_ψ), due au fait que ω^1 a abouti en premier, comme le cas des deux feux de forêts de l'exemple 5.5, ou une des occurrences composant le chemin causal ω^1, comme le cas des deux assassins et le voyageur dans le désert de l'exemple 5.4. Le premier cas de figure s'exprime formellement :

$$\exists (e_j^2, t_j^2) \in \omega^2,\ (e_\psi, t_\psi) \rightarrow \left(\neg pre(e_j^2), t_j^2\right).$$

Le deuxième s'exprime formellement :

$$\exists\left((e_i^1, t_i^1) \in \omega^1, (e_j^2, t_j^2) \in \omega^2\right),\ (e_i^1, t_i^1) \rightarrow \left(\neg pre(e_j^2), t_j^2\right).$$

La figure 5.6 correspond au premier cas et la figure 5.5 au deuxième. La première grande colonne du tableau 5.1 peut alors être divisée en deux sous-colonnes en fonction des ces deux cas de figure. En les confrontant chacun aux treize relations temporelles possibles, la définition formelle de deux types de cas de surdétermination peut déjà être proposée.

Définition 5.8 [Surdétermination préemptive précoce]**.** *Soit un cas de surdétermination. Il s'agit de surdétermination de type préemptive précoce si :*

$$\Omega^1 = \{\omega^1\} \quad \wedge \quad \Omega^2 = \varnothing \quad \wedge \quad \exists\left((e_i^1, t_i^1) \in \omega^1, (e_j^2, t_j^2) \in \omega^2\right), (e_i^1, t_i^1) \rightarrow \left(\neg pre(e_j^2), t_j^2\right).$$

Définition 5.9 [Surdétermination préemptive tardive]**.** *Soit un cas de surdétermination. Il s'agit de surdétermination de type préemptive tardive si :*

$$\Omega^1 = \{\omega^1\} \quad \wedge \quad \Omega^2 = \varnothing \quad \wedge \quad \exists(e_j^2, t_j^2) \in \omega^2, (e_\psi, t_\psi) \rightarrow \left(\neg pre(e_j^2), t_j^2\right).$$

La deuxième grande colonne du tableau 5.1 considère le cas où $\Omega^2 = \{\omega^2\}$. Dans cette colonne il n'y a aucune interruption de chemin causal. La décomposition faite de la première grande colonne n'est donc pas envisageable. D'après l'étude des exemples utilisés dans la littérature, il faut plutôt s'intéresser à la façon dont l'occurrence (e_ψ, t_ψ) est amenée. Si nous reprenons l'exemple du voyageur dans le désert, « la façon dont l'occurrence (e_ψ, t_ψ) est amenée » revient à se demander si le voyageur est mort par déshydratation ou par empoisonnement. Dans $\mathscr{S}_s$ et dans un STEE en général, la façon dont un évènement se produit est intrinsèquement reliée aux préconditions $\psi = pre(e_\psi)$ de l'évènement en question. Nous dirons que chaque évènement à des supports et que rendre vrai un de ces supports est une façon bien particulière de produire l'évènement. Plus précisément, les supports de e_ψ sont les impliquants premiers de ψ. Nous utilisons la notation W pour faire référence à un support. Dans l'exemple 5.1, il y a deux façons différentes de produire l'électrocution de la victime (e_ψ, t_ψ), soit par $W = \{f_1, f_3\}$, ou par $W' = \{f_2, f_3\}$. Chacune correspond à une des deux branches possibles pouvant être empruntées par le courant en fonction de l'état des interrupteurs. Nous noterons W^1 et W^2 les supports utilisés respectivement par ω^1 et ω^2 pour causer (e_ψ, t_ψ).
La deuxième grande colonne du tableau 5.1 peut alors être divisée en deux sous-colonnes. Une première correspondant au cas où ω^1 et ω^2 utilisent le même support $W^1 = W^2$, et une deuxième où ce n'est pas le cas $W^1 \neq W^2$. Les figures 5.3 et 5.4 correspondent au premier cas, les figures 5.8 et 5.9 au deuxième.
En confrontant chacun de ces deux cas aux treize relations temporelles possibles, il apparaît que les relations temporelles ont une influence importante sur le type de cas de surdétermination. Le cas extrême correspondrait à la situation où chaque ligne correspondrait à un type de surdétermination différent. Ce n'est pas le cas ici, toutes les relations temporelles ne sont pas discriminantes. Après analyse et comparaison avec les exemples de la littérature, il semble que c'est la relation entre t_1^1 et t_1^2 qui est pertinente, i.e. le temps auquel le chemin causale aboutit. Cette observation nous permet de proposer quelques nuances en proposant de scinder le cas symétrique/duplicatif en trois types de cas de surdétermination plus affinés. Leur définition formelle est la suivante.

Définition 5.10 [Surdétermination duplicative synchrone]**.** *Soit un cas de surdétermination. Il s'agit de surdétermination de type duplicative synchrone si :*

$$\Omega^1 = \{\omega^1\} \quad \wedge \quad \Omega^2 = \{\omega^2\} \quad \wedge \quad W^1 \neq W^2 \quad \wedge \quad t_1^1 = t_1^2.$$

Définition 5.11 [Surdétermination duplicative asynchrone]**.** *Soit un cas de surdétermination. Il s'agit de surdétermination de type duplicative asynchrone si :*

$$\Omega^1 = \{\omega^1\} \quad \wedge \quad \Omega^2 = \{\omega^2\} \quad \wedge \quad W^1 \neq W^2 \quad \wedge \quad t_1^1 < t_1^2.$$

Définition 5.12 [Surdétermination symétrique]**.** *Soit un cas de surdétermination. Il s'agit de surdétermination de type symétrique si :*

$$\Omega^1 = \{\omega^1\} \quad \wedge \quad \Omega^2 = \{\omega^2\} \quad \wedge \quad W^1 = W^2 \quad \wedge \quad t_1^1 = t_1^2.$$

La différence entre le cas duplicatif synchrone et asynchrone repose exclusivement sur une relation temporelle entre les chemins causaux. Les figures 5.8 et 5.9 correspondent respectivement à chacun de ces cas. Le cas symétrique doit son nom au fait que ce qui le caractérise est que le support utilisé par les chemins causaux est le même et qu'ils aboutissent simultanément. Suite au raffinement que nous proposons, la figure 5.3 présentée initialement comme un cas symétrique/duplicatif est considéré comme un cas symétrique.

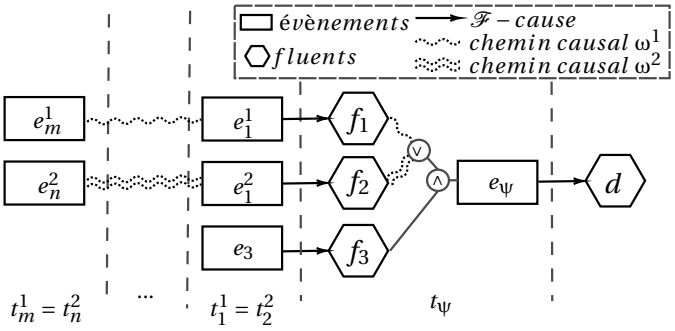


FIGURE 5.8 – Illustration d'un cas de surdétermination duplicative synchrone.

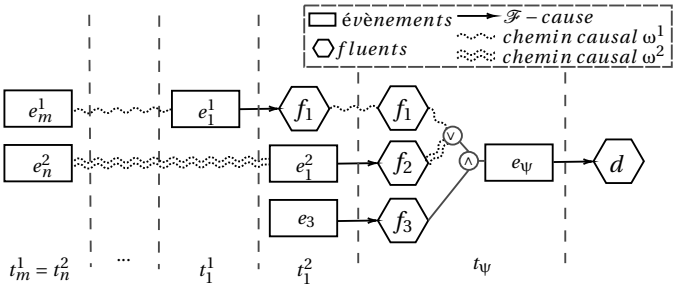


FIGURE 5.9 – Illustration d'un cas de surdétermination duplicative asynchrone.

Pour HALL et PAUL [2003], les cas dits de « trumping » dans la littérature n'existent pas réellement. Il écrit : « cases of trumping turn out on inspection to be nothing more than either cases of symmetric overdetermination in disguise or cases of late pre-emption in

disguise ». Grâce à l'étude formelle et approfondie de la surdétermination, nous sommes en mesure de proposer une définition qui correspond aux exemples classiquement utilisés pour parler de « trumping » et qui est distincte de tous les autres cas déjà introduits. Dans notre typologie nous avons appelé ce cas « imitatif ». Il doit son nom au fait que ce qui le caractérise est que le support utilisé par les chemins causaux est le même, comme le cas symétrique, mais ils n'aboutissent pas simultanément. Le qualifier d'imitatif est une analogie à un individu qui imiterait un autre, aussi bonne l'imitation puisse être, celui qui imite devant observer l'autre avant d'agir implique qu'il y a toujours un décalage entre les deux. La définition que nous donnons à ce cas dit de « trumping » dans la littérature est proche de la vision de McDermott [2002], Halpern et Pearl [2005] et Hitchcock [2007] qui soutiennent qu'il s'agit d'un cas plus proche du symétrique/duplicatif que de la préemption. La figure 5.4 correspond à ce cas.

Définition 5.13 [Surdétermination imitative]**.** *Soit un cas de surdétermination. Il s'agit de surdétermination de type imitative si :*

$$\Omega^1 = \{\omega^1\} \quad \wedge \quad \Omega^2 = \{\omega^2\} \quad \wedge \quad W^1 = W^2 \quad \wedge \quad t_1^1 < t_1^2.$$

Notez que dans les cas correspondant aux deux colonnes où $\Omega^1 = \{\omega^1\}$ et $\Omega^2 = \{\omega^2\}$, le type de cas de surdétermination n'est pas affecté par lequel des deux chemin causaux aboutit en premier. Toutefois, même si cet ordre ne change pas le type, selon la définition de causalité effective qui est défendue, il peut changer quels évènements seront retenus comme étant des causes. En l'occurrence, certaines approches considéreront que dans un cas asynchrone duplicatif, comme celui de la figure 5.9, c'est uniquement les occurrences du premier chemin causal à aboutir qui sont des causes. D'autres approches soutiendront le contraire et d'autres encore pourront dire que cela n'a aucune importance. Dans un souci de généralité, nous souhaitons qu'il ne puisse pas y avoir de divergence dans la réponse causale apportée par une même approche pour deux exemples appartenant au même type de cas de surdétermination. Autrement dit, si pour un premier exemple une théorie trouve que (e_m^1, t_m^1) est une cause et pas (e_n^2, t_n^2), alors, pour un autre exemple classé comme étant du même type que le premier par notre typologie, cette théorie devrait donner la même réponse.
Pour ce faire, tout en gardant le tableau 5.1 le plus concis possible, nous supposons dans ces deux dernières colonnes que ω^1 est toujours le chemin à aboutir en premier. Cette condition explique l'existence de cases grisées dans ces colonnes. Notez que si les cases avec « * » sont grisées c'est parce qu'il y a une vrai incohérence entre les relations causales à l'origine de l'interruption de ω^2 et les relations temporelles entre les deux chemins causaux. En l'occurrence, la relation causale $(e_\psi, t_\psi) \rightarrow \left(\neg pre(e_j^2), t_j^2\right)$ ne peut pas être vérifiée pour la treizième relation de Allen étant donné que ω^1 aboutit après ω^2. Rien d'autre ne pouvant causer l'occurrence (e_ψ, t_ψ) d'après nos hypothèses, ω^2 ne serait pas interrompu et donc il y aurait une contradiction avec l'entête de la colonne puisque $\Omega^2 \neq \varnothing$. Notez que si les cases avec « ** » sont grisées c'est parce qu'il y a une incohérence entre notre hypothèse selon laquelle ω^1 est toujours le chemin à aboutir en premier et les relations temporelles entre les deux chemins causaux. Ce deuxième cas de cases grisées est donc un choix de conception alors que le premier relève d'un cas qui ne pourrait tout simplement pas exister.

Le tableau 5.2 présente une version de la typologie obtenue en remplaçant les treize relations de Allen par le paramètre temporel identifié comme discriminant. Cette version à

	$\Omega^1 = \{\omega^1\}, \Omega^2 = \emptyset$		$\Omega^1 = \{\omega^1\}, \Omega^2 = \{\omega^2\}$	
	$(e_i^1, t_i^1) \rightarrow (\neg pre(e_j^2), t_j^2)$	$(e_\psi, t_\psi) \rightarrow (\neg pre(e_j^2), t_j^2)$	$W^1 \neq W^2$	$W^1 = W^2$
$t_1^1 < t_1^2$	Préemptive précoce	Préemptive tardive/Durative	Duplicative asynchrone	Imitative
$t_1^1 = t_1^2$	Préemptive précoce	*	Duplicative synchrone	Symétrique
$t_1^1 > t_1^2$	Préemptive précoce	*	**	**

TABLEAU 5.2 – Typologie formelle concise des cas de surdétermination prenant en compte les relations temporelles pertinentes entre deux chemins causaux. (*) Incohérence entre la relation causale à l'origine de l'interruption de ω^2 et la relation temporelle entre les intervalles. (**) Incohérence entre l'hypothèse que ω^1 est toujours le premier chemin causal à aboutir et la relation temporelle entre les intervalles.

l'avantage d'être plus concise, en revanche les lecteurs les plus attentifs remarqueront que la distinction entre préemption tardive et préemption tardive durative est perdue. En effet, le fait que la version durative ne puisse avoir lieu que dans un plus petit nombre de configurations temporelles, qui correspondent à $t_1^1 > t_n^2$, est perdu puisque les temps de début de chaîne ne sont pas gardés. Toutefois, comme il a été mentionné dans la section 3.1.2.1, la version durative ne peut être formalisée que dans une modélisation du monde non discrétisée, ce qui n'est ni le cas de $\mathscr{S}_s$, ni d'aucune approche causale dont nous avons connaissance. En conséquence, cette concession est bien moindre par rapport au gain en concision et clarté que cette version apporte.

5.3 Enseignements et autres résultats à partir de la typologie

Dans cette section nous abordons deux points qu'il semble important d'explorer suite à la proposition de la typologie. Le premier sera traité dans la section 5.3.1 et concerne la sensibilité de la causalité à la représentation du problème, et donc à quel point cette étape de représentation est importante. Le deuxième plus qu'un constat est une proposition que nous faisons. Ayant défini formellement six catégories bien distinctes couvrant les exemples de surdétermination utilisés dans la littérature, nous proposons une nouvelle façon de comparer les différentes approches causales entre elles. Ce deuxième point sera traité dans la section 5.3.2.

5.3.1 Sur l'importance de la représentation

Dans cette section nous montrons que la façon dont le problème est représenté est décisive d'un point de vue causal. L'appartenance d'un problème à un type de cas de surdétermination peut varier selon certains choix de représentation qui pourraient paraître anodins. C'est pourquoi nous défendons l'idée que lorsqu'il est question de problèmes complexes de surdétermination causale, il est important d'utiliser des langages suffisamment expressifs car ils forcent à rendre explicites certains de ces choix décisifs.

Commençons par montrer que les frontières entre les deux dernières colonnes sont sensibles à la représentation. Pour cela nous allons commencer par parler du cas imitatif qui, comme montré dans la section 3.1.2.1, paraît être le plus controversé. En effet, alors que des auteurs comme HALL et PAUL [2003] considèrent que ces cas ne constituent pas vraiment un type distinct des autres, d'autres comme BOCHMAN [2018a] considèrent qu'il s'agit de cas avec une importance particulière : « relevant cases of actual causation ». Un des exemples

de surdétermination imitative le plus utilisé est celui où un bateau sur un fleuve est obligé de s'arrêter du fait que la rivière est bloquée. Non seulement le pont A s'est effondré dans la rivière à quelques mètres du bateau, le pont B s'est également effondré un peu plus loin dans la rivière bloquant également le passage [WRIGHT, 2011].
Cet exemple peut être classé dans différents types selon la façon dont il est représenté. Si la raison pour laquelle le bateau s'arrête est que le bateau s'approche d'un obstacle, façon alternative de dire que la rivière est bloquée, alors cet exemple sera considéré comme un cas de préemption tardive. Si l'évènement correspondant à l'arrêt du bateau est noté e_ψ, sa précondition sera $\psi = obstacle\_en\_face$. Le chemin causal ω^1 associé à l'effondrement du pont A aboutit en premier par sa proximité spatiale et déclenche e_ψ. Alors, l'effet de e_ψ qui est d'arrêter le bateau interromps le chemin causal ω^2 associé à l'effondrement du pont B puisque étant à l'arrêt, il ne pourra pas arriver à l'endroit où le pont B s'est effondré. ω^2 ne sera donc pas un chemin causal et nous serons dans le cas de préemption tardive de la définition 5.9.
Par contre, si nous nous tenons à la description des faits et que nous représentons le fait que dès que la rivière est bloquée le bateau s'arrête, nous pouvons être dans un des quatre types des deux dernières colonnes du tableau 5.2, celles ayant comme entête $\Omega^1 = \{\omega^1\}$, $\Omega^2 = \{\omega^2\}$. Une première représentation consisterait à considérer que chaque effondrement de pont bloque la rivière d'une façon particulière qui doit être représentée différemment. Dans ce cas, la précondition serait $\psi = \text{A\_}bloque\_riviere \vee \text{B\_}bloque\_riviere$. En fonction de la temporalité des effondrements de ponts, nous sommes soit dans le cas duplicatif synchrone ou duplicatif asynchrone des définitions 5.10 et 5.11.
Une deuxième représentation consisterait à considérer qu'il n'y a aucune différence dans la façon dont les effondrements de pont bloquent la rivière, ou en tout cas que la différence n'a aucune importance. Dans ce cas, la précondition serait $\psi = riviere\_bloquee$. À nouveau, en fonction de la temporalité des effondrements de ponts, nous sommes dans des cas différents. Si l'effondrement est simultané, alors il s'agit du cas symétrique de la définition 5.12. Sinon, il s'agit du cas imitatif de la définition 5.13. Notez que représenter ce dernier cas implique de pouvoir exprimer le fait que malgré que le fluent $riviere\_bloquee$ ait été rendu vrai à un temps donné par l'effondrement d'un des ponts, un effondrement ultérieur sera également considéré comme ayant pour effet la véracité de ce fluent.
À travers cet exemple nous venons de montrer à quel point les frontières entre cas peuvent être sensibles à la représentation, notamment celles entre les types des deux dernières colonnes du tableau 5.2.

Nous montrons maintenant que les frontières entre les deux premières colonnes sont tout aussi sensibles à la représentation. Cette porosité est très bien expliquée par WRIGHT [2011]. Considérons le cas classique où un assassin A empoisonne la boisson de sa cible, mais qu'un assassin B tue la victime avec son arme à feu avant même que le poison puisse commencer à faire effet. Dans cette situation, la granularité de la représentation joue également un rôle très important. Selon une première représentation nous pourrions avoir un cas de préemption tardive ; le processus d'empoisonnement n'aboutit pas, il est interrompu par la mort de la victime comme dans l'exemple de la figure 5.6.
Toutefois, en représentant plus finement le problème et en faisant apparaître les processus biologiques sous-jacents, nous pourrions avoir un cas de préemption précoce. En effet, dans les détails, le chemin causal déclenché par l'assassin B a interrompu celui de l'empoisonnement déclenché par l'assassin A. La blessure causée par l'assassin B déclenche une

hémorragie et l'arrêt du coeur, deux occurrences qui empêchent le poison de se répandre dans le corps de la victime et donc d'agir. Ce cas peut donc être considéré du même type que celui de la figure 5.5.
La réécriture proposée par HALPERN [2016] de l'exemple 3.5 de Suzy et Billy consiste à ajouter deux variables au problème. Cet ajout est tout simplement une modification de la granularité dans la représentation du problème ; il transforme l'exemple d'origine qui est un cas de préemption tardive en un cas de préemption précoce.

Nous montrons maintenant que la sensibilité des frontières à la représentation n'est pas juste entre les deux premières colonnes et entre les deux dernières colonnes, les frontières entre les deux premières colonnes et les deux dernières le sont également. Avec le cas du bateau nous avons vu qu'un cas imitatif peut devenir un cas de préemption tardive. Ce n'est pas le seul cas de figure. En premier lieu, si la représentation du monde ne permet pas de considérer qu'un évènement a un fluent comme effet si celui était déjà vrai lorsque l'évènement se produit, alors les cas imitatifs deviennent des cas de préemption précoce.
En second lieu, modifier les préconditions de e_ψ peut faire d'un cas duplicatif asynchrone un cas de préemption tardive et vice versa. En l'occurrence, dans l'exemple 5.1, la présence ou non de $\neg d$ dans $pre(e_\psi)$ fait toute la différence. Qui plus est, cet exemple est parfait pour montrer qu'un tel choix peut modifier l'intuition sur les relations causales du cas étudié. Ce qui est considéré comme une cause n'est potentiellement pas pareil si l'électrocution entraîne la mort ou non. Alors que dans le premier cas, le premier agent qui ferme l'interrupteur préempte n'importe quel autre chemin causal ayant pu causer la mort a posteriori, car la victime doit être en vie pour être exécutée $\neg d \in pre(e_\psi)$; dans le deuxième cas, fermer l'interrupteur après un autre agent cause que du courant puisse passer par la branche en question et atteigne la victime, permettant donc des cas autres que la préemption.

La plupart des choix de représentation dont nous avons discuté dans cette section ne deviennent explicites qu'au moment de représenter l'exemple. Autrement, ces hypothèses restent implicites tout en influençant quand même l'intuition causale qui nous guide vers une réponse. Ayant montré la sensibilité de la causalité à la représentation, comme c'est le cas également pour l'éthique, il paraît d'autant plus nécessaire que dans cette thèse nous utilisions un langage permettant de rendre explicites les nuances, aussi bien de la causalité que de l'éthique.

5.3.2 Quelques propriétés pour qualifier et comparer les approches

Comme l'indique la citation de BECKERS [2021a] en début de chapitre, la plupart des articles dans le domaine de la causalité effective proposent une nouvelle définition de ce qu'est la causalité effective, puis essayent de montrer que leur définition est plus intuitive que les autres en la confrontant à un ensemble d'exemples de surdétermination connus comme étant complexes. D'après ce même auteur, cette stratégie est vouée à l'échec : « It is unrealistic to expect that this [...] strategy in and on itself can deliver a satisfactory account of causation, because there are too many examples and even more intuitions ». La question de savoir laquelle des définitions de causalité est la plus adéquate est en dehors du cadre de ce chapitre. Elle a été abordée dans le chapitre 3 et sera de nouveau abordée dans le chapitre suivant. Toutefois, nous partageons l'avis énoncé ci-dessus et souhaitons contribuer à la clarification du domaine.

Grâce à la typologie présentée dans la section 5.2.2, nous proposons d'aller au delà d'exemples individuels et de généraliser la comparaison entre approches. Ayant des définitions claires des différents types de cas de surdétermination, il est possible d'établir des propriétés qui caractérisent la façon dont une définition de causalité va considérer un type de surdétermination. Pour reprendre une formulation précédente, si pour un premier exemple une théorie trouve que (e_m^1, t_m^1) est une cause et pas (e_n^2, t_n^2), alors, pour un autre exemple classé comme étant du même type que le premier par notre typologie, cette théorie devrait donner la même réponse. Plutôt que de confronter les différentes approches à de multiples exemples pas nécessairement représentatifs de l'ensemble des cas possibles, nous pouvons à présent prouver quelles seront les occurrences d'évènements considérées par l'approche comme des causes, et cela pour tous les exemples d'un même type. Faire ce travail pour les six catégories proposées permet de couvrir les exemples de la littérature et permet de comparer plus facilement les approches entre elles. Voici un exemple de comment il serait possible de caractériser les différentes approches :

Définition 5.14 [Sensibilité des approches causales]**.** *Une approche en causalité effective est sensible à la surdétermination :*

- *préemptive (définition 5.8 et 5.9) : si dans ces cas elle considère* $(e_m^1, t_m^1) \in \omega^1$ *une cause effective de* (e_ψ, t_ψ)*, contrairement à* $(e_n^2, t_n^2) \in \omega^2$.
- *duplicative et symétrique (définition 5.10, 5.11 et 5.12) : si dans ces cas elle considère aussi bien* $(e_m^1, t_m^1) \in \omega^1$ *que* $(e_n^2, t_n^2) \in \omega^2$ *comme des causes effectives de* (e_ψ, t_ψ).

Notez que la définition 5.14 est simplement un exemple de ce qu'une approche causale pourrait considérer comme des causes dans chaque catégorie. Il s'agit là d'un point de vue parmi tous ceux qui peuvent être exprimés en s'appuyant sur la typologie proposée.

5.4 Conclusion

Dans ce chapitre nous avons apporté nos premières contributions au domaine de la causalité effective. Nous avons commencé par présenter la modélisation du monde dans laquelle s'inscrit cette première contribution de clarification. Il s'agit d'un STEE choisi par sa généralité de sorte à ce que toute représentation de l'action, du changement et de la causalité suffisamment proche puisse facilement trouver une équivalence avec notre modélisation. Puis, nous avons ensuite proposé une définition formelle de surdétermination et de chemin causal, deux concepts rarement définis, mais qui s'avèrent être indispensables pour pouvoir identifier précisément les différents cas de surdétermination. Ensuite, nous avons fait une proposition de typologie qui permet de classifier clairement les cas de surdétermination de la littérature dans six catégories distinctes. En nous appuyant sur cette typologie, nous avons montré la sensibilité de la causalité à la représentation, et donc à quel point cette étape est importante. Nous avons également pu mettre en évidence l'existence de tout un ensemble de cas intéressants qui a notre connaissance n'ont pas encore été explorés dans la littérature. Finalement, nous avons proposé une nouvelle façon de comparer les différentes approches causales entre elles.
Les clarifications qui ont été apportées dans ce chapitre sont d'une grande utilité pour le chapitre suivant dans lequel nous proposons une approche causale adaptée aux problèmes de l'éthique computationnelle.

Chapitre 6

Contribution : modélisation, représentation et automatisation de la causalité positive

« *Some [of the popular examples of actual causation formulated in philosophical literature] sound deceptively simple, but faithful modelling of them requires time, concurrency and natural actions.* »

BATUSOV et SOUTCHANSKI [2018]

Sommaire

La contribution de cette thèse au domaine de l'éthique computationnelle est principalement faite par le biais de la causalité. En effet, dans la section 4.4 il a été montré que la causalité était une pièce fondamentale lorsqu'il est question de formaliser une grande partie des théories morales vues dans le chapitre 1. Dans la section 3.2 il a été montré que les approches en causalité effective existantes ne permettent pas de répondre à un certain nombre d'attentes propres à l'éthique computationnelle. Il a donc été décidé de proposer une nouvelle approche causale commune pour l'éthique computationnelle permettant d'y répondre.

Ce chapitre se veut une présentation détaillée du cœur de l'approche que nous proposons. La présentation de l'ensemble de notre approche causale est distribuée entre ce chapitre et le chapitre 7. Ces deux chapitres correspondent à notre proposition de représentation de l'action, du changement et de la causalité. Cette proposition peut par exemple venir remplacer $\mathbb{AC}$ dans le cadre modulaire $\mathbb{ACE}$. L'ensemble de notre approche peut être vue comme un escalier à quatre marches où chaque marche correspond à un des défis dans le domaine identifiés dans la section 3.2, et surtout chaque marche prend appui sur les précédentes. Si nous parlons pour ce chapitre de cœur de l'approche, c'est parce que cette première marche représente les fondations essentielles au développement de toute autre marche.
Cette première marche traite de causalité positive. Une fois construite elle permet de s'interroger sur les causes de la véracité d'une partie de l'état du monde ou sur les causes qu'un évènement se soit produit. C'est en quelque sorte l'objectif premier lorsque nous pensons à la causalité. Les exemples utilisés pour illustrer la surdétermination dans le chapitre 5 s'intéressent tous à la causalité positive.

La construction de cette première marche demande l'introduction de plusieurs éléments. Le premier d'entre eux est un système de transition d'états étiqueté. Contrairement à $\mathscr{S}_e$ et $\mathscr{S}_s$, celui-ci ne sera pas le plus général possible mais une proposition que nous faisons où un certain nombre de choix ont été faits de sorte à ce qu'il convienne aux attentes propres à l'éthique. S'agissant d'une proposition bien spécifique ayant vocation à être implémentée, nous introduisons un langage de description d'action décrivant ce STEE. Cette représentation de l'action et du changement peut par exemple venir remplacer $\mathbb{A}$ dans le cadre modulaire $\mathbb{ACE}$.
Le deuxième élément est la définition de causalité positive que nous adoptons. Il consiste en une définition de ce que nous avons appelé dans le chapitre 5 des $\mathscr{F}$-causes et des causes effectives. La base de notre causalité sont les $\mathscr{F}$-causes directes, à partir de cette relation nous construisons le reste. Celle-ci se base sur le test NESS de Wright [2011] qui, comme nous l'avons vu dans la section 3.2.1.1, est la définition la plus adaptée à notre objectif par sa factualité due à la séparation forte qu'elle établit entre causalité et responsabilité. La définition des $\mathscr{F}$-causes directes que nous proposons peut donc être vue comme une représentation du test NESS dans le langage de description d'action que nous utilisons. La mise en commun de ces deux premiers éléments est la continuation logique de travaux ayant pour but de connecter les langages de description d'action et la causalité effective, comme ceux de Batusov et Soutchanski [2018]; Berreby et collab. [2018]; Hopkins et Pearl [2007]; LeBlanc et collab. [2019].
Le troisième et dernier élément est une implémentation complète et correcte en Answer Set Programming de la définition de causalité positive que nous proposons. Même si principalement pensé pour l'éthique computationnelle, l'implémentation de ce pont entre langages

de description d'action et causalité est un programme plus général permettant l'exploration de relations causales complexes dans un cadre de prise de décision. Le chapitre 8 est un exemple d'application de ces trois éléments pour une utilisation autre que l'éthique computationnelle. Plus exactement, il présente les bénéfices de l'utilisation de notre proposition pour l'explicabilité dans le domaine de l'argumentation abstraite.

Ce chapitre est divisé en quatre sections. La section 6.1 introduit le langage de description d'action que nous proposons. La section 6.2 présente la définition de causalité positive que nous proposons et montre ses propriétés par rapport à la typologie de la surdétermination du chapitre 5. Pour établir ces propriétés, nous montrons que $\mathscr{S}_c$ peut être vu comme une spécification du $\mathscr{S}_s$ introduit dans le chapitre 5. La section 6.3 décrit l'implémentation en Answer Set Programming de la définition de causalité positive que nous proposons et prouve qu'elle est complète et correcte. Finalement, la section 6.4 discute de l'expressivité de notre langage de description d'action par rapport à STRIPS et propose une version plus expressive ainsi qu'un pont entre PDDL et les langages de description d'action proposés. Les travaux présentés dans ce chapitre font l'objet d'une publication en 2022 dans le cadre de la « International Conference on Principles and Practice of Multi-Agent Systems » (PRIMA 2022) [SARMIENTO et collab., 2022] et de la soumission d'une extension au « Journal of Artificial Intelligence Research » (JAIR) [SARMIENTO et collab., 2023] sous modifications suite à de premiers retours encourageants.

6.1 Modélisation et représentation : langage de description d'action pour le raisonnement causal en éthique computationnelle $\mathscr{S}_c$

Le STEE, et le langage de description d'action le décrivant dont nous allons parler dans cette section, ont été conçus pour répondre à la nécessité de rendre explicites les nuances, aussi bien de la causalité que de l'éthique. Cette nécessité s'explique par la sensibilité de la causalité et de l'éthique à la représentation que nous avons montré dans les chapitres 1, 3 et 5. Nous notons $\mathscr{S}_c$ cette représentation de l'action et du changement, où l'indice c fait référence à la causalité.

Comme pour le chapitre 5, nous utilisons un exemple qui est déroulé tout au long du chapitre et qui est complété par d'autres si besoin. Pour cet exemple nous avons choisi de reprendre l'exemple 3.1 introduit plus tôt dans le document et qui traite de pollution. Pour le confort du lecteur, nous le rappelons ci-dessous.

Exemple 6.1 [pollution]**.** *Un village situé le long d'une rivière abrite plusieurs familles. L'eau potable utilisée par les habitants du village provient d'une usine de potabilisation qui puise l'eau dans la rivière, elle-même provenant d'un lac situé en amont. Cependant, la capacité de cette usine est limitée, elle ne peut traiter l'eau que si elle a un indice de pollution strictement inférieur à un seuil. Lorsque l'eau de la montagne atteint le lac, l'indice de pollution est nul. Il existe deux sources potentielles de pollution du lac : (i) les eaux usées industrielles d'une usine qui produit des enceintes connectées pour un célèbre site d'achat en ligne et qui donne du travail à au moins un membre de chaque famille du village ; (ii) les eaux usées industrielles d'une usine qui produit des médicaments vitaux pour des patients et qui a complètement automatisé sa ligne de production. En temps normal, les eaux usées industrielles de l'usine de médicaments ne polluent pas le lac car elles sont traitées rigoureusement par une station*

d'épuration avant d'y être déversées. Nous supposons que les déversements sont nécessaires au fonctionnement des deux usines et que la gestion de leur production est assurée par un agent.

Le langage de description d'action dont nous parlons dans cette section a été conçu pour permettre de déterminer l'évolution du monde étant donné un ensemble d'actions choisies délibérément par des agents. Autrement dit, il s'agit ici de projection temporelle, un type particulier de raisonnement temporel. Le choix de considérer ces actions comme une entrée s'explique par l'utilisation éthique que nous voulons faire de $\mathscr{S}_c$. En effet, comme exposé dans le chapitre 1, évaluer moralement une action consiste à déterminer son statut déontique. Les actions, et plus exactement les ensembles d'actions que sont les décisions, est ce que nous cherchons à évaluer moralement. L'occurrence de ces actions pouvant entraîner une réaction en chaîne d'évènements dits « naturels », il est donc nécessaire pour avoir une connaissance complète de l'évolution du monde, et donc de l'impact des actions, de s'intéresser aussi bien à l'évolution des états du monde qu'à l'occurrence des évènements. Notez que contrairement à $\mathscr{S}_e$ et $\mathscr{S}_s$, nous avons dans ce langage une distinction claire entre deux types d'évènements : des actions et des évènements naturels. Le premier est utilisé pour faire référence aux choix délibérés des agents, des évènements qui dépendent de la volition d'un agent et qu'il fait donc sens d'évaluer moralement. Le deuxième est plutôt utilisé pour faire référence à des évènements propres à l'aspect physique du monde, indépendants de la volonté des agents et donc qu'il ne fait pas sens d'évaluer moralement. L'emploi du terme « *évènements* » pour décrire l'ensemble des transitions, connotant la possibilité d'actions sans agents [RUSSELL et NORVIG, 2010, chap 12], prend ici tout son sens.
Comme $\mathscr{S}_s$, $\mathscr{S}_c$ adopte une structure classique. D'un côté, l'état du monde peut être défini comme un ensemble de variables le décrivant. Nous notons cet ensemble de fluents $\mathbb{F}$. De l'autre, les transitions entre états du monde sont le résultat de l'occurrence d'un ensemble d'évènements. Nous notons l'ensemble d'évènements $\mathbb{E}$. L'évolution du monde peut être vue comme une suite d'états qui s'enchaînent les uns après les autres au fur et à mesure que des évènements se produisent.

La présentation de notre langage est faite en quatre parties. La section 6.1.1 introduit la représentation des états du monde, la section 6.1.2 la représentation des évènements, la section 6.1.3 les transitions entre états et la section 6.1.4 le passage du STEE obtenu aux traces décrivant l'évolution du monde.

6.1.1 États

Un état du monde peut être défini par un ensemble de fluents qui le décrivent. Un *littéral de fluent* est soit un fluent $f \in \mathbb{F}$, ou sa négation $\neg f$. L'ensemble des littéraux de fluents dans $\mathbb{F}$ est noté $Lit_{\mathbb{F}}$, défini par $Lit_{\mathbb{F}} = \mathbb{F} \cup \{\neg f \mid f \in \mathbb{F}\}$. Le complément d'un littéral de fluent $l \in Lit_{\mathbb{F}}$, noté $\overline{l}$, est défini comme $\overline{l} = \neg f$ si $l = f$, ou $\overline{l} = f$ si $l = \neg f$. Par extension, le complément d'un ensemble de littéraux $\mathrm{L} \subseteq Lit_{\mathbb{F}}$, noté $\overline{\mathrm{L}}$, est défini comme $\overline{\mathrm{L}} = \left\{\overline{l} \mid l \in \mathrm{L}\right\}$.

Définition 6.1 [état S]**.** *L'ensemble* $\mathrm{L} \subseteq Lit_{\mathbb{F}}$ *est un état si :*

- *il est* cohérent *:* $\forall l \in \mathrm{L}, \overline{l} \notin \mathrm{L}$*;*
- *il est* complet *:* $\forall f \in \mathbb{F}, f \in \mathrm{L}$ *ou* $\neg f \in \mathrm{L}$.

Un état dans $\mathscr{S}_c$ est donc un ensemble $\mathrm{L} \subseteq Lit_{\mathbb{F}}$ donnant la valeur de chaque fluent décrivant le monde. Dans notre cadre, un ensemble L incohérent ne peut pas décrire la réalité.

L'indice de pollution ne peut pas avoir atteint le seuil et en même temps ne pas l'avoir atteint. Par contre, en l'absence d'information ou pour certaines notations, nous pourrons utiliser des descriptions partielles du monde. Nous parlons alors d'*états partiels* qui sont décrits par un ensemble $L \subseteq Lit_{\mathbb{F}}$ cohérent mais incomplet.

Nous notons $\mathscr{F}$ les *formules d'état*, qui pour rappel sont des formules de fluents construites en utilisant les opérateurs logiques classiques. Dans $\mathscr{S}_c$ les formules d'état peuvent prendre la forme suivante :

$$\psi ::= l | \psi_1 \wedge \psi_2 | \psi_1 \vee \psi_2$$

où $l \in Lit_{\mathbb{F}}$ et les opérateurs logiques $\wedge$ et $\vee$ ont leur sémantique classique de la logique du premier ordre. Étant donné une formule $\psi \in \mathscr{F}$ et un état partiel ou un état $L \subseteq Lit_{\mathbb{F}}$, la relation $L \vDash \psi$ est définie classiquement :

— $L \vDash l$ si $l \in L$;

— $L \vDash \psi_1 \wedge \psi_2$ si $L \vDash \psi_1$ et $L \vDash \psi_2$;

— $L \vDash \psi_1 \vee \psi_2$ si $L \vDash \psi_1$ ou $L \vDash \psi_2$.

Exemple 6.1 [suite]**.** *Les littéraux $o_e, o_m \in Lit^2_{\mathbb{F}}$ correspondent à l'existence d'eaux usées industrielles provenant respectivement de l'usine produisant des enceintes et de celle produisant des médicaments. Les littéraux $t_n, m_k \in Lit^2_{\mathbb{F}}$ décrivent respectivement que n personnes du village ont du travail et que k doses de médicaments vitaux sont disponibles. Le littéral $s_{\leq} \in Lit_{\mathbb{F}}$ indique que l'indice de pollution a atteint ou dépassé le seuil et $o_p \in Lit_{\mathbb{F}}$ que les habitants du village ont accès à l'eau potable. Finalement, le littéral $se_{hs} \in Lit_{\mathbb{F}}$ indique que la station d'épuration qui traite les eaux usées de l'usine de médicaments est hors service.*

Notez que comme dans $\mathscr{S}_s$, nous faisons dans $\mathscr{S}_c$ le choix de ne garder que des fluents ontiques et de ne pas inclure de fluents épistémiques. Pour $\mathscr{S}_s$ cette hypothèse se justifiait par le fait que la distinction n'avait pas d'impact sur le problème de surdétermination que nous étudiions. Pour l'objectif premier de cette proposition qui est de raisonner sur la causalité, ce choix n'a pas d'impact puisque cette distinction n'a aucun impact sur les aspects causaux. Toutefois, comme nous l'avons vu dans le chapitre 4, les fluents épistémiques sont nécessaires pour modéliser certaines théories morales. Manipuler ces fluents de façon satisfaisante représente un vrai défi auquel des domaines entiers de recherche sont consacrés. Cela dépasse le cadre de cette thèse, mais reste une voie importante à explorer dans de travaux futurs.

6.1.2 Évènements

Les transitions entre états sont le résultat de l'occurrence d'évènements. Deux ensembles disjoints $\mathbb{A}$ et $\mathbb{N}$ forment une partition de $\mathbb{E}$. Nous adoptons ici une distinction ontologique entre deux types d'évènements [Revaz, 2009]. D'un côté, $\mathbb{A}$ contient les *actions* : « conduites d'un humain (ou d'une entité anthropomorphisée) doté d'une raison d'agir (motif) et d'une intention ». De l'autre, $\mathbb{N}$ contient les *évènements naturels* : « phénomènes se produisant dans la nature sous l'effet d'une cause ». Par exemple, nous pouvons considérer que lâcher un objet est une action, mais que sa chute est due à la gravité qui agit sur l'objet et donc au principe fondamental de la dynamique qui serait un évènement naturel. Nous pouvons considérer qu'empoisonner un être sentient est une action, alors que les mécanismes biologiques qui font que la présence de poison entraîne la mort sont des évènements naturels.

Du point de vue de la représentation et de l'état final obtenu, il est équivalent dans ces exemples d'expliciter ces étapes intermédiaires avec des évènements naturels ou de s'en passer en mettant leurs effets directement dans les effets de l'action. Nous pourrions simplement dire que l'effet de lâcher un objet c'est qu'il tombe et l'effet d'empoissonner un individu c'est qu'il meurt. Toutefois, dans la section 5.3.1 il a été montré que deux représentations différentes du même problème peuvent avoir une influence sur le résultat causal attendu, notamment dans les cas de surdétermination. En outre, dans la section 4.4, il a aussi été montré que deux représentations différentes du même problème peuvent avoir une influence sur l'évaluation éthique qui est faite d'une action. Causalité et éthique étant le centre de cette thèse, nous ne pouvons proposer une représentation de l'action et du changement qui ne permette pas la nuance.
Les évènements naturels sont utilisés dans d'autres formalismes de représentation de l'action et du changement. En l'occurrence, ils existent en tant que « `:event` » en PDDL+ [Fox et Long, 2006] et sont modélisables avec des « triggered axioms » en Calcul des Évènements [Mueller, 2014]. Par ailleurs, l'adoption de cette dichotomie est ce qui nous distingue des langages de description d'action comme $\mathscr{A}_c$ [Baral et Gelfond, 1997], une version étendue de $\mathscr{A}$ qui autorise la cooccurrence.

Un évènement $e \in \mathbb{E}$ est une formule atomique. Comme vu précédemment, en causalité il est important de comprendre pourquoi un évènement e peut se produire et quelles sont ses conséquences. C'est pourquoi nous utilisons trois composantes pour caractériser les évènements : des *conditions de déclenchement* donnant toutes les conditions devant être satisfaites par l'état S pour que l'évènement puisse se produire ; des *préconditions* indiquant les fluents décrivant l'état physique du monde devant être satisfaits pour que l'évènement puisse se produire ; des *effets* précisant les changements de l'état du monde attendus si l'évènement se produit.
Dans $\mathscr{S}_c$ nous parlerons surtout de conditions de déclenchement pour les évènements naturels et exclusivement de préconditions pour les actions. Pour comprendre ce choix il faut revenir à la dichotomie entre les deux types d'évènements. Comme expliqué par Revaz [2009] l'occurrence d'un évènement naturel « [...] peut être expliquée par des lois alors que l'agir humain ne peut être que compris, c'est-à-dire interprété ». Si nous transposons à notre langage de description d'action, d'un côté, le déclenchement d'un évènement naturel est entièrement déterminé par la satisfaction de certaines conditions, d'un autre, le déclenchement d'une action repose sur la satisfaction de certaines conditions mais nécessite en plus d'autres facteurs propres à la volition de l'agent qui l'effectue. Contrairement au déclenchement des évènements naturels, celui des actions ne peut donc pas être entièrement expliqué par l'état du monde décrit avec des fluents ontiques. Ayant décidé de ne pas considérer l'existence de fluents épistémiques dans le cadre de cette thèse, il n'est pas possible d'exprimer les conditions de déclenchement pour une action, uniquement ses préconditions. L'aspect volitionnel manquant est donné comme une entrée du langage. Nous en parlons plus en détail dans la section 6.1.4.
Formellement, les fonctions qui associent aux évènements leurs préconditions, leurs conditions de déclenchement et leurs effets sont respectivement définies comme :

$$pre : \mathbb{E} \to \mathscr{F} \qquad tri : \mathbb{N} \to \mathscr{F} \qquad eff : \mathbb{E} \to 2^{Lit_{\mathbb{F}}}.$$

Pour résumer, d'un côté nous avons $\mathbb{A}$ qui contient les actions qui se déclenchent si leurs préconditions sont satisfaites $S \vDash pre(a)$ et si elles sont réalisées par un agent. Elles sont

donc soumises à leur volition. D'un autre côté nous avons $\mathbb{N}$ qui contient les évènements naturels qui se déclenchent aussitôt que leurs conditions de déclenchement sont satisfaites $S \models tri(en)$, sans qu'un agent n'ait besoin de les réaliser. Notez que pour les évènements naturels, il n'y a pas de différence entre les préconditions et les conditions de déclenchement $tri(en) = pre(en)$. En effet, le déclenchement de ces évènements est indépendant de tout état mental d'un agent, il ne dépend que de l'état physique du monde.

Exemple 6.1 [suite]**.** *L'évènement naturel correspondant au dysfonctionnement de l'usine de potabilisation est noté $dis_p \in \mathbb{N}$. Il est déclenché lorsque l'indice de pollution a atteint ou dépassé le seuil $s_{\leq}$ et a pour effet que les habitants du village n'ont plus accès à l'eau potable $\overline{o_p}$. L'évènement naturel correspondant au déversement des eaux usées industrielles existantes dans le lac est noté $dev_o \in \mathbb{N}$. Il est déclenché lorsqu'il y a des eaux usées industrielles provenant de l'usine d'enceintes o_e ou lorsqu'il y a des eaux usées industrielles provenant de l'usine de médicaments o_m et qu'elles ne sont pas traitées par la station d'épuration car hors service se_{hs}. Cet évènement a pour effet que l'indice de pollution atteigne ou dépasse le seuil $s_{\leq}$ et qu'il n'y ait plus d'eaux usées industrielles $\overline{o_m}, \overline{o_e}$.*
Les actions correspondant à lancer la production de médicaments ou d'enceintes sont respectivement notées $prod_m, prod_e \in \mathbb{A}^2$. Nous supposons que ces actions n'ont pas de préconditions, elles peuvent être déclenchées par l'agent en charge de la gestion à n'importe quel instant. La première action a pour effet la création d'eaux usées industrielles o_m et la disponibilité de k doses de médicaments vitaux m_k. La deuxième a pour effet la création d'eaux usées industrielles o_e et que n personnes du village aient du travail t_n. Toute l'information donnée ci-dessus peut être représentée ainsi :

$$pre(prod_m) = \top,\ \mathit{eff}(prod_m) = \{m_k, o_m\};$$
$$pre(prod_e) = \top,\ \mathit{eff}(prod_e) = \{t_n, o_e\};$$
$$tri(dis_p) = s_{\leq},\ \mathit{eff}(dis_p) = \{\overline{o_p}\};$$
$$tri(dev_o) = o_e \vee (o_m \wedge se_{hs}), \mathit{eff}(dev_o) = \{s_{\leq}, \overline{o_m}, \overline{o_e}\}.$$

La figure 6.1 illustre la représentation que nous venons de décrire de l'exemple 6.1 dans $\mathscr{S}_c$.

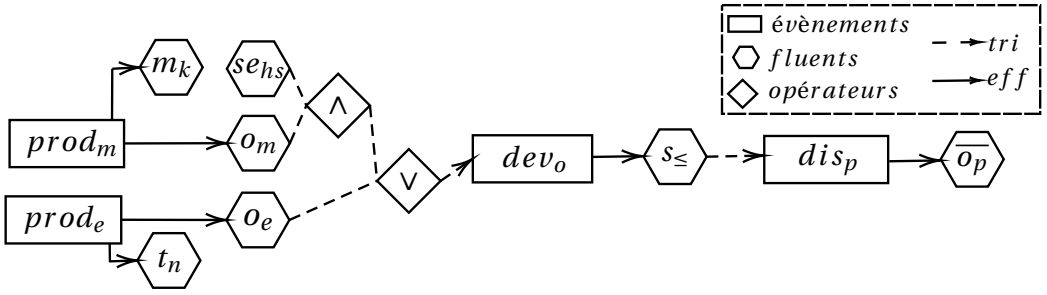


FIGURE 6.1 – Illustration de la représentation de l'exemple 6.1 dans $\mathscr{S}_c$.

Comme montré dans la section 3.1.2.1, lorsqu'il est question de surdétermination, il est inévitablement question de préconditions disjonctives et très souvent de cooccurrence d'évènements. $\mathscr{S}_c$ permet les préconditions disjonctives par la forme des formules d'état $\mathscr{F}$. Pour la cooccurrence d'évènements, nous notons $E \subseteq \mathbb{E}$ un ensemble d'évènements et considérons que $pre(E) = \bigcup_{e \in E} pre(e)$ et que $\mathit{eff}(E) = \bigcup_{e \in E} \mathit{eff}(e)$.

Toutefois, tous les évènements ne peuvent pas être faits en cooccurrence avec un autre. Nous considérons que deux évènements $e, e' \in \mathbb{E}^2$ sont *interférents* entre eux si l'ensemble $\{l | l \in \mathit{eff}(e) \cup \mathit{eff}(e')\}$ n'est pas cohérent au sens de la définition 6.1. Pour résoudre des conflits potentiels, nous choisissons de permettre d'établir des priorités entre le déclenchement d'évènements. Pour cela, un ordre partiel strict $\succ_{\mathbb{E}}$ est introduit ; il garantit la priorité de déclenchement d'un évènement par rapport à un autre.

6.1.3 Transitions entre états

L'évolution du monde peut être vue comme une suite d'états qui s'enchaînent les uns après les autres au fur et à mesure que des évènements se produisent. Dans cette section nous allons voir les règles qui régissent ces enchaînements.

Définition 6.2 [Système de transition d'états étiqueté $\mathscr{S}_c$]**.** *Le système de transition d'états étiqueté $\mathscr{S}_c$ est un triplet $\langle 2^{Lit_{\mathbb{F}}}, 2^{\mathbb{E}}, \tau \rangle$ où τ est l'ensemble des relations étiquetées de transition entre états notées* (S, E, S')*. Les triplets de cet ensemble vérifient :*

1. $S \subseteq Lit_{\mathbb{F}}$ *est un état au sens de la définition 6.1 ;*
2. $E \subseteq \mathbb{E}$ *vérifie :*
   (a) $\forall e \in E, S \vDash pre(e)$;
   (b) $\nexists (e, e') \in E^2,\ e \succ_{\mathbb{E}} e'$;
   (c) $\forall e \in \mathbb{E}$ *tel que* $S \vDash tri(e), e \in E$ *ou* $\exists e' \in E, e' \succ_{\mathbb{E}} e$;
3. $S' = \left\{ l \in S | \forall e \in E, \bar{l} \notin \mathit{eff}(e) \right\} \cup \left\{ l \in Lit_{\mathbb{F}} | \exists e \in E, l \in \mathit{eff}(e) \right\}$.

La condition 2a assure que tous les évènements dans l'ensemble E ont leurs préconditions satisfaites par l'état S. Nous dirons que (S, E, S') est *exécutable*. La condition 2b assure que toutes les priorités entre les évènements ont été respectées, cela empêche que deux évènements interférents puissent avoir lieu en même temps. Nous dirons que (S, E, S') est *correcte en cooccurrence*. La condition 2c assure que tous les évènements qui devaient être dans E le sont bien. Si un évènement a ses conditions de déclenchement satisfaites par S, alors il doit être dans E, à moins qu'un autre évènement de E lui soit prioritaire. Nous dirons que (S, E, S') est *correcte en déclenchement*. L'ensemble des relations étiquetées de transition entre états (S, E, S') dans τ sont dites *valides*, i.e. elles sont exécutables, correctes en cooccurrence et correctes en déclenchement.

Dans notre cadre déterministe, la condition 3 spécifie comment est déduit l'état suivant S' à partir uniquement de l'état précèdent S et des évènements qui s'y produisent E. Cette condition contient quelques subtilités qu'il est important d'aborder car elles ont une influence sur la causalité. Notamment, il faut noter que selon l'état S où un évènement e se produit, il peut ne pas avoir réellement tous les effets avec lesquels il a été représenté $\mathit{eff}(e)$, que nous appellerons dorénavant ses *effets intrinsèques*. Il peut en avoir plus ou moins. Nous allons voir pourquoi.
Comme mentionné lors de leur présentation dans la section 6.1.2, les effets indiquent simplement les changements de l'état du monde « attendus » si l'évènement se produit. Prenons l'exemple d'un agent qui veut allumer la lumière dans une pièce en appuyant sur un interrupteur. Dans un premier scénario il est possible que l'agent cause un court-circuit et déclenche un incendie car l'installation électrique était défectueuse. Lorsque l'action consistant à appuyer sur l'interrupteur est représentée, il n'est pas intuitif de prendre en compte

le court-circuit puis l'incendie comme faisant partie de ses effets intrinsèques. En plus d'affecter la généralité de la représentation, nous avons montré dans la section 4.4 que cela pourrait fausser l'évaluation éthique. Dans ces cas de figure, il est préférable de décomposer le processus en introduisant des évènements naturels. Dans notre exemple, il y aurait un évènement naturel correspondant au court-circuit ; ses effets seraient le déclenchement d'un incendie et ses conditions de déclenchement seraient une installation électrique défectueuse et le fait que le circuit soit fermé, donc que du courant y circule. Après l'insertion d'un tel évènement naturel, il est nécessaire de pouvoir construire une chaîne causale pour relier le fait d'appuyer sur l'interrupteur avec l'incendie. Une relation causale étant une relation binaire qui lie une cause à une conséquence, nous nous référons à tous les effets qu'un évènement a réellement eu comme ses *conséquences*. Dans le chapitre 4 nous avons montré que certaines théories morales évaluent les actions en fonction de leurs conséquences. Pour pouvoir évaluer éthiquement une action qui a plus de conséquences que d'effets intrinsèques avec ces théories, il est nécessaire de pouvoir raisonner causalement.
Toujours dans le même exemple, mais dans un deuxième scénario, il est possible que l'agent appuie sur l'interrupteur mais que la lumière soit déjà allumée. Cela n'empêche pas l'agent de réaliser l'action, celle-ci peut avoir d'autres conséquences que nous devons prendre en compte, comme modifier l'état du circuit électrique. Toutefois, l'état allumé de la lumière ne fait pas partie de ses conséquences étant donné l'état S dans lequel l'action a été réalisée. Ce deuxième scénario est possible étant donné que dans la condition 3 hypothèse est faite que la loi d'inertie s'applique à tous les fluents. Autrement dit, la valeur de vérité des fluents reste la même d'un état à l'autre tant qu'un évènement ne la change pas. L'inertie est prise en compte grâce au terme $\left\{l \in \mathrm{S} | \forall e \in \mathrm{E}, \bar{l} \notin \mathit{eff}(e)\right\}$. Si l'état de la lumière n'était pas soumis à la loi d'inertie, la lumière allumée à l'état S serait éteinte à l'état S′ à moins qu'un évènement dans E ne la maintienne allumée. Comme mentionné dans le chapitre 2, l'hypothèse inertielle est très souvent adoptée car elle permet de résoudre efficacement le problème du cadre formulé par McCarthy et Hayes [1969]. En définitive, pour pouvoir évaluer éthiquement une action qui a moins de conséquences que d'effets intrinsèques avec ces théories morales qui évaluent les actions en fonction de leurs conséquences, il est également nécessaire de pouvoir raisonner causalement.

Déterminer toutes les conséquences d'un évènement n'est pas une question anodine, la section 6.2 est entièrement consacrée à cela. Toutefois, il est déjà possible au niveau du $\mathscr{S}_c$ d'introduire des outils permettant d'extraire certaines informations causales simples comme : quelles conséquences un ensemble d'évènements E se produisant à L a parmi ses effets intrinsèques, nous appelons ces conséquences des effets effectifs ; ou, étant donné un état partiel L, quel est l'état partiel obtenu si un ensemble d'évènements E y avait lieu. Ces deux informations sont obtenues grâce au prédicat $\mathit{actual}\mathrm{Eff}(\mathrm{E},\mathrm{L})$ et l'opérateur de mise à jour $\mathrm{L} \triangleright \mathrm{E}$. Notez que si ces opérateurs sont définis pour des états partiels, ils s'appliquent tout aussi bien aux états.

Définition 6.3 [Effets effectifs $\mathit{actual}\mathrm{Eff}(\mathrm{E},\mathrm{L})$]**.** *Étant donné un état partiel* $\mathrm{L} \subseteq \mathit{Lit}_{\mathbb{F}}$ *et un ensemble d'évènements* $\mathrm{E} \subseteq \mathbb{E}$*, le prédicat* $\mathit{actual}\mathrm{Eff}(\mathrm{E},\mathrm{L})$ *associe à un couple* (E,L) *un état partiel étant les effets effectifs de* E*. Il est défini comme :*

$$\mathit{actual}\mathrm{Eff}(\mathrm{E},\mathrm{L}) = \left\{l \in \mathit{Lit}_{\mathbb{F}} | \exists e \in \mathrm{E}, l \in \mathit{eff}(e) \wedge l \notin \mathrm{L}\right\}.$$

Dit autrement, si les évènements de l'ensemble E avaient lieu dans l'état partiel L, leurs effets effectifs seraient l'union des effets effectifs des évènements faits individuellement :

$$actualEff(E,L) = \bigcup_{e \in E} actualEff(\{e\},L).$$

Définition 6.4 [Opérateur de mise à jour $\triangleright$]**.** *Étant donné un état partiel* $L \subseteq Lit_{\mathbb{F}}$ *et un ensemble d'évènements* $E \subseteq \mathbb{E}$*, l'opérateur de mise à jour indique l'état partiel obtenu si* E *avait lieu dans* L*. Il est défini comme :*

$$L \triangleright E = \left(L \setminus \overline{actualEff(E,L)}\right) \cup actualEff(E,L).$$

Les informations données par $actualEff(E,L)$ et $\triangleright$ peuvent être assimilées à des informations causales de base pouvant être extraites directement de la sémantique de $\mathscr{S}_c$. En plus d'être causale, cette information est directionnelle puisqu'il est inconcevable dans notre sémantique de dire que l'effet effectif d'un évènement est la cause de cet évènement. Deux résultats peuvent être déduits de ces constats. Le premier est que les critiques de Lewis [1973] sur les approches logiques qui ont été exposées dans le chapitre 3 ne peuvent pas nous être appliquées. Nous nous inscrivons dans le cadre des approches causales par inférence.
Le deuxième est que nous pouvons nous appuyer sur les évènements qui se produisent et leurs effets effectifs pour simuler l'évolution du monde à partir d'un état donné. La condition 3 de la définition 6.2 pourrait tout simplement s'écrire $S' = S \triangleright E$.

6.1.4 Traces décrivant l'évolution du monde

Comme dans le chapitre 5, du fait que nous proposons un STEE pour la causalité effective, nous nous intéressons à des séquences d'évènements et d'états particulières, et non pas au STEE dans son ensemble. Ces séquences peuvent être vues comme des chemins dans le STEE et un temps $t \in \mathbb{T}$, où $\mathbb{T} = \{-1, 0, \ldots, N\}$, peut être associé à chaque élément de la séquence. Nous notons S_0 l'état initial.
Il s'agit ici d'une représentation bornée dans le passé d'un problème réel qui lui n'est pas borné. Comme nous l'avons vu dans le chapitre 3, tout évènement naturel est considéré comme ayant une cause et il est possible de remonter ainsi jusqu'à la création de l'univers. Pour rester le plus fidèles possible à la conception philosophique de la causalité dans notre représentation du monde, tous les états précédant $t = 0$ sont recueillis dans un état $S_{-1} = Lit_{\mathbb{F}} \setminus S_0$ associé au temps $t = -1$. Celui-ci représente tous les états du monde avant l'état initial. Pour chaque littéral $l \in S_0$ nous introduisons un évènement $ini_l \in \mathbb{N}$ tel que $eff(ini_l) = l$. Nous avons donc le triplet $(S_{-1}, E_{-1}, S_0) \in \tau$ où $E_{-1} = \{ini_l \in \mathbb{N} | l \in S_0\}$ vérifie $S_0 = S_{-1} \triangleright E_{-1}$.

Définition 6.5 [contexte κ_c]**.** *Le contexte noté* κ_c *est l'octuple* $(\mathbb{E}, \mathbb{F}, pre, tri, eff, S_0, \succ_{\mathbb{E}}, \mathbb{T})$*, où* $\mathbb{E}, \mathbb{F}, pre, tri, eff, S_0, \succ_{\mathbb{E}}$ *et* $\mathbb{T}$ *ont été définis précédemment.*

Pour un contexte κ_c donné, il existe potentiellement plus d'un chemin possible dans le STEE. En effet, aucune spécification du moment où les actions sont réalisées n'est inclue dans le contexte et donc tout chemin où les τ sont valides est possible. Dans tous ces chemins il est vérifié que les préconditions des actions sont valides, mais l'aspect volitionnel est manquant. Comme mentionné précédemment, cet élément est donné comme une entrée du langage.

Définition 6.6 [Scénario σ]**.** *Le scénario noté* σ *est un ensemble d'actions couplées à un temps* $\sigma \subseteq \mathbb{A} \times \mathbb{T}$. *Il représente la volition des agents.*

Définition 6.7 [cadre causal χ]**.** *Dans* $\mathscr{S}_c$*, le cadre causal* χ *est le couple* (κ_c, σ)*, où* κ_c *le contexte et* σ *un scénario.*

Les chemins dans un STEE peuvent être représentés comme des traces où $S^{\chi}(t)$ sont les états traversés et $E^{\chi}(t)$ les transitions qui nous y ont emmenés.

Définition 6.8 [Traces d'évènements et d'états τ^e_χ et τ^s_χ]**.** *Étant donné un cadre causal* χ*, la trace d'évènements* τ^e_χ *et la trace d'états* τ^s_χ *sont respectivement les séquences d'évènements* $E^{\chi}(-1), E^{\chi}(0), \ldots, E^{\chi}(N)$ *et d'états* $S^{\chi}(-1), S^{\chi}(0), S^{\chi}(1), \ldots, S^{\chi}(N+1)$ *qui vérifient :*

1. $S^{\chi}(0) = S_0$;
2. $\forall t \in \mathbb{T},\ (S^{\chi}(t), E^{\chi}(t), S^{\chi}(t+1)) \in \tau$;
3. $\forall t \in \mathbb{T}, \forall a \in E^{\chi}(t),\ a \in \mathbb{A} \implies (a,t) \in \sigma$.

Bien que pour un contexte κ_c donné il existe potentiellement plus d'un chemin possible dans le STEE, l'ajout d'un scénario σ fait que, s'il existe un chemin, celui-ci est unique. Ce chemin est celui décrit par les traces τ^e_χ et τ^s_χ.

Proposition 6.1. *Étant donné un cadre causal* χ*, les traces* τ^e_χ *et* τ^s_χ *sont uniques.*

Démonstration. Nous prouvons par l'absurde l'unicité des traces pour un χ donné. Soit $\chi = (\kappa_c, \sigma)$ le cadre causal et τ^e_χ, $\tau^{e'}_\chi$ deux traces d'évènements. Par reductio ad absurdum nous supposons que $\tau^e_\chi \neq \tau^{e'}_\chi$.

Nous notons $E^{\chi}(t)$ et $S^{\chi}(t)$ les éléments associés à τ^e_χ et τ^s_χ, puis $E^{\chi'}(t)$ et $S^{\chi'}(t)$ les éléments associés à $\tau^{e'}_\chi$ et $\tau^{s'}_\chi$. D'après la définition 6.2, $S^{\chi}(t+1)$ est déduit de $S^{\chi}(t)$ et les évènements dans $E^{\chi}(t)$. De la même façon, $S^{\chi'}(t+1)$ est déduit de $S^{\chi'}(t)$ et les évènements dans $E^{\chi'}(t)$. Puis, étant donné que le contexte κ_c est commun à τ^e_χ et $\tau^{e'}_\chi$, alors $S^{\chi}(0) = S^{\chi'}(0)$ et $E^{\chi}(-1) = E^{\chi'}(-1)$. Par conséquent, la première différence entre τ^e_χ et $\tau^{e'}_\chi$ ne pourrait pas se trouver entre deux états, mais uniquement entre deux ensembles d'évènements.

Notons t_0 le premier point temporel où une différence serait observée entre τ^e_χ et $\tau^{e'}_\chi$. Nous aurions $\forall t < t_0$, $E^{\chi}(t) = E^{\chi'}(t)$ et $\forall t \leq t_0$, $S^{\chi}(t) = S^{\chi'}(t)$, i.e. une équivalence sur les deux traces avant ce premier point, et nous aurions à t_0, $E^{\chi}(t_0) \neq E^{\chi'}(t_0)$. Nous pouvons alors déduire que $\forall e \in \mathbb{E}$, $S^{\chi}(t_0) \models pre(e) \Leftrightarrow S^{\chi'}(t_0) \models pre(e)$.
Notons $D = E^{\chi}(t_0) \setminus E^{\chi'}(t_0) \cup E^{\chi'}(t_0) \setminus E^{\chi}(t_0)$, l'ensemble avec toutes les différences entre $E^{\chi}(t_0)$ et $E^{\chi'}(t_0)$. Puis, considérons un évènement $e_0 \in max_{\succ_{\mathbb{E}}}\{D\}$, i.e. l'événement dans l'ensemble D avec la priorité de déclenchement maximale. Sans perte de généralité par la symétrie χ et χ′, nous pouvons considérer que $e_0 \in E^{\chi'}(t_0) \setminus E^{\chi}(t_0)$, soit $e_0 \notin E^{\chi}(t_0)$ et $e_0 \in E^{\chi'}(t_0)$. Deux cas sont alors possibles, soit $e_0 \in \mathbb{N}$ ou $e_0 \in \mathbb{A}$.

i) Nous montrons par l'absurde que $e_0 \notin \mathbb{N}$. Par reductio ad absurdum nous supposons que $e_0 \in \mathbb{N}$. Comme $e_0 \in E^{\chi'}(t_0)$ et que $S^{\chi}(t_0) = S^{\chi'}(t_0)$, alors $S^{\chi}(t_0) \models tri(e_0)$. Par la condition 2c de la définition 6.2, nous savons que $e_0 \notin E^{\chi}(t_0)$ implique $\exists e \in E^{\chi}(t_0)$ tel que $e \succ_{\mathbb{E}} e_0$. Cela est en contradiction avec l'hypothèse que $e_0 \in max_{\succ_{\mathbb{E}}}\{D\}$. Nous pouvons conclure que $e_0 \in \mathbb{N}$ n'est pas possible et donc $e_0 \notin \mathbb{N}$.

ii) Nous montrons par l'absurde que $e_0 \notin \mathbb{A}$. Par reductio ad absurdum nous supposons que $e_0 \in \mathbb{A}$. Par la condition 3 de la définition 6.8, nous savons que $e_0 \in E^{\chi'}(t_0)$ implique $(e_0, t) \in \sigma$. Étant donné que σ est commun à τ^e_χ et $\tau^{e'}_\chi$, alors comme $S^{\chi}(t_0) = S^{\chi'}(t_0)$

il faut que $e_0 \in \mathrm{E}^{\chi}(t_0)$. Cela est à nouveau en contradiction avec l'hypothèse que $e_0 \notin \mathrm{E}^{\chi}(t_0)$ et $e_0 \in \mathrm{E}^{\chi'}(t_0)$. Nous pouvons conclure que $e_0 \in \mathbb{A}$ n'est pas possible et donc $e_0 \notin \mathbb{A}$.

L'hypothèse qu'il existerait un premier point temporel t_0 où une différence serait observée entre τ^e_χ et $\tau^{e'}_\chi$ mène à une contradiction pour les deux cas possibles, (i) et (ii). Il ne nous reste d'autre alternative que de rejeter cette hypothèse. Puisqu'il n'existe pas de premier point temporel t_0 où une différence serait observée entre τ^e_χ et $\tau^{e'}_\chi$, nous devons également rejeter $\tau^e_\chi \neq \tau^{e'}_\chi$. L'unicité des traces étant donné χ est alors prouvée. □

Dorénavant, lorsqu'il sera question d'évènements et d'états, il s'agira de ceux des traces uniques τ^e_χ et τ^s_χ d'un cadre causal χ donné. Ainsi, l'ensemble des évènements qui se sont effectivement produits à l'instant t est $\mathrm{E}^{\chi}(t) = \tau^e_\chi(t)$. De même, l'état réel à l'instant t est $\mathrm{S}^{\chi}(t) = \tau^s_\chi(t)$.

Exemple 6.1 [suite]**.** *L'information donnée jusqu'ici sur l'exemple correspond au contexte* κ_c*. Nous allons maintenant préciser un scénario* σ *possible auquel nous allons nous intéresser par la suite. Il s'agit du cas de surdétermination symétrique décrit dans l'exemple 3.4 où l'agent lance la production des deux usines simultanément. Le scénario correspondant s'écrit* $\sigma = \{(prod_m, 0), (prod_e, 0)\}$*. Voici les traces* τ^e_χ *et* τ^s_χ *qui correspondent au cadre causal* χ *:*

$$\mathrm{S}^{\chi}(-1) = \{\overline{se_{hs}}, m_k, o_m, o_e, t_n, s_{\leq}, \overline{o_p}\},$$
$$\mathrm{E}^{\chi}(-1) = \left\{ini_{se_{hs}}, ini_{\overline{m_k}}, ini_{\overline{o_m}}, ini_{\overline{o_e}}, ini_{\overline{t_n}}, ini_{\overline{s_{\leq}}}, ini_{o_p}\right\};$$
$$\mathrm{S}^{\chi}(0) = \{se_{hs}, \overline{m_k}, \overline{o_m}, \overline{o_e}, \overline{t_n}, \overline{s_{\leq}}, o_p\}, \mathrm{E}^{\chi}(0) = \{prod_m, prod_e\};$$
$$\mathrm{S}^{\chi}(1) = \{se_{hs}, m_k, o_m, o_e, t_n, \overline{s_{\leq}}, o_p\}, \mathrm{E}^{\chi}(1) = \{dev_o\};$$
$$\mathrm{S}^{\chi}(2) = \{se_{hs}, m_k, \overline{o_m}, \overline{o_e}, t_n, s_{\leq}, o_p\}, \mathrm{E}^{\chi}(2) = \{dis_p\};$$
$$\mathrm{S}^{\chi}(3) = \{se_{hs}, m_k, \overline{o_m}, \overline{o_e}, t_n, s_{\leq}, \overline{o_p}\}.$$

6.2 Modélisation et représentation : le test NESS comme définition de causalité positive

Nous avons à présent $\mathscr{S}_c$, un langage de description d'action adapté au raisonnement causal et éthique. Dans cette section nous proposons une définition de causalité positive pour $\mathscr{S}_c$. Ensemble, ils forment la première marche de notre approche permettant de raisonner sur la causalité. Comme montré dans le chapitre 3, il existe plusieurs approches permettant un tel raisonnement. Toutefois, comme montré dans la section 3.2, ces approches ne sont pas tout à fait adaptées aux problèmes éthiques qui intéressent l'éthique computationnelle. De plus, nous voulons une approche apte à être utilisée pour toutes les théories morales. Celle que nous proposons dans cette thèse a été conçue pour l'être, aussi bien par le langage de description d'action utilisé, que par la définition de causalité dont nous allons maintenant parler.

L'approche que nous adoptons s'inspire majoritairement de travaux dans le droit, et plus spécifiquement dans le droit pénal. De toutes les disciplines où la causalité a fait l'objet d'études, c'est celle où les travaux ont le plus de liens avec les besoins de l'éthique computationnelle. En effet, il y est question d'actions réalisées par des individus dans des situations bien spécifiques et qui rendent potentiellement l'individu qui les a réalisées responsable

d'un préjudice. Deux similitudes entre causalité en droit pénal et en éthique méritent d'être soulignées.
La première est que les actions sont au centre du débat, et il est bien question ici d'actions au sens de la distinction ontologique adoptée dans $\mathscr{S}_c$, i.e. des évènements réalisés par des agents et donc dépendants de leur volition. En droit pénal, l'enquête causale a comme but de déterminer s'il existe une quelconque relation causale entre des actions de l'individu jugé et le préjudice. En éthique, l'enquête causale a comme but de déterminer s'il existe une quelconque relation causale entre des actions et une partie de l'état du monde considéré comme étant bien ou mal. Dans d'autres situations où le raisonnement causal est utilisé, actions et évènements naturels ont la même importance et donc les distinguer ne semble pas nécessaire. Le chapitre 5 en est un exemple. Ce chapitre se voulant une contribution au domaine de la causalité dans sa globalité, cette distinction n'a pas été faite dans $\mathscr{S}_s$.
La deuxième similitude est la place qu'occupe l'enquête causale et les propriétés qu'elle doit avoir. Comme nous l'avons vu dans la section 3.1.2.2, en droit pénal, il est souhaitable qu'une claire distinction soit faite entre causalité et responsabilité. L'enquête causale est une des étapes nécessaires à déterminer s'il existe une responsabilité. Il s'agit de la seule étape qui peut ne pas faire appel qu'à des éléments factuels. Comme nous l'avons vu dans les chapitres 1 et 4, en éthique, une claire distinction est faite entre causalité et statut déontique de l'action. De nouveau, l'enquête causale est simplement là pour fournir des relations causales. Celles-ci seront utilisées différemment en fonction de la théorie morale. De la même façon qu'il y a différents points de vue sur si un individu qui a contribué à causer un préjudice doit être reconnu responsable de celui-ci, chaque théorie morale peut proposer une vision différente de quel doit être le statut déontique d'une action qui a contribué à causer du mal ou du bien. En l'occurrence, la théorie du droit naturel présentée dans la section 1.3.2 défend que l'intentionnalité et la proportionnalité entre bonnes et mauvaises conséquences est à prendre en compte, alors que l'utilitarisme de l'acte présenté dans la section 1.3.1.1 considère que l'intentionnalité n'a aucune importance. Comme pour le droit pénal, en éthique computationnelle il est donc indispensable que la définition de causalité choisie soit purement factuelle et ne fasse pas intervenir des questions de responsabilité. Dans le cas où elle ne l'était pas, cela reviendrait à intégrer un raisonnement causal qui fausserait la formalisation de la théorie morale. Une telle formalisation ne serait pas inutilisable, mais elle ne correspondrait pas réellement à la théorie morale dont elle est censée être la formalisation.

Une relation causale est une relation binaire qui lie une cause à une conséquence. Nous avons vu dans le chapitre 5 que dans un STEE, il est possible d'aussi bien vouloir déterminer les causes de $\psi \in \mathscr{F}$ étant vraie dans un état, ou les causes d'une occurrence d'évènement $e \in \mathbb{E}$. En conséquence, deux types de relations causales différentes sont nécessaires. Nous avons appelé les premières des $\mathscr{F}$-causes et les deuxièmes des causes effectives. Leur différence est illustrée sur la figure 5.2.
Parmi les $\mathscr{F}$-causes nous en avons distingué un type particulier que nous avons qualifié de directes. Nous en faisons la base de notre causalité, à partir de cette relation nous construisons le reste. La définition que nous en donnons est une représentation du test NESS de WRIGHT [2011] dans le langage de description d'action que nous utilisons. Comme nous l'avons vu dans le chapitre 3, cette définition se caractérise par sa factualité et la séparation forte qu'elle établit entre causalité et responsabilité, si bien qu'il s'agit de la définition la plus adaptée à notre objectif.

Cette section est divisée en quatre sections. La section 6.2.1 introduit les NESS-causes directes. Puis, la section 6.2.2 généralise cette définition et introduit les NESS-causes qui correspondent aux $\mathscr{F}$-causes. Reste alors pour finir avec notre définition de causalité positive à définir les causes effectives. C'est ce que fait la section 6.2.3. Finalement, la section 6.2.4 présente les propriétés de notre définition par rapport à la typologie de la surdétermination du chapitre 5. Pour établir ces propriétés, elle montre que $\mathscr{S}_c$ peut être vu comme une spécification du $\mathscr{S}_s$.

6.2.1 NESS-causes directes

Dans cette section nous introduisons les NESS-causes directes, la relation causale à partir de laquelle toutes les autres sont construites. Les NESS-causes directes donnent des informations essentielles en se basant sur les effets que l'occurrence d'un évènement a réellement eus. La définition que nous en donnons est une représentation du test NESS de WRIGHT [2011] dans le langage de description d'action que nous utilisons. Pour rappel, ce test stipule que s'il y a une conséquence ψ et un ensemble de conditions suffisantes W toutes réunies pour que ψ soit vraie, alors une cause c est un élément nécessaire à la suffisance de cet ensemble : « A condition c was a cause of a consequence [ψ] if and only if it was necessary for the sufficiency of a set of existing antecedent conditions [W] that was sufficient for the occurrence of [ψ] ». Comme nous l'avons vu dans le chapitre 3, la particularité de cette définition est qu'elle fait appel à la notion de nécessité, mais que celle-ci est subordonnée à la suffisance qui est au premier plan. Nous allons voir en détail comment.

Avant de passer à la définition, rappelons quelques notations et concepts qui nous sont utiles par la suite. Étant donné χ : une formule vraie dans un état correspond à un couple (ψ, t) qui vérifie $S^{\chi}(t) \vDash \psi$; une occurrence d'évènement correspond à un couple (e, t) qui vérifie $e \in E^{\chi}(t)$. Dans la suite de ce chapitre nous faisons référence à des ensembles d'occurrences d'évènements $C \subseteq \mathbb{E} \times \mathbb{T}$, comme par exemple $C \subseteq \{(e,t) | e \in E^{\chi}(t),\ t \in \mathbb{T}\}$. Lorsque ces ensembles C seront utilisés dans le prédicat $actualEff(E, L)$ ou l'opérateur $L \rhd E$ à la place de l'ensemble E, par abus de notation, nous ferons en réalité référence à l'ensemble d'évènements constitué uniquement des évènements des couples dans C, i.e. $\{e \in \mathbb{E} | \exists t, (e,t) \in C\}$. Plus précisément :

$$S \rhd C \stackrel{\text{def}}{=} S \rhd \{e \in \mathbb{E} | \exists t, (e,t) \in C\} \qquad actualEff(C,S) \stackrel{\text{def}}{=} actualEff(\{e \in \mathbb{E} | \exists t (e,t) \in C\}, S)$$

Le concept mathématique de partition d'ensembles allant être utilisé dans nos définitions de causalité et plusieurs preuves, il est utile de rappeler sa définition exacte.

Définition 6.9 [Partition d'un ensemble]**.** *Étant donné un ensemble* X*, les éléments* $X_1, \ldots, X_k$ *forment une partition de* X *ssi :*

1. $\forall i \in \{1, \ldots, k\}, X_i \neq \varnothing$;
2. $\bigcup_{i \in \{1,\ldots,k\}} X_i = X$;
3. $\forall i, j \in \{1, \ldots, k\}^2, i \neq j \implies X_i \cap X_j = \varnothing$.

Dans la définition 6.10 nous définissons ce qu'est une NESS-cause directe. Toutefois, pour arriver à définir ce qu'est une NESS-cause directe individuellement, il est pertinent de passer par ce que nous appelons un *ensemble suffisant de NESS-causes directes*. Si une NESS-cause directe de (ψ, t_ψ) est un élément qui contribue à la véracité de ψ dans $S^{\chi}(t_\psi)$,

un ensemble suffisant de NESS-causes directes de (ψ, t_ψ) est un ensemble suffisant pour expliquer sa véracité.

Exemple 6.1 [suite]**.** *Reprenons notre exemple de pollution, et plus exactement la condition de déclenchement du déversement des eaux usées* $tri(dev_o) = o_e \vee (o_m \wedge se_{hs})$*. L'occurrence* $(prod_m, 0)$ *contribue à* $(tri(dev_o), 1)$ *en rendant vrai* o_m*. Toutefois, cela n'est pas suffisant pour expliquer* $(tri(dev_o), 1)$*, sans* se_{hs} *la formule de préconditions ne serait pas vraie. Un ensemble suffisant pour expliquer* $(tri(dev_o), 1)$ *est alors* $\{(prod_m, 0), (ini_{se_{hs}}, -1)\}$*. Notez que ce dernier n'est pas le seul ensemble suffisant pour expliquer* $(tri(dev_o), 1)$*, le singleton* $\{(prod_e, 0)\}$ *en est un également. L'occurrence* $(prod_e, 0)$ *contribue à* $(tri(dev_o), 1)$ *en rendant vrai* o_e *et est en même temps un ensemble suffisant pour l'expliquer.*

Définition 6.10 [NESS-causes directes]**.** *Étant donné un cadre causal* χ *et une formule vraie dans un état* $(\psi, t_\psi) \in \mathscr{F} \times \mathbb{T}$*, l'ensemble d'occurrences d'évènements* $C \subseteq \{(e, t) | e \in E^\chi(t), t \in \mathbb{T}\}$ *est un ensemble suffisant de NESS-causes directes de* (ψ, t_ψ)*, relation que nous notons* $C \rightarrow (\psi, t_\psi)$*, ssi il existe un support* $W \subseteq Lit_\mathbb{F}$ *tel que les conditions suivantes soient vérifiées :*

1. *Suffisance causale et minimalité de* W *:* $W \vDash \psi$ *et* $\forall W' \subset W,\ W' \nvDash \psi$ *;*
2. *Il existe une séquence décroissante* $t_1, \ldots, t_k$ *et une partition* $W_1, \ldots, W_k$ *de* W *qui lui est associée tel que* $\forall i \in \{1, \ldots, k\}$*, étant donné* $C(t_i) = C \cap E^\chi(t_i)$ *:*
   (a) *Nécessité faible et minimalité de* C *à* t_i *:* $S^\chi(t_i) \rhd C(t_i) \vDash W_i$ *et* $\forall C' \subset C(t_i),\ S^\chi(t_i) \rhd C' \nvDash W_i$ *;*
   (b) *Persistance de la nécessité :* $\forall t \in \mathbb{T}$ *tel que* $t_i < t \leq t_\psi,\ S^\chi(t) \vDash W_i$ *;*
3. *Minimalité de* C *:* $C = \bigcup_{i \in \{1,\ldots,k\}} C(t_i)$*.*

(e, t) *est une NESS-cause directe de* (ψ, t_ψ) *ssi* $\exists C \subseteq \mathbb{E} \times \mathbb{T}$ *tel que* $(e, t) \in C$ *et* $C \rightarrow (\psi, t_\psi)$*. Lorsque cela est pertinent, le support* W *peut être rendu explicite dans la notation* $C \underset{W}{\rightarrow} (\psi, t_\psi)$*.*

Le test NESS de Wright [1985] repose sur quatre éléments essentiels. Le défi derrière la représentation de ce test dans $\mathscr{S}_c$ repose en la représentation de ces éléments. Ces éléments sont : (i) la suffisance causale des causes, (ii) la nécessité faible des causes, (iii) la minimalité aussi bien dans la suffisance que dans la nécessité et (iv) l'effectivité des causes. Regardons les subtilités derrière chaque élément et la façon dont elles sont satisfaites dans la définition 6.10.

(i) La suffisance causale des causes est garantie principalement par la condition 1. Les ensembles suffisants pour causer (ψ, t_ψ) sont les supports W. L'ensemble d'occurrences d'évènements C qui a rendu vrais les littéraux de fluents dans W hérite de cette propriété et est donc également considéré comme suffisant, d'où leur appellation d'ensembles suffisants de NESS-causes directes. En passant des littéraux de fluents aux occurrences d'évènements, la simple suffisance logique dans $W \vDash \psi$ devient une suffisance causale. D'après Wright [2011], cette dernière intègre l'aspect temporel de la causalité faisant que cause et conséquence ne peuvent jamais être confondues : « The successional nature of causation is incorporated in the concept of causal sufficiency, which is defined as the complete instantiation of all the conditions in the antecedent of the relevant causal law ». Ce caractère successif est ici assuré par l'utilisation de $\mathscr{S}_c$ intégrant des informations causales et directionnelles, qui ont été introduites dans la section 6.1.3. Toute la surcouche sémantique apportée par $\mathscr{S}_c$ nous permet donc d'échapper aux critiques formulées à l'égard des théories par régularité que nous avons vues dans le chapitre 3. Plus précisément, nous échappons

aux critiques qui en réalité, plus que des critiques à la vision causale, sont des critiques à l'utilisation de la logique comme support. Nous avons vu dans le chapitre 3 que l'approche de WRIGHT [2011] était à mi-chemin entre approches par régularité et par inférence. Nous proposons ici une version entièrement par inférence.

Exemple 6.1 [suite]**.** *La formule* $tri(dev_o) = o_e \vee (o_m \wedge se_{hs})$ *a deux supports qui satisfont la condition 1 :* $\mathrm{W} = \{o_e\}$ *et* $\mathrm{W}' = \{o_m, se_{hs}\}$*. Un ensemble d'occurrences d'évènements peut être suffisant causalement à* $(tri(dev_o), 1)$ *aussi bien en étant un ensemble suffisant pour expliquer la véracité de* $(\mathrm{W}, 1)$*, que celle de* $(\mathrm{W}', 1)$*. Dans* $\mathscr{S}_c$ *il est impensable d'envisager que les rôles entre causes et conséquences puissent être inversés, ne serait-ce que par notre cadre imposant qu'une cause précède toujours sa conséquence.*

(ii) La nécessité faible des causes est garantie par la condition 2a. Nous sommes loin de la nécessité forte d'autres approches causales qui pose tant de problèmes pour les cas de surdétermination, où en l'absence de la cause la conséquence ne peut pas être vraie. Ici la nécessité est dite faible car elle est subordonnée à la suffisance ; la cause n'est pas nécessaire à la conséquence directement, mais au fait que tous les littéraux de fluents d'un des ensembles suffisants W soient vrais : « a causally relevant factor need merely be necessary for the sufficiency of a set of conditions sufficient for the occurrence of the consequence, rather than being necessary for the consequence itself » [WRIGHT, 2011].
Dans la condition 2a, la partie $\mathrm{S}^{\chi}(t_i) \rhd \mathrm{C}(t_i) \models \mathrm{W}_i$ indique que mettre à jour $\mathrm{S}^{\chi}(t_i)$ par l'ensemble $\mathrm{C}(t_i) \subseteq \mathrm{C}$ mène à un état $\mathrm{S}^{\chi}(t_i + 1)$ où l'élément W_i de la partition de W associée au temps t_i est satisfait. En appliquant cela $\forall i \in \{1, \ldots, k\}$, et donc à chaque temps de la séquence décroissante $t_1, \ldots, t_k$, nous nous assurons que tous les éléments de W sont bien rendus vrais.
Dans cette même condition, la partie $\forall \mathrm{C}' \subset \mathrm{C}(t_i)$, $\mathrm{S}^{\chi}(t_i) \rhd \mathrm{C}' \not\models \mathrm{W}_i$ est également importante. En effet, elle implique une condition essentielle pour parler de nécessité qui n'apparaît pas explicitement dans la condition : $\mathrm{S}^{\chi}(t_i) \not\models \mathrm{W}_i$. Il n'est pas possible de dire que $\mathrm{C}(t_i)$ est nécessaire à $\mathrm{S}^{\chi}(t_i + 1) \models \mathrm{W}_i$ si $\mathrm{S}^{\chi}(t_i) \models \mathrm{W}_i$ et donc que $\mathrm{S}^{\chi}(t_i) \rhd \varnothing \models \mathrm{W}_i$. Dans la mesure où l'ensemble $\mathrm{C}' = \varnothing$ est un sous ensemble de $\mathrm{C}(t_i)$, la condition essentielle pour parler de nécessité $(\mathrm{S}^{\chi}(t_i) \rhd \varnothing = \mathrm{S}^{\chi}(t_i)) \not\models \mathrm{W}_i$ est bien couverte par la deuxième partie de la condition 2a.

Exemple 6.1 [suite]**.** *Si nous prenons le support* $\mathrm{W}' = \{o_m, se_{hs}\}$ *et que nous regardons la trace d'états, il est possible de créer une partition de ce support en fonction du moment où les éléments dans* W′ *ont été rendus vrais. Plus précisément, nous associons à chaque littéral de fluent dans* W′ *le temps où la transition qui les a rendus vrais s'est produite. Nous avons donc une séquence décroissante* $t_1 = 0, t_2 = -1$ *et la partition associée* $\mathrm{W}'_1 = \{o_m\}, \mathrm{W}'_2 = \{se_{hs}\}$*. L'ensemble* $\mathrm{C}' = \{(prod_m, 0), (ini_{se_{hs}}, -1)\}$ *satisfait la condition 2a vu que :*

$$\left(\mathrm{S}^{\chi}(0) \rhd \{(prod_m, 0)\} \models o_m\right) \wedge \left(\mathrm{S}^{\chi}(0) \rhd \varnothing \not\models o_m\right)$$

et que

$$\left(\mathrm{S}^{\chi}(-1) \rhd \{(ini_{se_{hs}}, -1)\} \models se_{hs}\right) \wedge \left(\mathrm{S}^{\chi}(-1) \rhd \varnothing \not\models se_{hs}\right).$$

En appliquant le même raisonnement pour le support $\mathrm{W} = \{o_e\}$*, nous trouvons l'ensemble* $\mathrm{C} = \{(prod_e, 0)\}$.

(iii) La minimalité, aussi bien dans la suffisance que dans la nécessité, n'apparaît pas explicitement dans la formulation du test NESS. Cependant, différents travaux ont montré qu'elle était indispensable aux théories causales par régularité ou inférence [ANDREAS

et GUENTHER, 2021; BAUMGARTNER, 2013; WRIGHT, 2011]. Cet élément essentiel est garanti par plusieurs conditions. La minimalité du support est vérifiée par la condition 1, plus exactement la partie $\forall W' \subset W,\ W' \not\models \psi$. La minimalité des différents $C(t_i)$ est vérifiée par la partie $\forall C' \subset C(t_i),\ S^X(t_i) \rhd C' \not\models W_i$ de la condition 2a. Cette deuxième partie de la condition est donc d'autant plus importante qu'elle est indispensable pour la nécessité et pour la minimalité à la fois. La minimalité de l'ensemble C est vérifiée par la condition 3. Cette condition assure que la minimalité dans la nécessité appliquée à chaque t_i se retrouve dans l'ensemble final. Dit autrement, elle empêche qu'une occurrence n'ayant pas eu lieu dans un temps de la séquence $t_1, \ldots, t_k$ se retrouve dans C.

(iv) L'effectivité des causes est garantie par la façon dont les $C(t_i)$ sont construits dans la condition 2. En effet, le fait que $C(t_i) = C \cap E^X(t_i)$ assure que les causes appartiennent bien à la trace d'évènements, et donc soient des évènements qui se sont réellement produits, étant donné que $C(t_i) \subseteq E^X(t_i)$ et que par construction $E^X(t_i) = \tau^e_\chi(t_i)$.

Exemple 6.1 [suite]**.** *Par la minimalité liée à la suffisance, il n'est pas possible d'avoir comme support* $W'' = \{o_m, se_{hs}, t_n\}$ *même si* $W'' \models tri(dev_o)$ *puisque* $W'' \subset W'$ *et pourtant nous avons* $W' \models tri(dev_o)$.
Par la minimalité liée à la nécessité, il n'est pas possible d'avoir $C''(0) = \{(prod_m, 0), (prod_e, 0)\}$ *associé à* $W'_1 = \{o_m\}$*, même si* $S^X(0) \rhd C''(0) \models o_m$*, puisque* $\{(prod_m, 0)\} \subset C''(0)$ *et pourtant* $S^X(0) \rhd \{(prod_m, 0)\} \models o_m$.
Par la minimalité liée à C*, il n'est pas possible d'avoir* $C'' = \{(prod_m, 0), (ini_{se_{hs}}, -1), (dev_o, 1)\}$ *étant donné que* $\nexists i \in \{1, \ldots, k\}$ *tel que* $t_i = 1$ *et donc il ne peut y avoir une occurrence à ce temps dans* C *puisque* $C = \bigcup_{i \in \{1,\ldots,k\}} C(t_i)$*,* $(dev_o, 1)$ *n'y a rien à faire.*
Par l'effectivité des causes, il est inenvisageable d'avoir $C'' = \{(prod_m, 2), (ini_{se_{hs}}, -1)\}$ *étant donné que* $prod_m \notin E^X(2)$.

Si la condition 2b n'a été mentionnée dans aucun des paragraphes précédents c'est parce que cette condition n'est pas propre au test NESS comme initialement conçu. L'absence de cette condition purement temporelle s'explique par le simple fait que le test NESS n'a pas été initialement pensé pour un formalisme représentant le temps explicitement. Il est toutefois clair dans les exemples utilisés qu'il était toujours présent dans le raisonnement de WRIGHT [2011]. Dans notre représentation où le temps est pris en compte, cette condition garantit qu'à l'ensemble d'évènements $C(t_i)$ auquel est attribué la satisfaction de W_i à $t_i + 1$, peut également être attribué la satisfaction de W_i à t_ψ. En d'autres termes, elle s'assure qu'un des éléments de W_i n'est pas devenu faux entre t_i et t_ψ, ce qui ferait qu'un ensemble d'évènements autre que $C(t_i)$ devrait se voir attribuer la satisfaction de W_i à t_ψ. Il suffit pour cela de vérifier que W_i a été satisfait par tous les états entre $t_i + 1$ et t_ψ.

Exemple 6.1 [suite]**.** *Dans le cas du support* $W' = \{o_m, se_{hs}\}$ *pour lequel nous avons trouvé l'ensemble* $C' = \{(prod_m, 0), (ini_{se_{hs}}, -1)\}$*, deux cas de figure possibles apparaissent. Le premier est associé à l'occurrence* $(prod_m, 0)$*. La condition 2b y est immédiatement satisfaite étant donné que* $t_i + 1 = t_\psi$*. Le deuxième est associé à l'occurrence* $(ini_{se_{hs}}, -1)$*. La condition 2b n'est pas immédiatement satisfaite étant donné que* $t_i + 1 \neq t_\psi$*. Dans ce cas la valeur de* se_{hs} *pourrait changer entre* t_i *et* t_ψ*. Si initialement la station d'épuration est hors service en raison d'une maintenance, mais qu'après sa remise en service un agent fait une action de sabotage qui la remet hors service, dans le cas où le déversement des eaux usées industrielles se produit après le sabotage, il serait erroné de dire que la maintenance a contribué au préjudice final. Notez tout de même que dans ce cas précis cela n'est pas possible quel que soit le STEE*

utilisé, car entre $t_i = -1$ et $t_\psi = 1$ il n'y a pas assez de pas temporels pour que se_{hs} devienne faux, puis vrai à nouveau. Cela n'est possible qu'à partir du moment où $t_\psi - t_i > 2$.

La définition 6.10 ne définit pas une procédure pour trouver les ensembles suffisants de NESS-causes directes, elle définit lorsqu'un ensemble d'occurrences d'évènements peut être considéré comme tel. Dans les faits, cette procédure est simple. Nous la présentons ci-dessous et nous l'illustrons à l'aide de l'exemple 6.2.

Exemple 6.2 [peloton d'exécution]**.** *Nous reprenons l'exemple 5.1 du peloton d'exécution auquel nous ajoutons un interrupteur. Faisons un bref rappel. La figure 6.2a illustre un circuit électrique pouvant correspondre à un peloton d'exécution où le condamné est soumis à la chaise électrique au lieu d'être fusillé. Le circuit est composé d'une batterie, d'un individu attaché et connecté à des électrodes, et de trois interrupteurs montés en parallèle. Nous supposons que cinq agents sont impliqués dans la situation : celui attaché et quatre autres, chacun pouvant contrôler les composants du circuit mais ignorant ce que font les autres.*
Les littéraux de fluents $l_1, l_2, l_3, l_4 \in Lit^4_{\mathbb{F}}$ correspondent chacun à l'état fermé d'un des éléments du circuit. Ainsi, lorsque le littéral est vrai, cela indique que le courant peut passer pour les interrupteurs et que du courant est généré pour la batterie.
La condition de déclenchement pour que l'individu connecté aux électrodes soit électrocuté, évènement noté $e_\psi \in \mathbb{N}$, est $\psi = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$, où $\psi \in \mathscr{F}$. Si cet évènement se produit, l'individu est considéré comme mort $d \in Lit_{\mathbb{F}}$. Les actions $e_1, e_2, e_3 \in \mathbb{A}^3$ ont comme effet intrinsèque de mettre un des interrupteurs dans son état fermé. L'action $e_4 \in \mathbb{A}$ fait de même pour la batterie. L'action $e_{\neg 1} \in \mathbb{A}$ a comme effet intrinsèque de mettre un des interrupteurs dans son état ouvert. Toute l'information donnée jusqu'ici fait partie du contexte κ_c de l'exemple. Elle peut se formaliser ainsi :

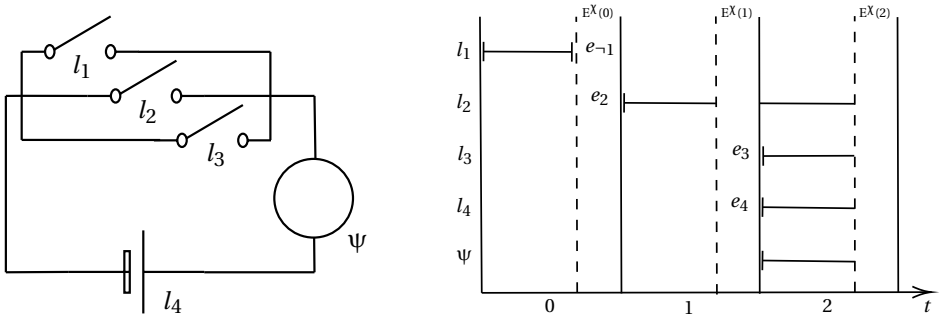
$$\forall i \in \{1,2,3,4\},\ pre(e_i) = \top,\ eff^+(e_i) = \{l_i\};$$
$$pre(e_{\neg 1}) = \top,\ eff^+(e_1) = \left\{\overline{l_1}\right\};$$
$$tri(e_\psi) = \psi,\ eff^+(e_\psi) = \{d\}.$$

Dans cet exemple nous considérons le scénario $\sigma = \{(e_{\neg 1}, 0), (e_2, 0), (e_3, 1), (e_4, 1)\}$. Ci-dessous les traces τ^e_χ et τ^s_χ obtenues étant donné le cadre causal χ décrit, elles correspondent à un cas de surdétermination et sont illustrées sur la figure 6.2b.

$$S^\chi(-1) = \left\{\overline{l_1}, l_2, l_3, l_4, d\right\}, E^\chi(-1) = \left\{ini_{l_1}, ini_{\overline{l_2}}, ini_{\overline{l_3}}, ini_{\overline{l_4}}, ini_{\overline{d}}\right\};$$
$$S^\chi(0) = \left\{l_1, \overline{l_2}, \overline{l_3}, \overline{l_4}, \overline{d}\right\}, E^\chi(0) = \{e_{\neg 1}, e_2\};$$
$$S^\chi(1) = \left\{\overline{l_1}, l_2, \overline{l_3}, \overline{l_4}, \overline{d}\right\}, E^\chi(1) = \{e_3, e_4\};$$
$$S^\chi(2) = \left\{\overline{l_1}, l_2, l_3, l_4, \overline{d}\right\}, E^\chi(2) = \left\{e_\psi\right\};$$
$$S^\chi(3) = \left\{\overline{l_1}, l_2, l_3, l_4, d\right\}.$$

En premier lieu, il faut commencer par identifier les supports W de la formule ψ pour lesquels les causes sont recherchées. Comme mentionné dans la section 5.2.2, chaque impliquant premier de ψ est un support. Il ne faut garder pour la suite que ceux où tous les littéraux de fluents les composant sont vrais au temps t_ψ.

Exemple 6.2 [suite]**.** *La question qui se pose dans l'exemple 6.2 est : quelles sont les causes de l'électrocution de l'individu au temps $t = 2$? Formellement, nous cherchons donc la cause de*



(a) Circuit électrique composé d'une batterie, d'un individu attaché et connecté à des électrodes, et de trois interrupteurs montés en parallèle.

(b) Illustration des traces d'évènements et d'états pour le cadre causal χ décrit dans l'exemple.

FIGURE 6.2 – Illustration de l'exemple 6.2 mettant en scène un circuit électrique pouvant correspondre à un peloton d'exécution où le condamné est soumis à la chaise électrique au lieu d'être fusillé.

la formule vraie $(\psi, 2)$.
Il existe trois supports possibles pour ψ *:* $W = \{l_1, l_4\}$, $W' = \{l_2, l_4\}$ *et* $W'' = \{l_3, l_4\}$. *Étant donné les traces* τ^e_χ *et* τ^s_χ, *nous avons* $(\psi, 2)$ *grâce aux supports* W' *et* W''. *Ce sont ceux que nous garderons pour la suite.* W *ne l'est pas dans la mesure où* $S^\chi(t_\psi) \not\models l_1$.

En second lieu, il faut établir une partition $W_1, \ldots, W_k$ de chaque W avec sa séquence décroissante $t_1, \ldots, t_k$. Pour cela, il est pertinent de commencer par t_ψ et de remonter dans le temps jusqu'à $t = -1$ en cherchant le moment où chaque littéral de fluent dans W a changé de valeur. Le temps à garder est le premier temps rencontré où le littéral est faux. Le fait de remonter le temps comme indiqué permet de s'assurer que la condition 2b est satisfaite dès cette étape. Une fois chaque littéral de fluent dans W associé à un temps, les regrouper en fonction du temps qui leur est associé revient à former la partition $W_1, \ldots, W_k$ et la séquence décroissante $t_1, \ldots, t_k$ voulue.

Exemple 6.2 [suite]**.** *Appliquons cette étape du processus à* $W' = \{l_2, l_4\}$. *Nous avons* l_4 *qui devient vrai au temps* $t = 2$, *au temps* $t = 1$ *il était faux. Puis, nous avons* l_2 *qui devient vrai au temps* $t = 1$, *au temps* $t = 0$ *il était faux. Nous avons donc la partition* $W'_1 = \{l_4\}, W'_2 = \{l_2\}$ *associée à la séquence décroissante* $t'_1 = 1, t'_2 = 0$.
Nous faisons de même pour $W'' = \{l_3, l_4\}$. *Nous avons* l_4 *qui devient vrai au temps* $t = 2$, *au temps* $t = 1$ *il était faux. Puis, nous avons* l_3 *qui devient également vrai au temps* $t = 2$, *au temps* $t = 1$ *il était faux. Nous avons donc la partition* $W''_1 = \{l_4, l_3\}$ *associée à la séquence décroissante* $t''_1 = 1$.

En dernier lieu, il faut s'intéresser à chaque élément de la partition et regarder quel ensemble d'occurrences d'évènements au temps qui lui est associé satisfait la condition 2a. En faisant l'union des ensembles trouvés pour chaque élément de la partition $W_1, \ldots, W_k$, ce qui correspond à la condition 3, un ensemble suffisant de NESS-causes directes de (ψ, t_ψ) est obtenu. Les éléments de cet ensemble sont chacun une NESS-cause directe de (ψ, t_ψ).

Exemple 6.2 [suite]**.** *Appliquons cette étape du processus à la partition* $W'_1 = \{l_4\}, W'_2 = \{l_2\}$ *associée à la séquence décroissante* $t'_1 = 1, t'_2 = 0$. *Commençons par* $W'_1 = \{l_4\}$. *Grâce à la trace*

et au contexte, donc uniquement aux informations données par $\mathscr{S}_c$, nous trouvons l'ensemble $C'(1) = \{(e_4, 1)\}$ *qui vérifie* $S^{\chi}(1) \triangleright C'(1) \models W'_1$ *et* $\forall C \subset C'(1),\ S^{\chi}(1) \triangleright C \not\models W'_1$. *En suivant le même raisonnement, nous trouvons pour* $W'_2 = \{l_2\}$ *l'ensemble d'occurrences* $C'(0) = \{(e_2, 0)\}$. *Nous pouvons alors former un ensemble suffisant de NESS-causes directes de* (ψ, t_ψ). *Cet ensemble est* $C' = C'(1) \cup C'(0) = \{(e_2, 0), (e_4, 1)\}$.
Nous faisons de même pour la partition $W''_1 = \{l_4, l_3\}$ *associée à la séquence décroissante* $t''_1 = 1$. *Avec le même raisonnement nous trouvons* $C''(1) = \{(e_3, 1), (e_4, 1)\}$ *qui vérifie* $S^{\chi}(1) \triangleright C''(1) \models W''_1$ *et* $\forall C \subset C''(1),\ S^{\chi}(1) \triangleright C \not\models W''_1$. *La partition étant composée d'un unique élément, l'ensemble suffisant de NESS-causes directes de* (ψ, t_ψ) *est pour ce support* $C'' = C''(1) = \{(e_3, 1), (e_4, 1)\}$.

Profitons de cet exemple pour comparer sur le même problème notre approche causale à d'autres existantes qui utilisent des langages de représentation de l'action et du changement. Nous choisissons de nous comparer à celle de BATUSOV et SOUTCHANSKI [2018] qui semble être la plus aboutie pour le moment et qui s'inspire de la définition de causalité d'HALPERN [2016] en Calcul des Situations. Pour cette comparaison nous devons toutefois ignorer l'évènement e_3 étant donné que l'approche de BATUSOV et SOUTCHANSKI [2018] ne gère pas la cooccurrence d'évènements. La comparaison devra donc uniquement être faite avec l'ensemble suffisant de NESS-causes directes $C' = \{(e_2, 0), (e_4, 1)\}$.
À la question : quelles sont les causes de l'électrocution de l'individu au temps $t = 2$, donc de la formule vraie $(\psi, 2)$? BATUSOV et SOUTCHANSKI [2018] répondraient que $(ini_{l_1}, -1)$ et $(e_4, 1)$ sont des « achievement causes » et $(e_2, 0)$ une « maintenance cause ». Le premier type de causes est celui pour lequel ils revendiquent une proximité avec HALPERN [2016]. Le deuxième est un ajout des auteurs qui qualifient ces causes de « causes responsible for protecting a previously achieved effect, despite potential threats that could destroy the effect ». Pour notre besoin de factualité, la présence de $(ini_{l_1}, -1)$ dans les causes semble inacceptable. Concrètement, $(ini_{l_1}, -1)$ ne joue aucun rôle sur la véracité de ψ au temps $t = 2$, puisque l_1 n'est plus vrai à ce temps là. De ce fait, $(ini_{l_1}, -1)$ ne peut pas être considéré comme une cause.
Cependant, derrière ce raisonnement il y a l'intuition de la nécessité forte : une cause est quelque chose qui a changé le résultat final et donc doit être soumis à un raisonnement contrefactuel. $(ini_{l_1}, -1)$ est ce qui a fait qu'au moins un des interrupteurs soit fermé, après cela les autres n'ont fait que maintenir cette condition vraie. Mettons nous dans le cas où l_1 n'est pas vrai dans l'état initial. Un raisonnement similaire nous mènerait à penser qu'entre $C' = \{(e_2, 0), (e_4, 1)\}$ et $C'' = \{(e_3, 1), (e_4, 1)\}$ le premier ensemble est « plus cause » que le second sous le prétexte que, bien que l_2 et l_3 sont tous les deux vrais à $t = 2$, l'antériorité de l_2 doit être prise en compte. Nous devons nous opposer à ce raisonnement qui revient à confondre causalité et responsabilité, nous cherchons les causes et non pas une sorte de « cause responsable ». Comme nous l'avons vu dans la section 3.1.2.2, s'aventurer dans le terrain de la responsabilité revient à quitter le cadre factuel dans lequel nous devons nous inscrire. En prenant en compte l'antériorité nous ne serions plus dans la « causal inquiry » de WRIGHT [1985], mais dans la « proximate-cause inquiry ». En effet, une fois les causes $(e_2, 0), (e_3, 1), (e_4, 1)$ déterminées, si nous voulions déterminer la responsabilité pénale des individus derrière chaque action, nous procéderions à une étape subjective où nous évaluerions si l'antériorité de $(e_2, 0)$ atténue ou élimine la responsabilité de $(e_3, 1)$ sur le préjudice. Nous soupçonnons donc que les choix de BATUSOV et SOUTCHANSKI [2018] les menant à considérer $(ini_{l_1}, -1)$ comme une cause sont le résultat de ce type de raisonnement, mais poussé encore plus loin.

Jusqu'ici, nous avons parlé de cas où pour chaque support W il n'y avait qu'un ensemble suffisant de NESS-causes directes. Comme nous avons pu le voir dans le chapitre 5, cela permet déjà l'existence de cas de surdétermination. Notez cependant qu'en autorisant la cooccurrence d'évènements, il peut exister plusieurs ensembles suffisants de NESS-causes directes pour un même support W.

Proposition 6.2 [Non unicité des ensembles suffisants de NESS-causes directes]**.** *Étant donné un cadre causal* χ *et la relation causale* $C \underset{W}{\rightarrow} (\psi, t_\psi)$, C *n'est pas nécessairement l'unique ensemble suffisant de NESS-causes directes de* (ψ, t_ψ) *d'après la définition 6.10.*

Démonstration. Soit $l \in Lit_{\mathbb{F}}$, la formule $\psi = l$, les ensembles d'occurrences d'évènements $C = \{(e,0)\}$ et $C' = \{(e',0)\}$ où $pre(e) = pre(e') = \top$, $eff(e) = eff(e') = l$ et les traces d'évènements et d'états τ^e_χ et τ^s_χ :

$S^\chi(-1) = \{l\}, E^\chi(-1) = \{ini_{\overline{l}}\}$;

$S^\chi(0) = \{\overline{l}\}, E^\chi(0) = \{e, e'\}$;

$S^\chi(1) = \{l\}$.

Considérons l'ensemble $W = \{l\}$ qui satisfait la condition 1 de la définition 6.10 et donc peut être considéré comme un support de la formule vraie $(l, 1)$. Il est possible d'y associer la partition $W_1 = \{l\}$ et la séquence $t_1 = 0$.
D'après la définition 6.10, $C \underset{W}{\rightarrow} (\psi, t_\psi)$, puisque $\forall i \in \{1\}, C(0) = C \cap E^\chi(0) = \{(e,0)\}$:

— $S^\chi(0) \rhd C(0) \models \{l\}$;

— $\forall C'' \subset C(0)$, $S^\chi(0) \rhd C'' \not\models \{l\}$;

— $\forall t$, $0 < t \leq 1$, $S^\chi(t) \models \{l\}$;

— $C = \bigcup_{i \in \{1\}} C(t_i)$.

Ces mêmes conditions sont satisfaites par l'ensemble C', donc d'après la définition 6.10, $C' \underset{W}{\rightarrow} (\psi, t_\psi)$. Nous venons donc de prouver la non-unicité des ensembles de NESS-causes directes, même lorsque le support W est le même. Dans ce cas particulier la non unicité est possible par la cooccurrence d'évènements. □

La définition 6.10 définit les conditions pour qu'un ensemble d'occurrences d'évènements puisse être considéré comme un ensemble suffisant de NESS-causes directes de (ψ, t_ψ). Pour savoir si une occurrence d'évènement est une NESS-cause directe de (ψ, t_ψ), il faut alors regarder si cette occurrence appartient à un des ensembles suffisants de NESS-causes directes de (ψ, t_ψ). Disposer des ensembles suffisants peut s'avérer intéressant d'un point de vue causal. Cependant, dans les réflexions sur la surdétermination du chapitre 5, ainsi que dans les besoins propres à l'éthique computationnelle, il est plus courant de parler des causes individuellement. Il est possible à partir de la définition 6.10 de déterminer des conditions permettant de savoir si une occurrence d'évènement peut être considérée comme une NESS-cause directe en fonction de la forme de la formule $\psi \in \mathscr{F}$, et cela sans devoir passer par les ensembles suffisants.
Le premier cas est celui où la conséquence qui nous intéresse est un littéral de fluent, i.e. $\psi = l \in Lit_{\mathbb{F}}$. Dans ce cas, la NESS-cause directe sera la dernière occurrence d'évènement à avoir rendu vrai l avant ou à t_ψ. Dans ce cas basique, le support W est le singleton dont l'unique élément est le littéral l.

Proposition 6.3 [NESS-cause directe d'un littéral]**.** *Étant donné un cadre causal* χ*, une occurrence d'évènement* (e,t) *et une formule vraie* (l,t_l) *où* $l \in Lit_{\mathbb{F}}$*,* (e,t) *est une NESS-cause directe de* (l,t_l) *ssi :*

$c_1)$ $S^\chi(t) \rhd \{e\} \models l$;

$c_2)$ $\forall t' \in \mathbb{T}, t < t' \leq t_l,\ S^\chi(t') \models l$;

$c_3)$ $S^\chi(t) \not\models l$.

Démonstration. [$\Longrightarrow$] Soit (e,t) une NESS-cause directe de (l,t_l). Par la définition 6.10, $\exists C \subseteq \mathbb{E} \times \mathbb{T}$, tel que $C \underset{W}{\rightarrow} (l,t_l)$ et $(e,t) \in C$. Le support W est nécessairement l'état partiel $L = \{l\}$ étant donné que L est l'unique ensemble qui satisfait la condition 1 de la définition 6.10. En effet, $L \models l$ et son seul sous ensemble, $\varnothing$, ne satisfait pas l.

D'après la définition 6.9, le singleton W a exactement une partition $W_1 = \{l\}$ à laquelle peut être associée le temps t_1. Par la définition 6.10, $C = C(t_1)$ et comme $(e,t) \in C$, $t_1 = t$.

L'unique partition étant $W_1 = \{l\}$, la condition 2b de la définition 6.10 peut s'exprimer $\forall t' \in \mathbb{T},\ t < t' \leq t_l,\ S^\chi(t') \models l$. Cela vérifie (c_2). La condition 2a de la définition 6.10 peut s'exprimer $S^\chi(t) \rhd C(t) \models l$ et $\forall C' \subset C(t),\ S^\chi(t) \rhd C' \not\models l$. La condition $S^\chi(t) \not\models l$ correspondant à (c_3) est inclue dans la condition 2a puisque dans le cas où $C' = \varnothing$, cela revient à $S^\chi(t) \rhd \varnothing \not\models l$ avec $S^\chi(t) = S^\chi(t) \rhd \varnothing$ par la définition 6.4. Comme $(e,t) \in C = C(t)$, $C(t) = \{(e,t)\} \cup C''$, où $C'' = \{(e'',t) | e'' \in E^\chi(t)\}$. De ce fait, la première partie de la condition 2a de la définition 6.10 peut s'écrire $S^\chi(t) \rhd (\{(e,t)\} \cup C'') \models l$, ce qui par les définitions 6.3 et 6.4 signifie :

$$l \in \left(S^\chi(t) \setminus \left(\overline{actualEff(\{(e,t)\}, S^\chi(t))} \cup \overline{actualEff(C'', S^\chi(t))}\right)\right) \cup actualEff(\{(e,t)\}, S^\chi(t)) \cup actualEff(C'', S^\chi(t)).$$

Trois cas peuvent être considérés. Soit l appartient à :
(i) $S^\chi(t) \setminus \left(\overline{actualEff(\{(e,t)\}, S^\chi(t))} \cup \overline{actualEff(C'', S^\chi(t))}\right)$;
(ii) $actualEff(C'', S^\chi(t))$;
ou (iii) $actualEff(\{(e,t)\}, S^\chi(t))$.
Le premier cas implique $l \in S^\chi(t)$, ce qui est en contradiction avec la condition $S^\chi(t) \not\models l$. Le deuxième cas implique $S^\chi(t) \rhd C'' \models l$, ce qui est en contradiction avec la condition 2a de la définition 6.10, $\forall C' \subset C(t),\ S^\chi(t) \rhd C' \not\models l$ étant donné que $C'' \subset C(t)$. Le troisième cas est donc le seul cas possible, ce qui implique $l \in actualEff(\{(e,t)\}, S^\chi(t))$, qui peut être écrit $S^\chi(t) \rhd \{e\} \models l$, correspondant à (c_1). Nous venons donc de prouver que lorsque (e,t) est une NESS-cause directe de (l,t_l), nous avons (c_1), (c_2) et (c_3).

[$\Longleftarrow$] Soit un littéral de fluent l vrai au temps t_l, (e,t) une occurrence d'évènement qui vérifie (c_1), (c_2) et (c_3), et un ensemble d'occurrences d'évènements $C = \{(e,t)\}$. Considérons un support hypothétique $W = \{l\}$ de la formule vraie (l,t_l). Ce support satisfait la condition 1 de la définition 6.10 puisque $\{l\} \models l$ et son seul sous ensemble, $\varnothing$, ne satisfait pas l.

Il existe une séquence décroissante $t_1 = t$ et une partition $W_1 = \{l\}$ de W telles que $\forall i \in \{1\}$, étant donné $C(t_i) = C \cap E^\chi(t_i)$:

— puisque $C = \{(e,t_i)\}$, $C = \bigcup_{i \in \{1,\ldots,k\}} C(t_i)$, ce qui correspond à la condition 3 de la définition 6.10 ;

— (c_1) peut s'écrire $S^\chi(t_i) \rhd C(t_i) \models W_i$ et (c_3) peut s'écrire $S^\chi(t_i) \not\models W_i$, ensemble cela correspond à la condition 2a de la définition 6.10 puisque $S^\chi(t_i) \rhd \varnothing \not\models W_i$ avec $\varnothing$ étant le seul sous-ensemble de $\{e\}$;

— (c_2) peut s'écrire $\forall t',\ t_i < t' \leq t_l,\ S^\chi(t') \models W_i$, ce qui correspond à la condition 2b de la définition 6.10.

Par conséquent, $\exists C \subseteq \mathbb{E} \times \mathbb{T}$ tel que $(e,t) \in C$ et $C \rightarrow (l, t_l)$. Nous venons donc de prouver que lorsque nous avons (c_1), (c_2) et (c_3), (e,t) est une NESS-cause directe de (l, t_l). □

Le deuxième cas est celui où la conséquence qui nous intéresse est une conjonction de littéraux de fluents $\psi = l_1 \wedge \cdots \wedge l_m$. Dans ce cas, toutes les NESS-causes directes de (l_j, t_ψ), avec l_j étant n'importe lequel des littéraux dans la conjonction, sont des NESS-causes directes de (ψ, t_ψ).

Proposition 6.4 [NESS-cause directe d'une conjonction]**.** *Étant donné un cadre causal* χ*, une occurrence d'évènement* (e,t) *et une formule vraie* (ψ, t_ψ) *où* $\psi = l_1 \wedge \cdots \wedge l_m$*,* (e,t) *est une NESS-cause directe de* (ψ, t_ψ) *ssi :*

c_1) $\exists j \in \{1,\ldots,m\},\ S^\chi(t) \rhd \{e\} \models l_j$;

c_2) $\forall t' \in \mathbb{T},\ t < t' \leq t_\psi,\ S^\chi(t') \models l_j$;

c_3) $S^\chi(t) \not\models l_j$.

Démonstration. [$\Longrightarrow$] Soit (e,t) une NESS-cause directe de (ψ, t_ψ). Par la définition 6.10, $\exists C \subseteq \mathbb{E} \times \mathbb{T}$, tel que $C \underset{W}{\rightarrow} (\psi, t_\psi)$ et $(e,t) \in C$. Le support W est nécessairement l'état partiel $L = \{l_1,\ldots,l_m\}$ étant donné que L est l'unique ensemble qui satisfait la condition 1 de la définition 6.10. En effet, $L \models \psi$ et $\forall L',\ L' \subset L, L' \not\models \psi$ étant donné que $\forall L',\ L' \subset L,\ \exists l_j \in \psi,\ l_j \notin L'$.

Pour chaque l_j, soit t_{l_j} le dernier pas de temps où l_j n'était pas vrai. Formellement, il s'agit de l'unique temps qui vérifie $S^\chi(t_{l_j}) \not\models l_j$ et $\forall t' \in \mathbb{T},\ t_{l_j} < t' \leq t_\psi,\ S^\chi(t') \models l_j$, temps qui existe nécessairement étant donné que $S^\chi(t_\psi) \models \psi \implies S^\chi(t_\psi) \models l_j$ et que $S^\chi(-1) = Lit_\mathbb{F} \setminus S^\chi(0)$. Donc soit l_j était vrai à l'état initial et $t_{l_j} = -1$, soit il a été rendu vrai après et $-1 < t_{l_j} < t_\psi$. Soit $T = \left\{ t_{l_j} \,|\, j \in \{1,\ldots,m\} \right\}$ et $t_1,\ldots,t_k$ la séquence décroissante correspondant à T, où $k \leq m$. La partition du support W peut alors être définie comme $\forall i \in \{1,\ldots,k\}, W_i = \left\{ l_j \in W \,|\, t_{l_j} = t_i \right\}$.

Comme $(e,t) \in C$ et que par la définition 6.10, $C(t_i) = C \cap E^\chi(t_i)$, alors $\exists i \in \{1,\ldots,k\},\ t_i = t$ et $S^\chi(t) \rhd C(t) \models W_i$ avec $C(t) = \{(e,t)\} \cup C''$, où $C'' = \left\{ (e'',t) | e'' \in E^\chi(t) \right\}$. De ce fait, la condition 2a de la définition 6.10 peut s'écrire $\forall l_j \in W_i,\ S^\chi(t) \rhd (\{(e,t)\} \cup C'') \models l_j$, ce qui par la définition 6.4 veut dire que $\forall l_j \in W_i$:

$$l \in \left(S^\chi(t) \setminus \left(\overline{actualEff(\{(e,t)\}, S^\chi(t))} \cup \overline{actualEff(C'', S^\chi(t))} \right) \right) \cup actualEff(\{(e,t)\}, S^\chi(t)) \cup actualEff(C'', S^\chi(t)).$$

Comme précédemment, trois cas peuvent être considérés. Soit l appartient à :
(i) $S^\chi(t) \setminus \left(\overline{actualEff(\{(e,t)\}, S^\chi(t))} \cup \overline{actualEff(C'', S^\chi(t))} \right)$;
(ii) $actualEff(C'', S^\chi(t))$;
ou (iii) $actualEff(\{(e,t)\}, S^\chi(t))$.
Le premier cas implique $l_j \in S^\chi(t)$, ce qui est en contradiction avec la condition $S^\chi(t) \not\models l_j$ inhérente par la façon dont est construite la partition du support, ce qui correspond à (c_3). l_j appartient donc nécessairement à $actualEff(\{(e,t)\}, S^\chi(t)) \cup actualEff(C'', S^\chi(t))$. Le cas où $\forall l_j \in W_i, l_j \notin actualEff(\{(e,t)\}, S^\chi(t))$ est contradictoire avec la condition 2a de la définition 6.10, $\forall C' \subset C(t),\ S^\chi(t) \rhd C' \not\models W_i$. En effet, cela dit $\forall l_j \in W_i, l_j \in actualEff(C'', S^\chi(t))$ ce qui est en contradiction avec cette condition puisque par la définition 6.10 nous savons

que $\{(e,t)\} \subset \mathrm{C}(t)$. Le cas où $\exists j \in \{1,\dots,m\},\ l_j \in actualEff(\{(e,t)\}, \mathrm{S}^\chi(t))$ est le seul cas possible. Par la définition 6.4, ce résultat peut s'écrire $\exists j \in \{1,\dots,m\},\ \mathrm{S}^\chi(t) \rhd \{e\} \vDash l_j$, ce qui correspond à (c_1). Finalement, pour le même $i \in \{1,\dots,k\}$ tel que $t_i = t$, la condition 2b de la définition 6.10 peut s'écrire comme $\forall l_j \in \mathrm{W}_i, \forall t' \in \mathbb{T},\ t < t' \leq t_\psi,\ \mathrm{S}^\chi(t') \vDash l_j$, ce qui correspond à (c_2). Nous venons donc de prouver que lorsque (e,t) est une NESS-cause directe de (ψ, t_ψ), nous avons (c_1), (c_2) et (c_3).

[$\Longleftarrow$] Soit une conjonction de littéraux de fluents ψ vrai au temps t_ψ et (e,t) une occurrence d'évènement qui vérifie (c_1), (c_2) et (c_3). Considérons un support hypothétique $\mathrm{W} = \{l_1,\dots,l_m\}$ de la formule vraie (ψ, t_ψ). Ce support satisfait la condition 1 de la définition 6.10 puisque $\{l_1,\dots,l_m\} \vDash \psi$ et que $\forall \mathrm{W}'$ tel que $\mathrm{W}' \subset \mathrm{W}, \mathrm{W}' \nvDash \psi$ étant donné que nous avons $\forall \mathrm{W}',\ \mathrm{W}' \subset \mathrm{W}, \exists l_j \in \psi,\ l_j \notin \mathrm{W}'$.

Comme $\mathrm{S}^\chi(t_\psi) \vDash \psi$ et $\mathrm{S}^\chi(-1) = Lit_\mathbb{F} \setminus \mathrm{S}^\chi(0)$, $\exists \mathrm{C} \subseteq \mathbb{E} \times \mathbb{T}$ tel que $\mathrm{C} \underset{\mathrm{W}}{\rightarrow} (\psi, t_\psi)$. Étant donné que (c_1) $\exists j \in \{1,\dots,m\},\ \mathrm{S}^\chi(t) \rhd \{e\} \vDash l_j$, (c_2) $\forall t' \in \mathbb{T},\ t < t' \leq t_\psi,\ \mathrm{S}^\chi(t') \vDash l_j$ et (c_3) $\mathrm{S}^\chi(t) \nvDash l_j$, d'après la proposition 6.3, (e,t) est une NESS-cause directe de (l_j, t_ψ). De ce fait, par la définition 6.10, $\exists \mathrm{C}' \subseteq \mathbb{E} \times \mathbb{T}$ tel que $(e,t) \in \mathrm{C}'$ et $\mathrm{C}' \underset{\mathrm{W}_j}{\rightarrow} (l_j, t_\psi)$. Comme montré dans la preuve de la proposition 6.3, le support de cette relation causale est nécessairement $\mathrm{W}_j = \{l_j\}$. Puisque $l_j \in \psi$ et que la partition de W satisfait par définition $\bigcup_{i \in \{1,\dots,k\}} \mathrm{W}_i = \mathrm{W}$, alors nous savons $\exists i \in \{1,\dots,k\}, \mathrm{W}_j \subseteq \mathrm{W}_i$ et $t_i = t$. Par conséquent, $\mathrm{C}' \subseteq \mathrm{C}(t)$ et donc $(e,t) \in \mathrm{C}$. Nous venons donc de prouver que lorsque nous avons (c_1), (c_2) et (c_3), (e,t) est une NESS-cause directe de (ψ, t_ψ). □

Le dernier cas est celui où la conséquence qui nous intéresse est sous forme normale disjonctive, i.e. une disjonction d'une ou plusieurs conjonctions d'un ou plusieurs littéraux de fluents $\psi = \psi_1 \vee \dots \vee \psi_m$. Ce cas est de loin le plus intéressant, derrière la grande majorité des cas de surdétermination il y a un ψ de cette forme. De fait, du moment où ψ est sous cette forme il existe un support W pour chaque élément disjoint, il existe donc autant de formes possibles de causer (ψ, t_ψ).

Proposition 6.5 [NESS-cause directe d'une FND]**.** *Étant donné un cadre causal* χ*, une occurrence d'évènement* (e,t) *et une formule vraie* (ψ, t_ψ) *où* $\psi = \psi_1 \vee \dots \vee \psi_m$ *est minimale par subsomption*[1] *et libre de tautologies,* (e,t) *est une NESS-cause directe de* (ψ, t_ψ) *ssi :*

- c_1) $\exists j \in \{1,\dots,m\},\ \mathrm{S}^\chi(t_\psi) \vDash \psi_j$;
- c_2) $\mathrm{S}^\chi(t) \rhd \{e\} \vDash l \in \psi_j$;
- c_3) $\forall t' \in \mathbb{T},\ t < t' \leq t_\psi,\ \mathrm{S}^\chi(t') \vDash l$;
- c_4) $\mathrm{S}^\chi(t) \nvDash l$.

Démonstration. [$\Longrightarrow$] Soit (e,t) une NESS-cause directe de (ψ, t_ψ). Par la définition 6.10, $\exists \mathrm{C} \subseteq \mathbb{E} \times \mathbb{T}$, tel que $\mathrm{C} \underset{\mathrm{W}}{\rightarrow} (\psi, t_\psi)$ et $(e,t) \in \mathrm{C}$. Étant donné que ψ est sous forme normale disjonctive, est minimale par subsomption et est libre de tautologies, les supports possibles $\mathrm{W}_1,\dots,\mathrm{W}_m$ sont nécessairement les états partiels $\mathrm{L}_1,\dots,\mathrm{L}_m$ qui correspondent à chaque conjonction de littéraux $\psi_1,\dots,\psi_m$, comme dans la preuve de la proposition 6.4. La relation causale dont nous connaissons l'existence par nos hypothèses, $\mathrm{C} \underset{\mathrm{W}}{\rightarrow} (\psi, t_\psi)$, est concernée par un seul des possibles supports. De ce fait, nous savons que $\exists j \in \{1,\dots,m\}, \mathrm{W}_j = \mathrm{W}$ et que $\mathrm{S}^\chi(t_\psi) \vDash \psi_j$, ce qui correspond à (c_1). L'élément disjoint associé à W_j est $\psi_j = l_1 \wedge \dots \wedge l_n$.

1. $\psi_{\mathrm{DNF}} = \psi_1 \vee \dots \vee \psi_m$ étant la forme normale disjonctive de ψ, $\forall i, j \in \{1,\dots,m\}^2, i \neq j, \psi_i \nvDash \psi_j$.

Avec cette information, nous pouvons déduire la relation causale : $C \underset{W_j}{\rightarrow} (\psi_j, t_\psi)$. D'après la proposition 6.4, étant donné que $(e,t) \in C$ et que ψ_j une conjonction de littéraux, nous déduisons que $\exists i \in \{1,\dots,n\}$, $S^\chi(t) \rhd \{e\} \models l_i \in \psi_j$, ce qui correspond à (c_2). Selon cette même proposition, nous pouvons déduire que $\forall t' \in \mathbb{T},\ t < t' \leq t_\psi$, $S^\chi(t') \models l_i$, ce qui correspond à (c_3), et que $S^\chi(t) \not\models l_i$, ce qui correspond à (c_4). Nous venons donc de prouver que lorsque (e,t) est une NESS-cause directe de (ψ, t_ψ), nous avons (c_1), (c_2), (c_3), et (c_4).

[$\Longleftarrow$] Soit une formule ψ sous FND vraie au temps t_ψ et (e,t) une occurrence d'évènement qui vérifie (c_1), (c_2), (c_3), et (c_4). Considérons les supports hypothétiques $W_1,\dots,W_m$ de (ψ, t_ψ), chacun correspondant à une des conjonctions de littéraux $\psi_1,\dots,\psi_m$. ψ étant sous forme normale disjonctive, étant minimale par subsomption et étant libre de tautologies, par la preuve de la proposition 6.4 nous pouvons déduire que les supports hypothétiques $W_1,\dots,W_m$ satisfont tous la condition 1 de la définition 6.10.

Comme (c_1) $\exists j \in \{1,\dots,m\}$, $S^\chi(t_\psi) \models \psi_j$ et $S^\chi(-1) = Lit_\mathbb{F} \setminus S^\chi(0)$, nous pouvons déduire que $\exists C \subseteq \mathbb{E} \times \mathbb{T}$ tel que $C \underset{W_j}{\rightarrow} (\psi_j, t_\psi)$. Puis, étant donné que nous savons (c_2) $S^\chi(t) \rhd \{e\} \models l \in \psi_j$, ainsi que (c_3) $\forall t' \in \mathbb{T},\ t < t' \leq t_\psi$ et (c_4) $S^\chi(t) \not\models l$, nous pouvons déduire avec la proposition 6.4 que (e,t) est une direct NESS-cause de (ψ_j, t_ψ). De ce fait, d'après la définition 6.10 nous savons que $(e,t) \in C$. Comme le support W_j dans la relation causale $C \underset{W_j}{\rightarrow} (\psi_j, t_\psi)$ satisfait également la condition 1 de la définition 6.10 pour la formule ψ, nous pouvons également déduire la relation causale $C \underset{W_j}{\rightarrow} (\psi, t_\psi)$. Nous venons donc de prouver que lorsque nous avons (c_1), (c_2), (c_3), et (c_4), (e,t) est une NESS-cause directe de (ψ, t_ψ). □

Les NESS-causes directes que nous avons présenté dans cette section sont l'essence de notre approche causale. Bien qu'elles soient nécessaires, ces relations ne sont pas suffisantes. En raison de la structure de $\mathscr{S}_c$, il se peut que les occurrences identifiées comme des NESS-causes directes ne nous permettent pas d'utiliser cette approche dans un raisonnement éthique.
Prenons l'exemple d'un agent qui lance une pierre vers une bouteille. Une possible trace d'évènements est illustrée sur la figure 6.3. Dans ce cas, si la conséquence qui nous intéresse est $(bouteille\_brisee, t+3)$, nous aurons l'occurrence $(bouteille\_se\_brise, t+2)$ comme NESS-cause directe. Du point de vue de l'éthique, et donc de l'éthique computationnelle, cette information n'est pas très intéressante. Comme ce que nous cherchons à évaluer sont les décisions des agents, il est essentiel de pouvoir établir une chaîne causale en remontant le temps de sorte à retrouver l'ensemble des actions qui sont derrière la conséquence. Dans notre exemple cela correspond à remonter le temps jusqu'à trouver l'occurrence $(agent\_lance\_pierre, t)$.

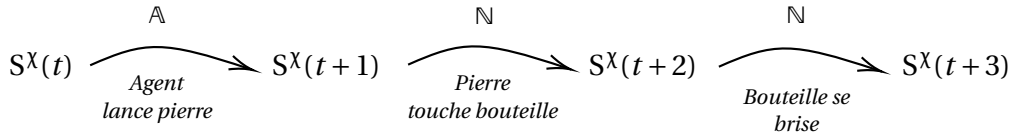


FIGURE 6.3 – Trace d'évènements d'un exemple où un agent lance une pierre vers une bouteille.

6.2.2 NESS-causes

Dans cette section nous introduisons les NESS-causes, une relation plus générale subsumant les NESS-causes directes. Par définition, les NESS-causes directes de (ψ, t_ψ) en sont également des NESS-causes. Cette relation est une spécification des $\mathscr{F}$-causes du chapitre 5. Ces relations causales permettent de retrouver la chaîne causale dont nous avons besoin pour le raisonnement éthique.
Si (ψ, t_ψ) est la formule vraie qui nous intéresse et que C est un de ses ensembles suffisants de NESS-causes directes, trouver les NESS-causes revient à s'intéresser aux causes de $(tri(C), t)$, où $t < t_\psi$ nécessairement. Dit autrement, il faut s'intéresser aux occurrences d'évènements qui ont causé le déclenchement des NESS-causes directes.

Dans la définition 6.11 nous définissons ce qu'est une NESS-cause. Comme pour la définition 6.10, nous définirons ce qu'est un *ensemble suffisant de NESS-causes*. Comme pour les NESS-causes directes, si une NESS-cause de (ψ, t_ψ) est un élément qui contribue à sa véracité, un ensemble suffisant de NESS-causes de (ψ, t_ψ) est un ensemble suffisant pour expliquer sa véracité.

Exemple 6.1 [suite]**.** *Si nous appliquons la définition 6.10 pour trouver les causes de* $(s_\leq, 2)$*, nous trouvons comme unique ensemble suffisant de NESS-causes directes* $C = \{(dev_o, 1)\}$*. Trouver un ensemble suffisant de NESS-causes revient à trouver un ensemble d'occurrences qui expliquent le déclenchement de* $(dev_o, 1)$*. Cela peut être vu comme chercher un ensemble qui pourrait se substituer à* $(dev_o, 1)$*. Pour trouver cet ensemble nous nous intéressons aux conditions de déclenchement de l'occurrence* $tri(dev_o) = o_e \vee (o_m \wedge se_{hs})$*. Nous avions vu précédemment dans l'exemple, lorsque nous avons introduit l'intuition de ce qu'était un ensemble suffisant de NESS-causes directes, que* $\{(prod_m, 0), (ini_{se_{hs}}, -1)\}$*, ainsi que* $\{(prod_e, 0)\}$ *sont tous deux des ensembles suffisants pour expliquer* $(tri(dev_o), 1)$*. Ces deux ensembles seront considérés alors comme des ensembles suffisants de NESS-causes de* $(s_\leq, 2)$*. Pour être exhaustif, il faut ajouter à ces deux ensembles* $C = \{(dev_o, 1)\}$*, puisque par définition nous avons dit que les NESS-causes directes sont aussi des NESS-causes.*

Dans la mesure où la recherche d'ensembles suffisants de NESS-causes peut être assimilée à un processus de substitution, nous utilisons des ensembles d'« *occurrences d'évènements retirables* » et des ensembles d'« *occurrences d'évènements substituantes* », notés respectivement C_R et C_S. Nous avons vu que pouvoir substituer un évènement passe par la recherche des évènements qui l'ont déclenché. Plus exactement, la recherche des causes de ses conditions de déclenchement qui pour rappel sont attribuées aux évènements par la fonction tri. En conséquence, puisque seuls les évènements naturels sont formalisés comme ayant des conditions de déclenchement, $tri : \mathbb{N} \to \mathscr{F}$, un ensemble d'occurrences d'évènements retirables ne peut contenir que des évènements naturels. Nous excluons également ceux au temps $t = -1$ car, notre formalisation étant bornée dans le passé, il n'est pas possible d'aller au delà, ces évènements représentent déjà tout ce qui a eu lieu avant. Formellement, $C_R \subseteq (\mathbb{N} \setminus E^X(-1)) \times \mathbb{T}$. Étant donné deux ensembles suffisants, un de NESS-causes directes C et l'autre de NESS-causes C', un ensemble d'occurrences d'évènements retirables est défini comme $C_R = C \setminus C'$. D'après la vision adoptée selon laquelle une occurrence d'évènement naturel peut être entièrement expliquée [Revaz, 2009], nous pouvons déduire qu'à tout ensemble C_R correspond un ensemble d'occurrences d'évènements substituantes C_S. Celui-ci est défini en conséquence comme $C_S = C' \setminus C$.

Définition 6.11 [NESS-causes]**.** *Étant donné un cadre causal* χ *et la relation* $C \underset{W}{\rightarrow} (\psi, t_\psi)$*, l'ensemble d'occurrences d'évènements* $C' = \{(e,t) | e \in E^\chi(t), t \in \mathbb{T}\}$ *est un ensemble suffisant de NESS-causes de* (ψ, t_ψ)*, relation que nous notons* $C' \dashrightarrow (\psi, t_\psi)$*, ssi un des cas suivants est vérifié :*

— *Cas base :* $C' = C$ *;*

— *Cas récursif : Soit un ensemble non vide d'occurrences d'évènements retirables* $C_R = C \setminus C'$ *et trois partitions associées à la séquence* $t_1, \ldots, t_k$ *:* $W_1, \ldots, W_k$ *de* W, $C(t_1), \ldots, C(t_k)$ *de* C *et* $C_R(t_1), \ldots, C_R(t_k)$ *de* C_R*. Il existe une séquence* $C_{S_1}, \ldots, C_{S_k}$*, non nécessairement monotone en temps, telle que :*

1. $C_S = \bigcup_{i \in \{1,\ldots,k\}} C_{S_i}$ *;*
2. $\forall i \in \{1, \ldots, k\}, C_R(t_i) = \varnothing \implies C_{S_i} = \varnothing$ *;*
3. $\forall i \in \{1, \ldots, k\}, C_R(t_i) \neq \varnothing \implies C_{S_i} \dashrightarrow (tri(C_R(t_i)), t_i)$.

(e,t) *est une NESS-cause de* (ψ, t_ψ) *ssi* $\exists C' \subseteq \mathbb{E} \times \mathbb{T}$ *tel que* $(e,t) \in C'$ *et* $C' \dashrightarrow (\psi, t_\psi)$.

Le cas de base de la définition 6.11 est celui indiquant qu'un ensemble suffisant de NESS-causes directes de (ψ, t_ψ) est également un de ses ensembles suffisants de NESS-causes. Le cas récursif de la définition 6.11 est essentiellement une spécification des cas où il est possible de trouver une relation de transitivité, avec la contrainte supplémentaire que nous voulons construire les ensembles suffisants. Étant donné le choix de n'inclure que des évènements naturels dans les ensembles d'occurrences d'évènements retirables, nous intégrons l'intuition venant de l'éthique que la transitivité est brisée par les actions, ou en tout cas qu'elle n'a pas la même force [Berreby et collab., 2018]. Nous verrons dans la section 7.4 que cette intuition est également retrouvée dans les discussions dans le domaine de la causalité.
Un ensemble de NESS-causes noté D est appelé un ensemble de causes volitionnelles si $D \subseteq (E^\chi(-1) \cup \mathbb{A}) \times \mathbb{T}$. Il s'agit d'un ensemble suffisant de NESS-causes où tous les évènements naturels ont été retirés et substitués soit par des actions, soit par les évènements dans $E^\chi(-1)$, donc les évènements qui représentent le passé et dont nous avons une connaissance incomplète. Si une action est une NESS-cause de (ψ, t_ψ), elle sera nécessairement dans un ensemble de causes volitionnelles. Si elle n'est dans aucun, elle ne l'est pas.

Comme dans la section précédente, la définition 6.11 ne définit pas une procédure pour trouver les ensembles suffisants de NESS-causes, elle définit lorsqu'un ensemble d'occurrences d'évènements peut être considéré comme tel. Dans les faits, cette procédure est simple. Nous la présentons ci-dessous et l'illustrons cette fois-ci à l'aide de l'exemple 6.1.
En premier lieu, il faut commencer par identifier tous les ensembles suffisants de NESS-causes directes de (ψ, t_ψ). À partir de chacun de ces ensembles que nous noterons C, il est possible de construire grâce au cas de base un ensemble suffisant de NESS-causes de (ψ, t_ψ), $C' = C$.

Exemple 6.1 [suite]**.** *La question qui se pose dans l'exemple 6.1 est : quelles sont les causes des habitants n'ayant plus accès à l'eau potable au temps* $t = 3$ *? Formellement, nous cherchons donc la cause de la formule vraie* $(\overline{o_p}, 3)$.
Si nous appliquons la définition 6.10 à cet exemple nous trouvons un unique ensemble suffisant de NESS-causes directes, $C = \{(dis_p, 2)\}$*. Nous pouvons donc déduire un premier ensemble suffisant de NESS-causes,* $C' = C = \{(dis_p, 2)\}$.

En second lieu, il faut former les trois partitions nécessaires au traitement du cas récursif. La première partition est celle associée au support W, essentiel à la relation $C \underset{W}{\rightarrow} (\psi, t_\psi)$. La partition $W_1, \ldots, W_k$ associée au temps $t_1, \ldots, t_k$ est la même que pour la définition des NESS-causes directes. La partition $C(t_1), \ldots, C(t_k)$ associée au temps $t_1, \ldots, t_k$ peut également être reprise du raisonnement sur les NESS-causes directes. La seule partition qui doit être construite est donc celle de C_R. Cela se fait en deux étapes.
La première consiste à trouver C_R. Pour cela, nous prenons le sous-ensemble de C ne comprenant que les évènements naturels en dehors de ceux à $t = -1$, $C_R = C \cap (\mathbb{N} \setminus E^X(-1) \times \mathbb{T})$.
La deuxième étape consiste à regrouper les occurrences d'évènements dans C_R en fonction des temps $t_1, \ldots, t_k$. En réalité $C_R(t_1), \ldots, C_R(t_k)$ n'est pas vraiment une partition contrairement aux deux précédentes. Celle-ci respecte les conditions 2 et 3 de la définition 6.9, mais ne respecte pas la condition 1. En effet, il est possible que certains éléments de la partition soient vides. C'est le cas si $\exists i \in \{1, \ldots, k\}, C(t_i) \subseteq (\mathbb{A} \cup E^X(-1)) \times \mathbb{T}$. Nous continuons à l'appeler une partition par abus de notation. Par contre, dans la mesure où $C_R \subseteq C$ et que C vérifie la condition 3 de la définition 6.10, il est impossible qu'un élément de C_R ne puisse pas être regroupé selon un des temps dans $t_1, \ldots, t_k$.

Exemple 6.1 [suite]**.** *Par la définition 6.10, nous savons que le support de la relation causale* $C \underset{W}{\rightarrow} (\psi, t_\psi)$*, où* $C = \{(dis_p, 2)\}$*, est* $W = \{\overline{o_p}\}$*, dont la partition est* $W_1 = \{\overline{o_p}\}$ *et la séquence associée* $t_1 = 2$*. La partition de* C *est* $C(t_1) = \{(dis_p, 2)\}$*.*
Maintenant, dans la mesure où $dis_p \in (\mathbb{N} \setminus E^X(-1))$*, nous avons* $C_R = \{(dis_p, 2)\}$ *et il est immédiat de construire* $C_R(t_1) = C_R = \{(dis_p, 2)\}$ *vu que* C_R *est un singleton.*

En dernier lieu, il faut s'intéresser à chaque élément de $C_R(t_1), \ldots, C_R(t_k)$ et trouver pour chacun un ensemble d'occurrences d'évènements substituantes C_{S_i} correspondant. Deux cas sont possibles. Le premier est celui où $C_R(t_i) = \varnothing$. Dans ce cas simple, le C_{S_i} correspondant est tout simplement $C_{S_i} = \varnothing$, comme l'indique la condition 2. Le deuxième cas est celui où $C_R(t_i) \neq \varnothing$. Nous retrouvons là le cas de figure récursif, le C_{S_i} correspondant est un ensemble suffisant de NESS-causes de $(tri(C_R(t_i)), t_i)$, donc $C_{S_i} \dashrightarrow (tri(C_R(t_i)), t_i)$ comme l'indique la condition 3. En dehors du cas limite dont nous parlerons par la suite, obtenir un ensemble suffisant de NESS-causes revient à faire l'union des C_{S_i} trouvés, comme l'indique la condition 1.
Notez que les C_{S_i} ne sont pas nécessairement monotones en temps contrairement aux $C_R(t_i)$. L'ensemble suffisant de NESS-causes de $(tri(C_R(t_i)), t_i)$ peut être composé d'occurrences d'évènements s'étant produites à différents temps.

Exemple 6.1 [suite]**.** *Nous avons l'ensemble* $C_R(t_1) = \{(dis_p, 2)\}$*, nous sommes donc dans le cas où nous avons* $C_R(t_1) \neq \varnothing$*. Nous devons alors trouver un* C_{S_1} *tel que* $C_{S_1} \dashrightarrow (tri(dis_p), 2)$*, avec* $tri(dis_p) = s_\leq$*. D'après la définition 6.10,* $C'_1 = \{(dev_o, 1)\}$ *est un ensemble suffisant de NESS-causes directes de* $(s_\leq, 2)$*, donc un ensemble suffisant de NESS-causes également par le cas de base de la définition 6.11. Nous avons donc une première solution :* $C_{S_1} = C'_1 = \{(dev_o, 1)\}$.

Le cas d'arrêt dans le processus récursif est lorsque $C_R = \varnothing$, donc $C_S = \varnothing$, et par conséquent le seul ensemble suffisant de NESS-causes de $(tri(C_R(t_i)), t_i)$ est son ensemble de NESS-causes directes par le cas base. Arriver à ce cas pour tous les éléments $C_R(t_1), \ldots, C_R(t_k)$ revient à construire les ensembles de causes volitionnelles D.

Exemple 6.1 [suite]**.** *Comme précisé,* $C_{S_1} = C'_1 = \{(dev_o, 1)\}$ *n'est qu'une première solution, celle-ci ne correspond pas au cas limite et donc n'est pas la solution qui nous servirait en éthique computationnelle. Réitérons le processus en appliquant le cas récursif pour trouver un autre* C_{S_1} *qui satisfasse* $C_{S_1} \dashrightarrow (s_{\leq}, 2)$*. Pour rester concis et ne pas introduire trop de notations, nous réutilisons les notations utilisées à l'itération précédente mais nous les réinitialisons.*
Reprenons depuis la première étape. Formellement, nous cherchons donc la cause de la formule vraie $(s_{\leq}, 2)$*.*
Si nous appliquons la définition 6.10 à cet exemple nous trouvons un unique ensemble suffisant de NESS-causes directes, $C = \{(dev_o, 1)\}$*. Nous pouvons donc déduire une solution supplémentaire,* $C_{S_1} = C = \{(dev_o, 1)\}$*.*

Passons à la deuxième étape. Par la définition 6.10, le support de la relation $C \underset{W}{\rightarrow} (\psi, t_\psi)$*, où* $C = \{(dev_o, 1)\}$*, est* $W = \{s_{\leq}\}$*, dont la partition est* $W_1 = \{s_{\leq}\}$ *et la séquence associée* $t_1 = 1$*. La partition de* C *est* $C(t_1) = \{(dev_o, 1)\}$*.*
Maintenant, dans la mesure où $dev_o \in (\mathbb{N} \setminus E^{\chi}(-1))$*, nous avons* $C_R = \{(dev_o, 1)\}$ *et il est immédiat de construire* $C_R(t_1) = C_R = \{(dev_o, 1)\}$ *vu que* C_R *est un singleton.*

Finalement, la dernière étape. Nous avons $C_R(t_1) = \{(dev_o, 1)\}$*, nous sommes donc dans le cas où* $C_R(t_1) \neq \varnothing$*. Nous devons alors trouver un ensemble* C_{S_1} *tel que* $C_{S_1} \dashrightarrow (tri(dev_o), 1)$*, où* $tri(dev_o) = o_e \vee (o_m \wedge se_{hs})$*. D'après la définition 6.10,* $C'_2 = \{(prod_m, 0), (ini_{se_{hs}}, -1)\}$ *et* $C'_3 = \{(prod_e, 0)\}$ *sont deux ensembles suffisants de NESS-causes directes de* $(o_e \vee (o_m \wedge se_{hs}), 1)$*, donc deux ensembles suffisants de NESS-causes également par le cas de base de la définition 6.11. Nous avons deux solutions supplémentaires :* $C_{S_1} = C'_2 = \{(prod_m, 0), (ini_{se_{hs}}, -1)\}$ *et* $C_{S_1} = C'_3 = \{(prod_e, 0)\}$*.*

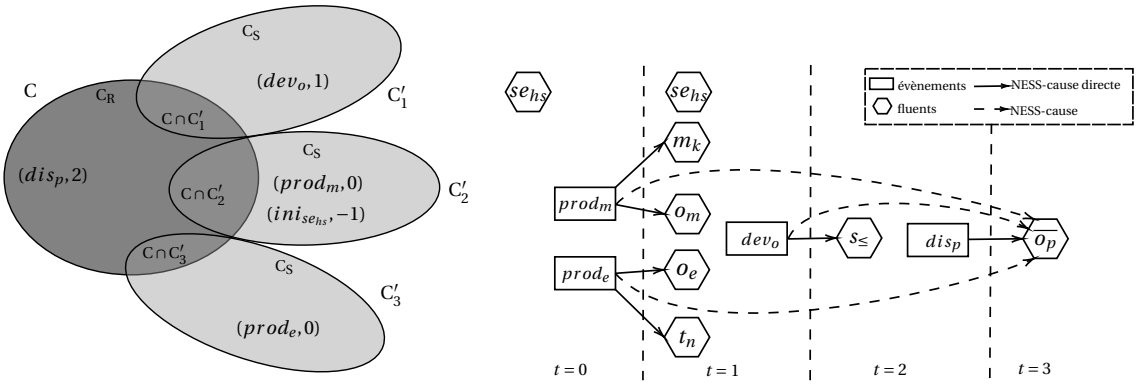
Si nous récapitulons nous avons quatre ensembles suffisants de NESS-causes de $(\overline{o_p}, 3)$ *:* $C = \{(dis_p, 2)\}$*,* $C'_1 = \{(dev_o, 1)\}$*,* $C'_2 = \{(prod_m, 0), (ini_{se_{hs}}, -1)\}$ *et* $C'_3 = \{(prod_e, 0)\}$*.*
D'après la définition 6.11, il n'existe pas d'autres ensembles suffisants de NESS-causes pour la formule vraie $(\overline{o_p}, 3)$*. En effet, nous avons atteint le cas limite pour tous les* C_R *issus de* C*. Le cas récursif et de base a été appliqué à* C *et à* C'_1*, et seul le cas de base peut être appliqué à* C'_2 *et* C'_3 *puisque* $C'_2 \subseteq (E^{\chi}(-1) \cup \mathbb{A}) \times \mathbb{T}$ *et* $C'_3 \subseteq (\mathbb{A} \times \mathbb{T})$*. Les seuls* C_R *pouvant en être tirés seront des ensembles vides ne permettant pas de vérifier la condition 1 de la définition 6.11 étant donné que* $C_R = \varnothing \implies C_S = \varnothing$*.* C'_2 *et* C'_3 *sont tous deux des ensembles de causes volitionnelles.*
La figure 6.4a est une représentation ensembliste du raisonnement que nous venons de réaliser pour trouver les quatre ensembles suffisants de NESS-causes pour $(\overline{o_p}, 3)$*. La figure 6.4b est une représentation des traces d'évènements et d'états auxquelles ont été ajoutées les NESS-causes de* $(\overline{o_p}, 3)$*.*

Proposition 6.6 [Non unicité des ensembles suffisants de NESS-causes]**.** *Étant donné un cadre causal* χ *et la relation causale* $C' \dashrightarrow (\psi, t_\psi)$*,* C' *n'est pas nécessairement l'unique ensemble suffisant de NESS-causes de* (ψ, t_ψ) *d'après la définition 6.11.*

Démonstration. La preuve est immédiate à partir de la définition 6.11 et la proposition 6.2. L'exemple 6.1 en est l'illustration. □

6.2.3 Causes effectives

Dans cette section nous introduisons les causes effectives. Cette relation est une spécification des causes effectives du chapitre 5. Comme nous l'avons vu à ce moment là, certains



(a) Représentation ensembliste du raisonnement permettant de trouver les quatre ensembles suffisants de NESS-causes de $(\overline{o_p}, 3)$ dans l'exemple 6.1.

(b) Représentation des traces d'évènements et d'états dans l'exemple 6.1 auxquelles ont été ajoutées les NESS-causes de $(\overline{o_p}, 3)$.

FIGURE 6.4 – Illustration de l'exemple 6.1 mettant en scène deux usines produisant des biens différents et ayant un impact sur l'accès à l'eau potable des habitants d'un village.

des formalismes ne font pas la distinction entre fluents et évènements. En l'occurrence, dans les approches basées sur les équations structurelles, toutes les variables sont considérées comme des évènements. Afin de permettre la comparaison avec le plus grand nombre d'approches possibles et pouvoir utiliser la typologie proposée précédemment, nous faisons le choix de définir ce type de relations causales commun à la plupart des approches.
Contrairement aux approches représentant uniquement des évènements, dans un STEE nous avons des états qui s'intercalent entre l'occurrence d'évènements. Qui plus est, ces occurrences dépendent de ces états. Dans les sections précédentes nous avons spécifié les $\mathscr{F}$-causes, relations causales reliant des occurrences d'évènements à des formules vraies. Les causes effectives qui relient deux occurrences d'évènements entre elles peuvent maintenant être définies en s'appuyant sur les $\mathscr{F}$-causes. L'occurrence d'un premier évènement est considérée comme une cause effective de l'occurrence d'un second si l'occurrence du premier est une NESS-cause des conditions de déclenchement du deuxième.

Définition 6.12 [causes effectives]**.** *Étant donné un cadre causal* χ *et deux occurrences d'évènements* (e,t) *et* (e_ψ, t_ψ)*,* (e,t) *est une cause effective de* (e_ψ, t_ψ)*, relation que nous notons* $(e,t) \rightsquigarrow (e_\psi, t_\psi)$*, ssi* (e,t) *est une NESS-cause de* $(tri(e_\psi), t_\psi)$*.*

Cette relation complète notre définition de causalité positive pour $\mathscr{S}_c$. L'ensemble de nos relations nous permet à présent de construire des chaînes causales. Nous pouvons par exemple déduire que, si l'occurrence (e', t_2) est une NESS-cause directe de (ψ, t_3) et que l'occurrence (e, t_1) est une cause effective de (e', t_2) avec $t_1 < t_2 < t_3$, alors l'occurrence (e, t_1) est une NESS-cause de (ψ, t_3). Voyons ce que cela nous permet de déduire pour l'exemple 6.1.

Exemple 6.1 [suite]**.** *Nous avons vu précédemment qu'il existe quatre ensembles suffisants de NESS-causes de* $(\overline{o_p}, 3)$ *:* $C = \{(dis_p, 2)\}$, $C'_1 = \{(dev_o, 1)\}$, $C'_2 = \{(prod_m, 0), (ini_{se_{hs}}, -1)\}$ *et* $C'_3 = \{(prod_e, 0)\}$*. D'après la définition 6.11, toutes les occurrences d'évènements dans ces ensembles sont individuellement des NESS-causes de* $(\overline{o_p}, 3)$*. Si nous excluons* C *qui est le seul des ensembles suffisants à être à la fois de NESS-causes et de NESS-causes directes, toutes les*

occurrences d'évènements dans les trois autres ensembles sont individuellement des NESS-causes de $(tri(dis_p),2)$. *Il en découle d'après la définition 6.12 que les occurrences d'évènements* $(dev_o,1),(prod_m,0),(prod_e,0),(ini_{se_{hs}},-1)$ *sont toutes des causes effectives de* $(dis_p,2)$. *La figure 6.5 représente les traces d'évènements et d'états de l'exemple et ajoute toutes les relations causales pouvant en être déduites.*

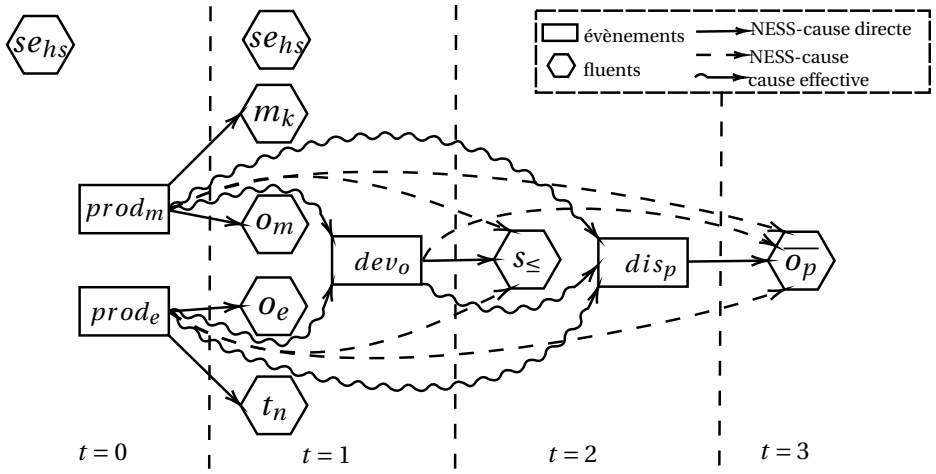


FIGURE 6.5 – Représentation des traces d'évènements et d'états dans l'exemple 6.1 auxquelles ont été ajoutées toutes les relations causales pouvant en être déduites.

6.2.4 Propriétés face à la surdétermination

Dans cette section nous montrons les propriétés de notre définition par rapport à la typologie de la surdétermination du chapitre 5. Plus spécifiquement nous montrons que l'approche présentée dans ce chapitre est sensible à la surdétermination préemptive, duplicative et symétrique au sens de la définition 5.14. Pour pouvoir soumettre notre approche, ou n'importe quelle approche, à cette typologie et prouver ses propriétés face à la surdétermination, il faut d'abord établir les liens nécessaires entre le formalisme utilisé et $\mathscr{S}_s$, ainsi qu'entre les relations causales de l'approche et celles de la section 5.1.2.

Lemme 6.1. *$\mathscr{S}_c$ peut être vu comme une spécification de $\mathscr{S}_s$ et les relations causales des définitions 6.11 et 6.12 sont respectivement des spécifications des relations causales des définitions 5.4 et 5.5.*

Démonstration. Pour pouvoir confronter l'approche causale que nous proposons à la typologie obtenue dans le chapitre 5, il faut que nous montrions que $\mathscr{S}_c$ peut être vu comme une spécification de $\mathscr{S}_s$ et que les relations causales définies correspondent bien aux deux types de relations utilisées pour définir la typologie.

Commençons par prouver que $\mathscr{S}_c$ peut être vu comme une spécification de $\mathscr{S}_s$. D'un point de vue global, les deux représentations sont des STEE et possèdent donc la même structure. Il faut alors montrer que $\mathscr{S}_s$ est plus général que $\mathscr{S}_c$, i.e. tout problème dans $\mathscr{S}_c$ doit pouvoir se représenter dans $\mathscr{S}_s$ même si quelques nuances se perdent une fois passés dans le dernier.

Si nous commençons par les états décrits dans la section 6.1.1 nous constatons quelques différences. Dans $\mathscr{S}_c$ des littéraux de fluents sont utilisés, il y a plus de conditions pour qu'un ensemble de fluents soit considéré un état et il y a des contraintes sur la syntaxe des formules d'état. La première différence est uniquement une question de notations et ne modifie en rien ce qui peut être représenté. Quant à la deuxième différence, elle implique que des états dans $\mathscr{S}_s$ n'en seront peut-être pas dans $\mathscr{S}_c$. Toutefois, comme cette différence ne change pas le fait que les états dans $\mathscr{S}_c$ en sont dans $\mathscr{S}_s$, alors $\mathscr{S}_s$ reste plus général que $\mathscr{S}_c$. Le même raisonnement peut être appliqué à la troisième différence. Toutes les formules d'état $\mathscr{F}$ dans $\mathscr{S}_c$ le sont dans $\mathscr{S}_s$.
Parlons maintenant des évènements présentées dans la section 6.1.2. Il existe également quelques différences. Dans $\mathscr{S}_c$ la fonction tri a été ajoutée, les effets positifs et négatifs sont donnés par la même fonction *eff*, il est possible d'établir des priorités entre évènements pour leur déclenchement et il y a une distinction entre actions et évènements naturels. Comme précédemment, soit ces différences reposent uniquement sur une question de notation, soit elles apportent des subtilités dans $\mathscr{S}_c$ mais ne changent pas le fait que $\mathscr{S}_s$ reste plus général que $\mathscr{S}_c$.
En ce qui concerne les transitions entre états introduites dans la section 6.1.2, la seule différence est la présence du point de temps $t = -1$ dans $\mathscr{S}_c$. Cette différence ne fait qu'ajouter un point de temps supplémentaire dont les informations sur l'état correspondant S_{-1} et les évènements qui s'y produisent E_{-1} peuvent être entièrement déduits de l'état initial S_0 et donc représentables sans difficulté dans $\mathscr{S}_s$.

À ce stade nous avons parlé de toutes les différences concernant les contextes κ_s et κ_c. Nous avons pu voir que κ_s est plus général que κ_c qui peut donc être vu comme une spécification. La grande différence entre ces deux STEE est la définition du cadre causal χ. Alors que $\mathscr{S}_s$ utilise une politique π_s, $\mathscr{S}_c$ utilise plutôt un scénario σ pour définir le couple χ. Pour montrer que $\mathscr{S}_c$ peut être vu comme une spécification de $\mathscr{S}_s$, il nous faut alors montrer qu'il existe une politique, au sens de la définition 5.2, pouvant remplacer σ dans le χ de $\mathscr{S}_c$.
Soit une politique $\pi_{\mathbb{N}} = \{e \in \mathbb{N} | S \vDash tri(e)\}$ qui à chaque état associe l'ensemble des évènements naturels dont les conditions de déclenchement y sont satisfaites. Le scénario σ ne peut pas simplement être remplacé par la politique $\pi_{\mathbb{N}}$, celle-ci ne gère que le cas des évènements naturels. En effet, elle couvre la condition 2c dans la définition 6.2 dans le cas sans priorités. Pour gérer les actions, il faut introduire une politique qui en plus couvre la condition 3 dans la définition 6.8. Cette politique que nous notons π_σ est définie comme suit : $\forall S, \forall t,\ \pi_\sigma(S,t) = \pi_{\mathbb{N}}(S,t) \cup \{a \in \mathbb{A} | S \vDash pre(a) \wedge (a,t) \in \sigma\}$. Cette politique peut être définie $\forall \sigma' \subseteq \sigma$. Cela nous permet alors de construire également des politiques contractuelles dans $\mathscr{S}_c$. La politique permettant le raisonnement contrefactuel peut être définie comme $\pi_*(S,t) = \bigcup_{\sigma' \subseteq \sigma} \pi_{\sigma'}(S,t)$.
Venant de montrer qu'il existe une politique, au sens de la définition 5.2, pouvant remplacer σ dans le χ de $\mathscr{S}_c$, nous pouvons alors conclure que $\mathscr{S}_c$ peut être vu comme une spécification de $\mathscr{S}_s$.

Montrons maintenant que les relations causales définies correspondent bien aux deux types de relations utilisées pour définir la typologie. Les NESS-causes définies dans la définition 6.10 et 6.11 sont bien un type de relations où une occurrence d'évènement (e,t) est la cause et une formule vraie (ψ, t_ψ) est la conséquence. Nous avons donc là une spécification de $\mathscr{F}$-causes comme définies dans la définition 5.4. Les causes effectives dans la définition 6.12 sont bien un type de relation où une occurrence d'évènement (e,t) est la cause

et une autre occurrence d'évènements (e', t') est la conséquence. Nous avons donc là une spécification des causes effectives comme définies dans la définition 5.5. □

Lemme 6.2. *Étant donné un cadre causal* χ*, un chemin causal complet* ω *qui relie* (e_m, t_m) *à* (e_ψ, t_ψ) *et une occurrence d'évènement* $(e_i, t_i) \in \omega$, (e_i, t_i) *est une NESS-cause de* (ψ, t_ψ).

Démonstration. Soit $P(i)$ la proposition suivante : étant donné un chemin causal ω qui relie (e_m, t_m) à (e_ψ, t_ψ) et une occurrence d'évènement $(e_i, t_i) \in \omega$, (e_i, t_i) est une NESS-cause de (ψ, t_ψ). Nous faisons une preuve par récurrence sur i.

Cas base : Soit le cas où $i = 1$. S'agissant d'un chemin causal complet, nous savons par la définition 5.6 que (e_1, t_1) est une NESS-cause directe de (ψ, t_ψ). Donc, d'après la définition 6.10, $\exists C \subseteq \mathbb{E} \times \mathbb{T}$ tel que $(e_1, t_1) \in C$ et C un ensemble suffisant de NESS-causes directes de (ψ, t_ψ). D'après le cas base de la définition 6.11 sur les NESS-causes, un ensemble suffisant de NESS-causes directes de (ψ, t_ψ) est également un ensemble suffisant de NESS-causes de (ψ, t_ψ). Donc, comme C est un ensemble suffisant de NESS-causes de (ψ, t_ψ) et que $(e_1, t_1) \in C$, alors (e_1, t_1) est une NESS-cause de (ψ, t_ψ). $P(1)$ est donc vraie.

Cas récursif : Nous montrons maintenant que pour tout $k \geq 1$, si notre hypothèse de récursion $P(k)$ est vraie, alors notre hypothèse $P(k+1)$ l'est également. Notre hypothèse de récursion $P(k)$ dit que : étant donné un chemin causal ω qui relie (e_m, t_m) à (e_ψ, t_ψ) et une occurrence $(e_k, t_k) \in \omega$, (e_k, t_k) est une NESS-cause de (ψ, t_ψ). La proposition $P(k+1)$ dit que : étant donné un chemin causal ω qui relie (e_m, t_m) à (e_ψ, t_ψ) et une occurrence $(e_{k+1}, t_{k+1}) \in \omega$, (e_{k+1}, t_{k+1}) est une NESS-cause de (ψ, t_ψ).

S'agissant d'un chemin causal complet, nous savons par la définition 5.6 que (e_{k+1}, t_{k+1}) est une NESS-cause directe de $(tri(e_k), t_k)$. De ce fait, $\exists C' \subseteq \mathbb{E} \times \mathbb{T}$ tel que $(e_{k+1}, t_{k+1}) \in C'$ et C' un ensemble suffisant de NESS-causes directes de $(tri(e_k), t_k)$. Par le même raisonnement que celui dans le cas base, nous savons que C' est également un ensemble suffisant de NESS-causes de $(tri(e_k), t_k)$.

Étant donné que (e_k, t_k) est une NESS-cause de (ψ, t_ψ) d'après notre hypothèse de récursion, $\exists C'' \subseteq \mathbb{E} \times \mathbb{T}$ tel que $(e_k, t_k) \in C''$ et C'' un ensemble suffisant de NESS-causes de (ψ, t_ψ).

Nous avons alors : (i) C un ensemble suffisant de NESS-causes directes de (ψ, t_ψ) ; (ii) C' un ensemble suffisant de NESS-causes de $(tri(e_k), t_k)$ et $(e_{k+1}, t_{k+1}) \in C'$; (iii) C'' un ensemble suffisant de NESS-causes de (ψ, t_ψ) et $(e_k, t_k) \in C''$. Nous considérons un nouvel ensemble d'occurrences d'évènements, $C''' = \big(C'' \setminus \{(e_k, t_k)\}\big) \cup C'$. Nous voulons prouver que cet ensemble est un ensemble suffisant de NESS-causes de (ψ, t_ψ). Deux cas peuvent être considérés : soit $C'' = C$, ou non.

Dans le premier cas, nous avons $(e_k, t_k) \in C$ et donc (e_k, t_k) appartient à un des éléments de la partition de l'ensemble d'occurrences retirables C_R. Dans ce cas, le retrait de (e_k, t_k) de l'ensemble C_R est compensé par l'union avec C' qui est un ensemble suffisant de NESS-causes de $(tri(e_k), t_k)$. Donc, la suffisance est préservée.

Dans le deuxième cas, nous avons (e_k, t_k) appartenant à un sous-ensemble de l'ensemble d'occurrences d'évènements substituantes C_{S_i} qui a donc substitué un ensemble d'occurrences d'évènements retirables $C_R(t_i)$. Comme précédemment, le retrait de (e_k, t_k) de C_{S_i} est compensé par l'union avec C' qui est un ensemble suffisant de NESS-causes de $(tri(e_k), t_k)$. Donc, la suffisance est préservée.

Par conséquent, C''' est un ensemble suffisant de NESS-causes de (ψ, t_ψ). Comme nous avons $(e_{k+1}, t_{k+1}) \in C'''$, cette occurrence d'évènement est une NESS-cause de (ψ, t_ψ). Nous venons alors de prouver que si $P(k)$ est vraie, alors $P(k+1)$ l'est aussi.

Conclusion : Étant donné que le cas base et le cas récursif sont prouvés vrais, par récursion mathématique nous savons que la proposition $P(i)$ est vraie pour tous les i. □

Théorème 6.1. *L'approche de causalité effective présentée dans ce chapitre est sensible à la surdétermination préemptive, duplicative et symétrique au sens de la définition 5.14.*

Démonstration. Nous nous plaçons dans le cadre causal χ et dans un cas de surdétermination. Par la définition 5.7, nous avons alors $e_\psi \in E^\chi(t_\psi)$, $(e^1_m, t^1_m), (e^2_n, t^2_n)$ deux occurrences d'évènements, et trois cadres causaux contrefactuels :

$$\chi^1_I = (\pi^\sigma \setminus \{e^2_n\}, \kappa_c), \qquad \chi^2_I = (\pi^\sigma \setminus \{e^1_m\}, \kappa_c), \qquad \chi_- = (\pi^\sigma \setminus \{e^1_m, e^2_n\}, \kappa_c).$$

De plus, nous considérons deux chemins causaux : ω^1 dans χ^1_I qui relie (e^1_m, t^1_m) à (e_ψ, t_ψ) et ω^2 dans χ^2_I qui relie (e^2_n, t^2_n) à (e_ψ, t_ψ). Pour rappel, la typologie est construite avec les hypothèses que $|\Omega^1_I| = |\Omega^2_I| = 1$, i.e. qu'il y a un unique chemin causal par ensemble de chemins causaux individuels, $\Omega^i \setminus \Omega^i_I = \varnothing$, i.e. qu'aucun chemin causal n'est créé de l'interaction entre ω^1 et ω^2 dans χ. De ce fait, $\Omega^1_I = \Omega^1 = \{\omega^1\}$ et $\Omega^2_I = \{\omega^2\}$. Deux cas sont alors possibles : soit $\Omega^2 = \{\omega^2\}$, ou $\Omega^2 = \varnothing$.

[Cas surdétermination symétrique et duplicative] : nous sommes dans le cas $\Omega^2 = \{\omega^2\}$. Nous voulons prouver que notre approche de causalité effective est sensible à la surdétermination symétrique et duplicative. D'après la définition 5.14, cela veut dire que dans ces deux cas de surdétermination, notre approche causale considère aussi bien (e^1_m, t^1_m) que (e^2_n, t^2_n) comme des causes effectives de (e_ψ, t_ψ).

D'après le lemme 6.2, étant donné que ω^1 et ω^2 sont tous deux des chemins causaux dans χ, les occurrences d'évènements (e^1_m, t^1_m) et (e^2_n, t^2_n) sont considérées des NESS-causes de (ψ, t_ψ). Par conséquent, d'après la définition 6.12, aussi bien (e^1_m, t^1_m) que (e^2_n, t^2_n) sont considérées comme des causes effectives de (e_ψ, t_ψ).

Compte tenu des relations causales $(e^1_1, t^1_1) \underset{W^1}{\rightarrow} (\psi, t_\psi)$ et $(e^2_1, t^2_1) \underset{W^2}{\rightarrow} (\psi, t_\psi)$, ce résultat est vrai que ce soit dans le cas où $W^1 = W^2$ ou dans le cas où $W^1 \neq W^2$, et cela indépendamment de la relation temporelle entre t^1_1 et t^2_1. De ce fait, notre approche causale est sensible aux cas de surdétermination duplicative asynchrone, duplicative synchrone et symétrique.

[Cas surdétermination préemptive] : nous sommes dans le cas où $\Omega^2 = \varnothing$. Nous voulons prouver que notre approche de causalité effective est sensible à la surdétermination préemptive. D'après la définition 5.14, cela veut dire que dans ces cas de surdétermination, notre approche causale considère (e^1_m, t^1_m) comme une cause effective de (e_ψ, t_ψ), mais pas (e^2_n, t^2_n).

D'après le lemme 6.2, étant donné que ω^1 est un chemin causal dans χ, l'occurrence d'évènement (e^1_m, t^1_m) est considérée une NESS-cause de (ψ, t_ψ). Par conséquent, d'après la définition 6.12, (e^1_m, t^1_m) est considérée comme une cause effective de (e_ψ, t_ψ).

Considérons maintenant le cas de (e^2_n, t^2_n). Étant donné que ω^2 n'est pas un chemin causal dans χ, nous savons que : soit $\exists (e^2_i, t^2_i) \in \omega^2$ tel que, (e^2_1, t^2_1) n'est pas une NESS-cause directe de (ψ, t_ψ), soit (e^2_i, t^2_i) n'est pas une NESS-cause directe de $(tri(e^2_{i-1}), t^2_{i-1})$. En reprenant le raisonnement utilisé dans la preuve du lemme 6.2, nous pouvons déduire qu'étant donné nos hypothèses, $|\Omega^2_I| = 1$ et $\Omega^2 \setminus \Omega^2_I = \varnothing$, dans aucun des deux cas (e^2_n, t^2_n) ne peut être considérée une NESS-cause de (ψ, t_ψ).
Dans le premier cas, les NESS-causes étant construites en remontant le temps à partir d'un ensemble suffisant de NESS-causes directes, l'absence d'un tel ensemble rend impossible

le cas base et le cas récursif dans la définition 6.11.
Dans le deuxième cas, comme (e_i^2, t_i^2) n'est pas une NESS-cause directe de $(tri(e_{i-1}^2), t_{i-1}^2)$, il n'y a pas de moyen que (e_i^2, t_i^2) soit une NESS-cause de $(tri(e_{i-1}^2), t_{i-1}^2)$. En effet, le même raisonnement que pour le cas précédent peut être appliqué avec $(tri(e_{i-1}^2), t_{i-1}^2)$ étant le nouveau (ψ, t_ψ). Par conséquent, d'après la définition 6.12, comme (e_n^2, t_n^2) n'est pas une NESS-cause de (ψ, t_ψ), (e_n^2, t_n^2) n'est pas non plus une cause effective de (e_ψ, t_ψ).

Ces résultats étant applicables aussi bien dans le cas où :

$$\exists (e_j^2, t_j^2) \in \omega^2, (e_\psi, t_\psi) \rightarrow (\neg tri(e_j^2), t_j^2),$$

que dans le cas où :

$$\exists \Big((e_i^1, t_i^1) \in \omega^1, (e_j^2, t_j^2) \in \omega^2 \Big), (e_i^1, t_i^1) \rightarrow (\neg tri(e_j^2), t_j^2),$$

notre approche causale est sensible aux cas de surdétermination préemptive précoce et préemptive tardive. □

Notez que les cas de surdétermination imitative ne peuvent pas être formalisés dans $\mathscr{S}_c$ étant donné qu'il y est impossible de considérer que l'occurrence d'un évènement ait comme effet effectif un littéral de fluent qui était déjà vrai.

Avant de finir cette section sur la surdétermination, il est pertinent de préciser qu'être en mesure de traiter les différents cas de surdétermination n'est pas simplement une question causale, il faut également être en mesure de représenter les problèmes. Pour ce faire, il est indispensable de permettre des préconditions disjonctives et la cooccurrence d'évènements, et il est conseillé de disposer d'évènements naturels. Le tableau 6.1 est une comparaison de différentes approches faisant le lien entre langages de représentation de l'action et du changement et causalité effective en fonction de la présence ou non de ces éléments. À terme, si le travail réalisé dans cette section est fait pour toutes ces approches, il serait également possible de comparer formellement leurs divergences sur la définition de causalité.

	Évènements naturels ou axiomes	Cooccurrence d'évènements	Préconditions disjonctives
Batusov et Soutchanski [2018]			X
Berreby et collab. [2018]	X		
LeBlanc et collab. [2019]	X	X	
Sarmiento et collab. [2023]	X	X	X

TABLEAU 6.1 – Comparaison de différentes approches faisant le lien entre langages de représentation de l'action et du changement et causalité effective en fonction de la présence ou non de préconditions disjonctives, cooccurrence d'évènements et évènements naturels.

6.3 Automatisation : implémentation complète et correcte en ASP

Nous avons à présent $\mathscr{S}_c$, un langage de description d'action adapté au raisonnement causal et éthique et une définition de causalité positive qui y est adaptée. Ensemble, ils forment la première marche de notre approche permettant de raisonner sur la causalité. Powers et Ganascia [2020] identifient plusieurs défis auxquels il faut faire face pour réussir à modéliser un raisonnement éthique ; « évaluer les conséquences des actions » en est un et cette première marche permet d'y répondre.
Dans cette section nous parlerons principalement de l'implémentation de cette première

marche en programmation logique. Plus précisément, nous présentons une implémentation complète et correcte de nos définitions de causalité. Le choix de l'Answer Set Programming (ASP) pour l'implémentation a été motivé par un autre défi identifié par POWERS et GANASCIA [2020] : être capables de « surmonter les contradictions logiques » qui se présentent couramment dans le raisonnement éthique en raison de conflits potentiels entre règles. L'ASP a en effet montré être une forme de programmation déclarative non monotone adaptée à traiter les questions éthiques [GANASCIA, 2015].
Le programme logique obtenu permet de raisonner sur des situations causales complexes. Un tel programme permet à l'éthique computationnelle de pouvoir traiter des cas qui ne l'étaient pas auparavant.

Avant de rentrer dans les détails de l'implémentation, faisons un bref récapitulatif de la sémantique et la syntaxe de l'ASP. Dans ce langage de programmation logique, les problèmes sont encodés sous forme de programmes disjonctifs étendus [GELFOND et LIFSCHITZ, 1991]. Un programme disjonctif étendu est un ensemble de règles r de la forme :

$$\mathrm{L}_1;\ldots;\mathrm{L}_l \leftarrow \mathrm{L}_{l+1},\ldots,\mathrm{L}_m,\ not\ \mathrm{L}_{m+1},\ldots,not\ \mathrm{L}_n,$$

où chaque $\mathrm{L}_i \in Lit$ est un littéral positif ou négatif, not est la *négation par l'échec*, « ; » représente la disjonction, « , » représente la conjonction et $n \geq m \geq l \geq 0$. Notez que c'est à la présence de la négation par l'échec que cette forme de programmation déclarative doit le qualificatif de non monotone, elle permet de raisonner en l'absence d'information.
Dans les règles nous distinguons trois ensembles de littéraux : d'un côté la tête de la règle, $head(r) = \{\mathrm{L}_1,\ldots,\mathrm{L}_l\}$, de l'autre le corps décomposable en deux, $body^+(r) = \{\mathrm{L}_{l+1},\ldots,\mathrm{L}_m\}$ et $body^-(r) = \{\mathrm{L}_{m+1},\ldots,\mathrm{L}_n\}$. Une règle est considérée une contrainte d'intégrité si $head(r) = \varnothing$ et un fait si $body(r) = \varnothing$.

Nous notons Π un programme en ASP. La sémantique d'un programme disjonctif étendu est celle des modèles stables. En voici un aperçu. Un ensemble $\mathrm{S} \subseteq Lit$ satisfait une règle r si $body^+(r) \subseteq \mathrm{S}$ et $body^-(r) \cap \mathrm{S} = \varnothing$ impliquent $head(r) \cap \mathrm{S} \neq \varnothing$. Il est dit que S satisfait un programme Π si S satisfait toutes les règles dans ce programme.
Soit Π un programme tel que $\forall r \in \Pi,\ body^-(r) = \varnothing$. Un ensemble $\mathrm{S} \subset Lit$ est un modèle stable cohérent de Π, ce que nous notons $\mathrm{S} \in \mathrm{AS}(\Pi)$, si S est un ensemble minimal qui satisfait toutes les règles de Π et que S ne contient pas un littéral et son complément. Maintenant le cas général. Soit Π un programme disjonctif étendu et $\mathrm{S} \subseteq Lit$. Pour chaque règle r dans Π, la règle $r^{\mathrm{S}} : head(r) \leftarrow body^+(r)$ est incluse dans la réduction Π^{S} si $body^-(r) \cap \mathrm{S} = \varnothing$. Alors, $\mathrm{S} \in \mathrm{AS}(\Pi)$ si $\mathrm{S} \in \mathrm{AS}(\Pi^{\mathrm{S}})$.

Proposition 6.7. *Étant donné un programme* Π*, un modèle stable cohérent* $\mathrm{S} \in \mathrm{AS}(\Pi)$ *et un littéral* ρ *:*

$$\rho \in \mathrm{S} \implies \exists r \in \Pi,\ \big(head(r) = \rho\big) \wedge \big(body^+(r) \subseteq \mathrm{S}\big) \wedge \big(body^-(r) \cap \mathrm{S} = \varnothing\big).$$

La causalité effective étant le cœur de ce chapitre, nous détaillerons principalement l'implémentation de notre approche de causalité effective définie dans la section 6.2 pour obtenir le programme $\pi_{\mathbb{C}}$ en ASP.
En ce qui concerne l'implémentation du langage de description d'action qui décrit $\mathscr{S}_c$ introduit dans la section 6.1 pour obtenir le programme $\pi_{\mathbb{A}}$ en ASP, nous ne détaillerons que les règles essentielles à la compréhension générale de son fonctionnement et nécessaires à

la compréhension totale de $\pi_{\mathbb{C}}$ [2].
Il est important de noter que, comme dans $\mathbb{ACE}$ [BOURGNE et collab., 2021], aussi bien $\pi_{\mathbb{A}}$ que $\pi_{\mathbb{C}}$ sont indépendants du domaine, quel que soit le problème traité, ces deux programmes n'ont pas besoin d'être modifiés. Ils utilisent tous deux des variables de trois types : des variables de temps $\mathbb{T}$, de fluents $\mathbb{F}$ et d'évènements $\mathbb{E}$.

Si $\pi_{\mathbb{A}}$ et $\pi_{\mathbb{C}}$ sont indépendants du domaine, il faut que l'information propre au domaine soit donnée par d'autres programmes. Plus exactement deux autres programmes. $\pi_{sce}(\sigma)$ est le programme ASP obtenu en traduisant le scénario $\sigma \subseteq \mathbb{A} \times \mathbb{T}$. Cette traduction consiste à représenter chaque couple du scénario $(a,t) \in \sigma$ par le prédicat `performs(a,t)`.
$\pi_{con}(\kappa_c)$ est le programme ASP obtenu en traduisant $\kappa_c = (\mathbb{E}, \mathbb{F}, pre, tri, eff, S_0, \succ_{\mathbb{E}}, \mathbb{T})$. Comme pour $\pi_{\mathbb{A}}$, nous ne détaillerons que les règles essentielles à la compréhension générale de son fonctionnement et nécessaires à la compréhension totale de $\pi_{\mathbb{C}}$. L'ensemble $\mathbb{T}$ est représenté par `time(0..N)`. Chaque fluent $f \in \mathbb{F}$ est représenté par le prédicat `fluent(f)` et les fluents $f \in S_0$ sont en plus représentés par le prédicat `initially(f)`. Chaque action et évènement naturel dans $\mathbb{E}$ est respectivement traduit comme :
`action(action_name,action_pre,action_eff)`,
`auto(natu_event_name,natu_event_tri,natu_event_eff)`,
où `action_name`$\in \mathbb{A}$, `natu_event_name`$\in \mathbb{N}$, `action_pre` un « goal descriptor (GD) » comme en PDDL permettant d'identifier les préconditions des évènements comme pre, `natu_event_tri` un GD permettant d'identifier les conditions de déclenchement des évènements comme tri et `action_eff` et `natu_event_eff` deux GD permettant d'identifier les effets intrinsèques des évènements comme eff. Les prédicats `conj(GD)`, `disj(GD)` et `in(GD,GD_L)` permettent la construction de formules de $\mathscr{F}$.

Exemple 6.1 [suite]**.** *Voyons comment sont traduits les évènements en prenant l'exemple de l'évènement naturel représentant le fait de déverser des eaux usées industrielles dans le lac $dev_o \in \mathbb{N}$. Pour rappel, $tri(dev_o) = o_e \vee (o_m \wedge se_{hs})$ et $eff(dev_o) = s_{\leq}$. Dans $\pi_{con}(\kappa_c)$ cet évènement est représenté de la façon suivante :*

$auto(dev\_o, dev\_oCond, dev\_oEff).$
$disj(dev\_oCond).$
$in(dev\_oCond, o\_e).$
$in(dev\_oCond, dev\_oCond1).$
$conj(dev\_oCond1).$
$in(dev\_oCond1, o\_m).$
$in(dev\_oCond1, se\_hs).$
$conj(dev\_oEff).$
$in(dev\_oEff, s\_\leq).$

Cette section est divisée en deux sections. La section 6.3.1 présente le programme $\pi_{\mathbb{A}}$. Puis, la section 6.3.2 introduit le programme $\pi_{\mathbb{C}}$ et prouve que la traduction pour l'obtenir est complète et correcte.

6.3.1 Le programme $\pi_{\mathbb{A}}$

Le premier groupe de règles détermine le prédicat `holds(F,T)` indiquant que $f \in S^X(t)$ avec $f \in \mathbb{F}$. (6.1) indique que $f \in S^X(0)$ si $\pi_{con}(\kappa_c) \models$ `initially(f)`. (6.2) indique $f \in S^X(t+1)$ si f a été initié par un évènement au temps t. Finalement, (6.3) indique que $f \in S^X(t+1)$

2. Pour les lecteurs intéressés, le code complet est retranscrit dans l'annexe B et disponible sur `https://gitlab.lip6.fr/sarmiento/jair2022`.

si $f \in \mathrm{S}^{\chi}(t)$ et qu'il n'a été terminé par aucun évènement. Dans cette dernière règle nous retrouvons notre hypothèse que la loi d'inertie s'applique à tous les fluents.

$$holds(\mathrm{F},0) :- initially(\mathrm{F}), fluent(\mathrm{F}). \tag{6.1}$$
$$holds(\mathrm{F},\mathrm{T}+1) :- initiated(\mathrm{E},\mathrm{F},\mathrm{T}). \tag{6.2}$$
$$holds(\mathrm{F},\mathrm{T}+1) :- holds(\mathrm{F},\mathrm{T}), fluent(\mathrm{F}), time(\mathrm{T}), not\ terminated(\mathrm{E},\mathrm{F},\mathrm{T}) : event(\mathrm{E}). \tag{6.3}$$

Comme nous l'avons vu dans le chapitre 3, dès les travaux de MILL [1843], toutes les approches par régularité et inférence, dont celle de WRIGHT [1985], adoptent l'idée que la causalité ne peut être suffisante que si nous tenons compte, en plus des conditions devant être vraies, des conditions devant être fausses à la survenue du résultat. Les évènements étant des causes de leur absence sont également des causes du résultat. En travaillant sur les littéraux de fluents, notre définition de la causalité tient déjà compte de cette notion.
Les règles suivantes étendent les prédicats `holds` et `initially` pour gérer les littéraux de fluents. (6.4) indique que $\pi_{con}(\kappa_c) \models$`initially(-f)` si $\pi_{con}(\kappa_c) \not\models$`initially(f)`. (6.5) indique que $\neg f \in \mathrm{S}^{\chi}(t)$ si $f \notin \mathrm{S}^{\chi}(t)$.

$$initially(neg(\mathrm{F})) :- not\ initially(\mathrm{F}), fluent(\mathrm{F}). \tag{6.4}$$
$$holds(neg(\mathrm{F}),\mathrm{T}) :- not\ holds(\mathrm{F},\mathrm{T}), fluent(\mathrm{F}), time(\mathrm{T}). \tag{6.5}$$

Le groupe suivant de règles prend en charge les changements qui ont lieu au niveau des fluents. (6.6) indique que l'évènement $e \in \mathrm{E}^{\chi}(t)$ a initié la véracité du fluent f, noté `initiated(E,F,T)`, si les effets `Effect`$= eff(e)$ sont appliqués, $f \in eff(e)$ et $f \notin \mathrm{S}^{\chi}(t)$. (6.7) indique que l'évènement $e \in \mathrm{E}^{\chi}(t)$ a terminé la véracité du fluent f, noté `terminated(E,F,T)`, si les effets `Effect`$= eff(e)$ sont appliqués, $\neg f \in eff(e)$ et $f \in \mathrm{S}^{\chi}(t)$. Un des roles du prédicat `apply(E,Effect,T)` est de s'assurer que $e \in \mathrm{E}^{\chi}(t)$, ce qui implique plusieurs choses. Par exemple, dans la mesure où d'après la définition 6.2 l'ensemble des relations étiquetées de transition entre états $(\mathrm{S},\mathrm{E},\mathrm{S}')$ dans τ sont valides, cela implique de vérifier que $\mathrm{S}^{\chi}(t) \models pre(e)$ et $\nexists e' \in \mathrm{E}^{\chi}(t), e' \succ_{\mathbb{E}} e$. Les règles en ASP correspondantes sont retranscrites dans l'annexe B. Les implications de $e \in \mathrm{E}^{\chi}(t)$ comme celles-ci sont gérées par le prédicat `happens(E,GD,T)` qui se retrouve dans les règles (6.8) et (6.9).

$$initiated(\mathrm{E},\mathrm{F},\mathrm{T}) :- apply(\mathrm{E},\mathrm{E}ffect,\mathrm{T}), in(\mathrm{E}ffect,\mathrm{F}), not\ holds(\mathrm{F},\mathrm{T}), fluent(\mathrm{F}). \tag{6.6}$$
$$terminated(\mathrm{E},\mathrm{F},\mathrm{T}) :- apply(\mathrm{E},\mathrm{E}ffect,\mathrm{T}), in(\mathrm{E}ffect,neg(\mathrm{F})), holds(\mathrm{F},\mathrm{T}), fluent(\mathrm{F}). \tag{6.7}$$

Ensemble, les prédicats `initiated(E,F,T)` et `terminated(E,F,T)` occupent dans $\pi_{\mathbb{A}}$ le même rôle que $actual\mathrm{E}ff$ dans $\mathscr{S}_c$. En effet, les règles les déterminant vérifient pour un $e \in \mathrm{E}^{\chi}(t)$ et un $\mathrm{S}^{\chi}(t)$ donnés que les deux conditions de la définition 6.3 sont satisfaites, à savoir (i) $l_i \in eff(e)$ et (ii) $l_i \notin \mathrm{S}^{\chi}(t)$. Par conséquent, `initiated(E,F,T)` et `terminated(E,F,T)` ensemble traduisent en ASP $actual\mathrm{E}ff(e, \mathrm{S}^{\chi}(t))$, où $e \in \mathrm{E}^{\chi}(t)$.

$$apply(\mathrm{A}, \mathrm{E}ffect, \mathrm{T}) :- happens(\mathrm{A}, \mathrm{GD}, \mathrm{T}), action(\mathrm{A}, \mathrm{GD}, \mathrm{E}ffect). \tag{6.8}$$

$$apply(\mathrm{U}, \mathrm{E}ffect, \mathrm{T}) :- happens(\mathrm{U}, \mathrm{GD}, \mathrm{T}), auto(\mathrm{U}, \mathrm{GD}, \mathrm{E}ffect). \tag{6.9}$$

6.3.2 Le programme $\pi_{\mathbb{C}}$

Dans le programme $\pi_{\mathbb{C}}$, nous représentons les occurrences d'évènements $(e,t) \in \mathbb{E} \times \mathbb{T}$ par le prédicat `o(e,t)` et les formules vraies $(\psi,t) \in \mathbb{F} \times \mathbb{T}$ par le prédicat `h(ψ,t)`.

6.3.2.1 Fluent à Fluent

Le premier groupe de règles détermine les relations de base qu'il peut y avoir entre les éléments de $\mathbb{F}$. (6.10) indique qu'il y a une relation d'inertie entre `h(l,t)` et `h(l,t+1)` si $l \in Lit_{\mathbb{F}}$, $l \in \mathrm{S}^{\chi}(t)$ et $l \in \mathrm{S}^{\chi}(t+1)$. Cette règle est nécessaire étant donné notre hypothèse selon laquelle la loi d'inertie s'applique à tous les fluents.

$$inertia(h(\mathrm{L}, \mathrm{T}), h(\mathrm{L}, \mathrm{T}+1)) :- holds(\mathrm{L}, \mathrm{T}), holds(\mathrm{L}, \mathrm{T}+1), literal(\mathrm{L}). \tag{6.10}$$

(6.11) indique qu'il y a une relation entre `h(ψ,t)` et `h(ψ∧,t)` si $\psi_{\wedge}$ est une conjonction, d'où le choix de notation, $\psi_{\wedge} \in \mathrm{S}^{\chi}(t)$ et $\psi \in \psi_{\wedge}$. Notez l'utilisation de la variable `GD_L` pour représenter ψ, Ce choix s'explique par la façon dont nous avons choisi de construire les formules de $\mathscr{F}$ dans $\pi_{con}(\kappa_c)$. (6.12) indique qu'il y a une relation entre `h(ψ,t)` et `h(ψ∨,t)` si $\psi_{\vee}$ est une disjonction, d'où le choix de notation, $\psi_{\vee} \in \mathrm{S}^{\chi}(t)$, $\psi \in \psi_{\vee}$ et $\psi \in \mathrm{S}^{\chi}(t)$.

$$r\_hh(h(\mathrm{GD\_L}, \mathrm{T}), h(\mathrm{C}, \mathrm{T})) :- conj(\mathrm{C}), holds(\mathrm{C}, \mathrm{T}), in(\mathrm{C}, \mathrm{GD\_L}). \tag{6.11}$$

$$r\_hh(h(\mathrm{GD\_L}, \mathrm{T}), h(\mathrm{D}, \mathrm{T})) :- disj(\mathrm{D}), holds(\mathrm{D}, \mathrm{T}), in(\mathrm{D}, \mathrm{GD\_L}), holds(\mathrm{GD\_L}, \mathrm{T}). \tag{6.12}$$

6.3.2.2 NESS-causes directes

Le groupe suivant de règles détermine les NESS-causes directes comme définies dans la définition 6.10. (6.13) indique que `o(ini`$_l$`,-1)` est une NESS-cause directe de `h(l,0)` si $l \in \mathrm{S}^{\chi}(0)$. De cette manière, nous sommes fidèles à l'idée philosophique qu'une chaîne causale peut être remontée indéfiniment en faisant apparaître les évènements qui sont au-delà de notre passé borné dans les relations causales.

$$direct\_ness(o(ini(\mathrm{L}), -1), h(\mathrm{L}, 0)) :- initially(\mathrm{L}). \tag{6.13}$$

(6.14) indique que `o(e,t)` est une NESS-cause directe de `h(f,t+1)` si est vérifiée la condition $f \in actual\mathrm{E}ff(e, \mathrm{S}^{\chi}(t))$, avec $e \in \mathrm{E}^{\chi}(t)$. De façon similaire, (6.15) indique que `o(e,t)` est une NESS-cause directe de `h(¬f,t+1)` si $\neg f \in actual\mathrm{E}ff(e, \mathrm{S}^{\chi}(t))$ avec $e \in \mathrm{E}^{\chi}(t)$. Ces règles formalisent le fait que $actual\mathrm{E}ff$ donne les informations causales basiques contenues dans $\mathscr{S}_c$.

$$direct\_ness(o(\mathrm{E}, \mathrm{T}), h(\mathrm{F}, \mathrm{T}+1)) :- initiated(\mathrm{E}, \mathrm{F}, \mathrm{T}). \tag{6.14}$$

$$direct\_ness(o(\mathrm{E}, \mathrm{T}), h(neg(\mathrm{F}), \mathrm{T}+1)) :- terminated(\mathrm{E}, \mathrm{F}, \mathrm{T}). \tag{6.15}$$

(6.16) indique que l'occurrence d'évènement notée `Event` est une NESS-cause directe de `h(l,t+1)` si `Event` est une NESS-cause directe de `h(l,t)` et qu'il y a une relation d'inertie entre `h(l,t)` et `h(l,t+1)`. Cette règle reflète la transitivité par inertie : si (e,t) est une NESS-cause directe de (l,t), elle sera une NESS-cause directe de la véracité de l à tous les temps après t jusqu'à ce que l soit rendu faux. D'une certaine façon, nous retrouvons là la condition 2b de la définition 6.10. L'utilisation de la variable `Event` rend la relation de transitivité plus claire car cela met l'accent sur le fait que contrairement aux règles précédentes, le temps de la cause et celui de la conséquence n'ont pas de contrainte spécifique, à part que la cause précède la conséquence.

$$direct\_ness(\mathrm{E}vent, h(\mathrm{L},\mathrm{T}+1)) :- direct\_ness(\mathrm{E}vent, h(\mathrm{L},\mathrm{T})), \\ inertia(h(\mathrm{L},\mathrm{T}), h(\mathrm{L},\mathrm{T}+1)). \tag{6.16}$$

Finalement, (6.17) indique que l'occurrence d'évènement notée `Event` est une NESS-cause directe de `h(ψ,t)` si `Event` est une NESS-cause directe de `h(ψ',t)` et qu'il existe une relation entre `h(ψ',t)` et `h(ψ,t)`. Cette règle reflète également la transitivité, mais cette fois-ci par rapport aux opérateurs logiques et au même temps.

$$direct\_ness(\mathrm{E}vent, h(\mathrm{GD},\mathrm{T})) :- direct\_ness(\mathrm{E}vent, h(\mathrm{GD\_L},\mathrm{T})), \\ r\_hh(h(\mathrm{GD\_L},\mathrm{T}), h(\mathrm{GD},\mathrm{T})). \tag{6.17}$$

Notez que, bien qu'il y ait une transitivité par l'inertie pour les littéraux, celle-ci n'existe pas pour les formules plus complexes. Si nous prenons l'exemple 6.2 et que nous le modifions légèrement, nous obtenons un exemple montrant que cette transitivité n'est pas désirable. Imaginons $ini_{l_4} \in \mathrm{E}^{\chi}(-1)$, nous aurions alors ψ vraie à tous les temps. Si la transitivité par inertie était vraie pour les formules comme ψ, notre programme considérerait $(ini_{l_1}, -1)$ non seulement comme une NESS-cause directe de $(\psi, 0)$, ce qui serait vrai, mais comme une NESS-cause directe de $(\psi, 2)$ également. Dans la section 6.2.1 nous avons discuté du fait que cela n'était factuellement pas nécessairement vrai et donc en contradiction avec notre conception de la causalité effective.

Définition 6.13 [Programme de causalité effective $\Pi(\chi)$]**.** *Étant donné un scénario* $\pi_{sce}(\sigma)$*, un contexte* $\pi_{con}(\kappa_c)$*, un langage de description d'action* $\pi_{\mathbb{A}}$ *et une définition de causalité effective* $\pi_{\mathbb{C}}$*, le programme de causalité effective est* $\Pi(\chi) = \pi_{sce}(\sigma) \cup \pi_{con}(\kappa_c) \cup \pi_{\mathbb{A}} \cup \pi_{\mathbb{C}}$.

Proposition 6.8 [Complétude si ψ un littéral]**.** *Étant donné une occurrence* $(e,t) \in \mathbb{E} \times \mathbb{T}$ *étant une NESS-cause directe de la formule vraie* (l, t_l)*,* $\Pi(\chi) \models$ `direct_ness(o(e,t),h(l,t_l))` [3].

Démonstration. Étant donné (e,t) une NESS-cause directe de (l, t_l), nous savons d'après la proposition 6.3 que : (c_1) $\mathrm{S}^{\chi}(t) \rhd \{e\} \models l$, (c_2) $\forall t',\ t < t' \leq t_l, \mathrm{S}^{\chi}(t') \models l$ et (c_3) $\mathrm{S}^{\chi}(t) \not\models l$. Trois cas peuvent alors être considérés.

Cas 1 : Dans le cas où $t_l = 0$, la seule façon par laquelle $\mathrm{S}^{\chi}(0) \models l$ est que $l \in \mathrm{S}_0$. De ce fait, par construction, $\Pi(\chi) \models initially(l)$. Donc, étant donné la règle :

$$(6.13) : direct\_ness(o(ini(\mathrm{L}), -1), h(\mathrm{L}, 0)) \ :- initially(\mathrm{L}).$$

3. Notez que le dernier $\models$ correspond à l'implication au sens des programmes logiques, ce qui doit être distingué de celui utilisée dans les STEE.

nous avons $\Pi(\chi) \models direct\_ness(o(ini(l),-1),h(l,0))$ où $ini(l) = e$ et $t = -1$.

Cas 2 : Dans le cas où $t_l = t+1$, étant donné que $e \in \mathrm{E}^{\chi}(t)$, (c_1) et (c_3), nous avons soit $\Pi(\chi) \models initiated(e,l,t)$, soit $\Pi(\chi) \models terminated(e,\overline{l},t)$. De ce fait, étant donné les règles :

$$(6.14) : direct\_ness(o(\mathrm{E},\mathrm{T}),h(\mathrm{L},\mathrm{T}+1)) \ :- \ initiated(\mathrm{E},\mathrm{L},\mathrm{T}).$$

$$(6.15) : direct\_ness(o(\mathrm{E},\mathrm{T}),h(neg(\mathrm{L}),\mathrm{T}+1)) \ :- \ terminated(\mathrm{E},\mathrm{L},\mathrm{T}).$$

nous avons $\Pi(\chi) \models direct\_ness(o(e,t),h(l,t+1))$.

Cas 3 : Dans tous les autres cas, nous avons $t_l > t+1$. En effet, $t_l \neq 0$ et $t_l \neq t+1$ étant donné que $t_l > t$ puisque nous assumons que les causes précèdent toujours temporellement leurs conséquences. Comme précédemment, $\Pi(\chi) \models direct\_ness(o(e,t),h(l,t+1))$. De plus, comme $\forall t',\ t < t' \leq t_l,\ \mathrm{S}^{\chi}(t') \models l$, nous savons que $\Pi(\chi) \models holds(l,t')$, où $t < t' \leq t_l$. Par conséquent, étant donné la règle :

$$(6.10) : inertia(h(\mathrm{L},\mathrm{T}),h(\mathrm{L},\mathrm{T}+1)) :- \ holds(\mathrm{L},\mathrm{T}),holds(\mathrm{L},\mathrm{T}+1),literal(\mathrm{L}).$$

nous avons donc la chaîne complète depuis $\Pi(\chi) \models inertia(h(l,t+1),h(l,t+2))$ jusqu'à $\Pi(\chi) \models inertia(h(l,t_l-1),h(l,t_l))$. Avec $\Pi(\chi) \models direct\_ness(o(e,t),h(l,t+1))$ et l'existence de la règle :

$$(6.16) : direct\_ness(\mathrm{E}vent,h(\mathrm{L},\mathrm{T}+1)) \ :- \ direct\_ness(\mathrm{E}vent,h(\mathrm{L},\mathrm{T})), \\ inertia(h(\mathrm{L},\mathrm{T}),h(\mathrm{L},\mathrm{T}+1)).$$

nous avons la chaîne complète depuis $\Pi(\chi) \models direct\_ness(o(e,t),h(l,t+2))$ jusqu'au résultat recherché : $\Pi(\chi) \models direct\_ness(o(e,t),h(l,t_l))$. □

Proposition 6.9 [Complétude si ψ une conjonction de littéraux]**.** *Étant donné une occurrence d'évènement* $(e,t) \in \mathbb{E} \times \mathbb{T}$ *une NESS-cause directe de la formule vraie* (ψ, t_ψ) *où la formule* $\psi = l_1 \wedge \cdots \wedge l_m$, $\Pi(\chi) \models \texttt{direct\_ness(o(e,t),h(}\psi\texttt{,}t_\psi\texttt{))}$.

Démonstration. Étant donné (e,t) une NESS-cause directe de (ψ, t_ψ), nous savons d'après la proposition 6.4 que : (c_1) $\exists j \in \{1,\ldots,m\}$, $\mathrm{S}^{\chi}(t) \rhd \{e\} \models l_j$, (c_2) $\forall t',\ t < t' \leq t_\psi$, $\mathrm{S}^{\chi}(t') \models l_j$ et (c_3) $\mathrm{S}^{\chi}(t) \not\models l_j$.

La conjonction de littéraux $\psi = l_1 \wedge \cdots \wedge l_m$ étant vraie à t_ψ, nous avons $\Pi(\chi) \models conj(\psi)$, $\Pi(\chi) \models in(\psi,l_1)$, ..., $\Pi(\chi) \models in(\psi,l_m)$ et $\Pi(\chi) \models holds(\psi,t_\psi)$. De ce fait, étant donnée la règle :

$$(6.11) : r\_hh(h(\mathrm{L},\mathrm{T}),h(\mathrm{C},\mathrm{T})) \ :- \ conj(\mathrm{C}),holds(\mathrm{C},\mathrm{T}),in(\mathrm{C},\mathrm{L}).$$ [4]

nous avons $\Pi(\chi) \models r\_hh(h(l_1,t_\psi),h(\psi,t_\psi)),\ldots$ et $\Pi(\chi) \models r\_hh(h(l_m,t_\psi),h(\psi,t_\psi))$. Étant donné les conditions (c_1), (c_2) et (c_3), d'après les propositions 6.3 et 6.8, (e,t) est une NESS-cause directe de (l_j,t_ψ) et donc $\Pi(\chi) \models direct\_ness(o(e,t),h(l_j,t_\psi))$. Par conséquent, étant donné la règle :

$$(6.17) : direct\_ness(\mathrm{E}vent,h(\mathrm{GD},\mathrm{T})) \ :- \ direct\_ness(\mathrm{E}vent,h(\mathrm{L},\mathrm{T})), \\ r\_hh(h(\mathrm{L},\mathrm{T}),h(\mathrm{GD},\mathrm{T})).$$

nous avons $\Pi(\chi) \models direct\_ness(o(e,t),h(\psi,t_\psi))$. □

4. Par souci de clarté, puisque nous savons que ψ est une conjonction de littéraux, nous remplaçons la variable GD_L par L.

Proposition 6.10 [Complétude si ψ une FND]**.** *Étant donné une occurrence* $(e,t) \in \mathbb{E} \times \mathbb{T}$ *une NESS-cause directe de la formule vraie* (ψ, t_ψ) *où* $\psi = \psi_1 \vee \cdots \vee \psi_m$ *est minimale par subsomption et libre de tautologies,* $\Pi(\chi) \vDash \texttt{direct\_ness(o(e,t),h(}\psi\texttt{,}t_\psi\texttt{))}$.

Démonstration. Étant donné (e,t) une NESS-cause directe de (ψ, t_ψ), nous savons d'après la proposition 6.5 : (c_1) $\exists j \in \{1,\ldots,m\}$, $S^\chi(t_\psi) \vDash \psi_j$, (c_2) $S^\chi(t) \rhd \{e\} \vDash l \in \psi_j$, (c_3) $\forall t'$, $t < t' \leq t_\psi$, $S^\chi(t') \vDash l$ et (c_4) $S^\chi(t) \nvDash l$.

La formule sous forme normale disjonctive $\psi = \psi_1 \vee \cdots \vee \psi_m$ étant vraie à t_ψ, nous avons $\Pi(\chi) \vDash disj(\psi)$, $\Pi(\chi) \vDash in(\psi, \psi_1)$, ..., $\Pi(\chi) \vDash in(\psi, \psi_m)$ et $\Pi(\chi) \vDash holds(\psi, t_\psi)$. Par la condition (c_1), nous avons également $\Pi(\chi) \vDash holds(\psi_j, t_\psi)$. De ce fait, étant donné la règle :

$$(6.12) : r\_hh(h(\text{GD\_L},\text{T}), h(\text{D},\text{T})) \ :- \ disj(\text{D}), holds(\text{D},\text{T}), in(\text{D},\text{GD\_L}), holds(\text{GD\_L},\text{T}).$$

nous avons $\Pi(\chi) \vDash r\_hh(h(\psi_j, t_\psi), h(\psi, t_\psi))$. Étant donné les condition (c_2), (c_3) et (c_4), selon les propositions 6.4 et 6.9, (e,t) est une NESS-cause directe de (ψ_j, t_ψ) et donc nous avons $\Pi(\chi) \vDash direct\_ness(o(e,t), h(\psi_j, t_\psi))$. Par conséquent, étant donné la règle :

$$(6.17) : direct\_ness(\text{E}vent, h(\text{GD},\text{T})) \ :- \ direct\_ness(\text{E}vent, h(\text{L},\text{T})), \\ r\_hh(h(\text{L},\text{T}), h(\text{GD},\text{T})).$$

nous avons $\Pi(\chi) \vDash direct\_ness(o(e,t), h(\psi, t_\psi))$. □

Proposition 6.11 [Correction si ψ un littéral]**.** *Étant donné une occurrence* $(e,t) \in \mathbb{E} \times \mathbb{T}$ *et* $\Pi(\chi) \vDash \texttt{direct\_ness(o(e,t),h(}l\texttt{,}t_l\texttt{))}$, (e,t) *est une NESS-cause directe de la formule vraie* (l, t_l).

Démonstration. Nous considérons l'occurrence d'évènement (e,t) telle que $\Pi(\chi) \vDash \rho$ avec $\rho = direct\_ness(o(e,t), h(l,t_l))$. D'après la proposition 6.7, étant donné que $\exists S \in AS(\Pi(\chi))$ tel que $\rho \in S$, alors $\exists r \in \Pi(\chi)$ tel que $head(r) = \rho$, $body^+(r) \subseteq S$ et $body^-(r) \cap S = \varnothing$.

Nous notons $t_l = t + n$, avec $n \in \mathbb{N}^*$ de par le fait que nous modélisons le temps de façon discrète et que le cas $n \leq 0$ contredirait que $t_l > t$. Soit $P(n)$ la proposition suivante : étant donné $(e,t) \in \mathbb{E} \times \mathbb{T}$ et $\Pi(\chi) \vDash direct\_ness(o(e,t), h(l, t+n))$, (e,t) est une NESS-cause directe de $(l, t+n)$. Nous faisons une preuve par récursion sur n.

Cas base : Dans le cas où $n = 1$, deux cas sont possibles : soit $t = -1$, soit $t \geq 0$. Nous considérons chacun de ces cas séparément.

Cas 1 : lorsque $t = -1$, $(l, t+n)$ est $(l, 0)$, et donc $\rho = direct\_ness(o(e,-1), h(l,0))$. La seule règle r qui peut avoir été utilisée pour obtenir ρ est :

$$(6.13) : direct\_ness(o(ini(l), -1), h(l,0)) \ :- \ initially(l).$$

De ce fait, nous avons nécessairement $e = ini_l$, et comme $body^+(r) \subseteq S$, nous avons également $initially(l) \in S$, donc $l \in S_0$. De ce fait, par la sémantique de $\mathscr{S}_c$, $ini_l \in E^\chi(-1)$ et $S^\chi(-1) \rhd \{ini_l\} \vDash l$. De plus, comme $t = -1$ et $t + n = 0$, la condition $\forall t'$, $-1 < t' \leq 0$, $S^\chi(t') \vDash l$ est satisfaite état donné que $S^\chi(0) \vDash l$. Pour finir, la condition $S^\chi(-1) \nvDash l$ est satisfaite par le fait que $S^\chi(-1) = Lit_\mathbb{F} \setminus S^\chi(0)$. Par conséquent, d'après la proposition 6.3, $(ini_l, -1)$ est une NESS-cause directe de $(l,0)$.

Cas 2 : $(l, t+n)$ est soit $(f, t+1)$ ou $(\neg f, t+1)$, donc $\rho = direct\_ness(o(e,t), h(f, t+1))$, ou $\rho = direct\_ness(o(e,t), h(neg(f), t+1))$. Dans chacun de ces cas, une unique règle r peut avoir été utilisée pour obtenir ρ. Respectivement, ces règles sont :

$$(6.14) : direct\_ness(o(e,t), h(f, t+1)) \ :- \ initiated(e, f, t).$$

$$(6.15) : direct\_ness(o(e,t), h(neg(f), t+1)) \ :- \ terminated(e,f,t).$$

Comme $body^+(r) \subseteq S$, nous avons soit $initiated(e,f,t) \in S$, soit $terminated(e,f,t) \in S$, et donc dans tous les cas $S^\chi(t) \rhd \{e\} \models l$ et $S^\chi(t) \not\models l$. De plus, comme $t_l = t+1$, la condition $\forall t',\ t < t' \leq t+1,\ S^\chi(t') \models l$ est satisfaite du fait que $S^\chi(t) \rhd \{e\} \models l$ et $e \in E^\chi(t)$. Par conséquent, d'après la proposition 6.3, (e,t) est une NESS-cause directe de $(l, t+1)$. P(1) est donc vraie pour les deux cas possibles.

Cas récursif : Nous montrons maintenant que pour tout $k \geq 1$, si notre hypothèse de récursion P(k) est vraie, alors la proposition P($k+1$) l'est également. La proposition P(k) est : étant donné l'occurrence $(e,t) \in \mathbb{E} \times \mathbb{T}$ et $\Pi(\chi) \models direct\_ness(o(e,t), h(l, t+k))$, (e,t) est une NESS-cause directe de $(l, t+k)$. La proposition P($k+1$) est alors : étant donné l'occurrence $(e,t) \in \mathbb{E} \times \mathbb{T}$ et $\Pi(\chi) \models direct\_ness(o(e,t), h(l, t+k+1))$, (e,t) est une NESS-cause directe de $(l, t+k+1)$.

Deux règles r peuvent être utilisées pour obtenir $\rho = direct\_ness(o(e,t), h(l, t+k+1))$ avec $k = n \in \mathbb{N}^*$.

Cas 1 :

$$(6.17) : direct\_ness(o(e,t), h(l, t+k+1)) \ :- \ direct\_ness(o(e,t), h(l', t+k+1)), \\ r\_hh(h(l', t+k+1), h(l, t+k+1)).$$

Dans ce cas, comme $body^+(r) \subseteq S$, nous aurions $r\_hh(h(l', t+k+1), h(l, t+k+1))$. Cela est impossible du fait que $r\_hh(h(l', t+k+1), h(l, t+k+1))$ peut uniquement être obtenu par les règles (6.11) et (6.12) qui demandent que l soit une conjonction, $conj(l)$, ou une disjonction $disj(l)$.

Cas 2 :

$$(6.16) : direct\_ness(o(e,t), h(l, t+k+1)) \ :- \ direct\_ness(o(e,t), h(l, t+k)), \\ inertia(h(l, t+k), h(l, t+k+1)).$$

Dans ce cas, comme $body^+(r) \subseteq S$, nous avons $direct\_ness(o(e,t), h(l, t+k)) \in S$, ainsi que $inertia(h(l, t+k), h(l, t+k+1)) \in S$. De ce fait, d'après notre hypothèse de récursion, l'occurrence (e,t) est une NESS-cause directe de $(l, t+k)$. D'après la proposition 6.3, cela veut dire que $S^\chi(t) \rhd \{e\} \models l$, mais aussi que $\forall t',\ t < t' \leq t+k$, $S^\chi(t') \models l$ et que $S^\chi(t) \not\models l$. De plus, comme $inertia(h(l, t+k), h(l, t+k+1)) \in S$ et que nous avons la règle :

$$(6.10) : inertia(h(l, t+k), h(l, t+k+1)) \ :- \ holds(l, t+k), holds(l, t+k+1), literal(l).$$

nous pouvons déduire que $S^\chi(t+k+1) \models l$ et donc $\forall t',\ t < t' \leq t+k+1, S^\chi(t') \models l$. D'après la proposition 6.3 cela nous permeet de dire que (e,t) est une NESS-cause directe de $(l, t+k+1)$.

Conclusion : Étant donné que le cas base et le cas récursif sont prouvés vrais, par récursion mathématique nous savons que la proposition P(n) est vraie pour tous les $n \in \mathbb{N}^*$. □

Proposition 6.12 [Correction si ψ une conjonction de littéraux]**.** *Étant donné une occurrence d'évènement* $(e,t) \in \mathbb{E} \times \mathbb{T}$ *et* $\Pi(\chi) \models direct\_ness(o(e,t), h(\psi, t_\psi))$ *où* $\psi = l_1 \wedge \cdots \wedge l_m$, (e,t) *est une NESS-cause directe de la formule vraie* (ψ, t_ψ).

Démonstration. La seule règle permettant d'obtenir $\rho = direct\_ness(o(e,t), h(\psi, t_\psi))$ étant donné que $\psi = l_1 \wedge \cdots \wedge l_m$ est :

$$(6.17) : direct\_ness(o(e,t), h(\psi, t_\psi)) \ :- \ direct\_ness(o(e,t), h(l_j, t_\psi)), \\ r\_hh(h(l_j, t_\psi), h(\psi, t_\psi)).$$

Dans ce cas, comme $body^+(r) \subseteq S$, nous avons $direct\_ness(o(e,t), h(l_j, t_\psi)) \in S$, ainsi que $r\_hh(h(l_j,t_\psi),\ h(\psi,t_\psi)) \in S$. De ce fait, d'après les propositions 6.11 et 6.3, l'occurrence (e,t) est une NESS-cause directe de (l_j, t_ψ), ce qui veut dire que $S^\chi(t) \rhd \{e\} \models l_j$, mais aussi que $\forall t',\ t < t' \leq t_\psi$, $S^\chi(t') \models l_j$ et que $S^\chi(t) \not\models l_j$. Comme $r\_hh(h(l_j,t_\psi), h(\psi,t_\psi)) \in S$ et étant donné la règle :

$$(6.11) : r\_hh(h(l_j,t_\psi), h(\psi,t_\psi)) \ :- \ conj(\psi), holds(\psi,t_\psi), in(\psi,l_j).$$

nous avons $conj(\psi) \in S$, $holds(\psi,t_\psi) \in S$ et $in(\psi,l_j) \in S$. Par conséquent, $j \in \{1,\ldots,m\}$ et donc, d'après la proposition 6.4, (e,t) est une NESS-cause directe de (ψ, t_ψ). □

Proposition 6.13 [Correction si ψ une FND]**.** *Étant donné une occurrence* $(e,t) \in \mathbb{E} \times \mathbb{T}$ *et* $\Pi(\chi) \models direct\_ness(o(e,t), h(\psi, t_\psi))$ *où* $\psi = \psi_1 \vee \cdots \vee \psi_m$ *est minimale par subsomption et libre de tautologies,* (e,t) *est une NESS-cause directe de la formule vraie* (ψ, t_ψ).

Démonstration. La seule règle permettant d'obtenir $\rho = direct\_ness(o(e,t), h(\psi,t_\psi))$ étant donné que $\psi = \psi_1 \vee \cdots \vee \psi_m$ est :

$$(6.17) : direct\_ness(o(e,t), h(\psi,t_\psi)) \ :- \ direct\_ness(o(e,t), h(\psi_j,t_\psi)), \\ r\_hh(h(\psi_j,t_\psi), h(\psi,t_\psi)).$$

Dans ce cas, comme $body^+(r) \subseteq S$, nous avons $direct\_ness(o(e,t), h(\psi_j,t_\psi)) \in S$, ainsi que $r\_hh(h(\psi_j,t_\psi),\ h(\psi,t_\psi)) \in S$. De ce fait, d'après les propositions 6.12 et 6.4, l'occurrence (e,t) est une NESS-cause directe de (ψ_j, t_ψ) . Étant donné que $\psi_j = l_1 \wedge \cdots \wedge l_n$, nous savons que $\exists i \in \{1,\ldots,n\}$, $S^\chi(t) \rhd \{e\} \models l_i$, mais aussi que $\forall t',\ t < t' \leq t_\psi$, $S^\chi(t') \models l_i$ et que $S^\chi(t) \not\models l_i$. Comme $r\_hh(h(\psi_j,t_\psi), h(\psi,t_\psi)) \in S$ et étant donné la règle :

$$(6.12) : r\_hh(h(\psi_j,t_\psi), h(\psi,t_\psi)) \ :- \ disj(\psi), holds(\psi,t_\psi), in(\psi,\psi_j), holds(\psi_j,t_\psi).$$

nous avons $disj(\psi) \in S$, $holds(\psi,t_\psi) \in S$, $in(\psi,\psi_j) \in S$ et $holds(\psi_j,t_\psi) \in S$. Par conséquent, $j \in \{1,\ldots,m\}$ et $S^\chi(t_\psi) \models \psi_j$, ce qui veut dire, d'après la proposition 6.5, que (e,t) est une NESS-cause directe de (ψ, t_ψ). □

Théorème 6.2 [Complétude et correction des NESS-causes directes]**.** *Étant donné un cadre causal* χ *et une formule sous FND, minimale par subsomption et libre de tautologies* $\psi \in \mathscr{F}$*, l'occurrence d'évènements* $(e,t) \in \mathbb{E} \times \mathbb{T}$ *est une NESS-cause directe de la formule vraie* (ψ, t_ψ) *ssi* $\Pi(\chi) \models direct\_ness(o(e,t), h(\psi, t_\psi))$.

Démonstration. Le théorème 6.2 découle des propositions 6.8 à 6.13. □

6.3.2.3 NESS-causes

Le groupe suivant de règles détermine les NESS-causes comme définies dans la définition 6.11. (6.18) indique que $o(e_1, t_1)$ est une NESS-cause de $h(\psi, t_\psi)$ si $o(e_1, t_1)$ est une NESS-cause directe de $h(\psi, t_\psi)$. Il s'agit ici du cas base de la définition 6.11.

$$ness(o(\text{E1},\text{T1}), h(\text{GD\_L},\text{T2})) :- direct\_ness(o(\text{E1},\text{T1}), h(\text{GD\_L},\text{T2})). \tag{6.18}$$

(6.19) indique que $o(e_1, t_1)$ est une NESS-cause de $h(\psi, t_\psi)$ si $o(e_1, t_1)$ est une cause effective de $o(e_2, t_2)$ et $o(e_2, t_2)$ est une NESS-cause de $h(\psi, t_\psi)$. Cela correspond au cas

récursif de la définition 6.11 où nous nous posons la question de quelles ont été les causes du déclenchement des causes déjà identifiées. Pour reprendre les notations de la définition 6.11, il s'agit du cas où $o(e_1, t_1)$ est une NESS-cause de $tri(C_R(t_i))$.

$$ness(o(\text{E1},\text{T1}), h(\text{GD\_L},\text{T3})) :- actual(o(\text{E1},\text{T1}), o(\text{E2},\text{T2})), \\ ness(o(\text{E2},\text{T2}), h(\text{GD\_L},\text{T3})). \tag{6.19}$$

Théorème 6.3 [Complétude et correction des NESS-causes]**.** *Étant donné un cadre causal* χ *et une formule sous FND, minimale par subsomption et libre de tautologies* $\psi \in \mathscr{F}$*, l'occurrence d'évènements* $(e,t) \in \mathbb{E} \times \mathbb{T}$ *est une NESS-cause de la formule vraie* (ψ, t_ψ) *ssi nous avons* $\Pi(\chi) \vDash$ `ness(o(e,t),h(ψ,t_ψ))`.

Démonstration. [$\Longrightarrow$] Soit (e,t) une NESS-cause de (ψ, t_ψ). D'après la définition 6.11, cela veut dire que $\exists C' \subseteq \mathbb{E} \times \mathbb{T}$ tel que $(e,t) \in C'$, où C' est un ensemble suffisant de NESS-causes de (ψ, t_ψ), et que $\exists C \subseteq \mathbb{E} \times \mathbb{T}$ tel que $C \rightarrow (\psi, t_\psi)$. Deux cas sont alors possibles : soit nous sommes dans le cas base de la définition 6.11, soit nous sommes dans le cas récursif. Nous traitons chaque cas séparément.

Cas base : Dans ce cas où $C' = C$, $(e,t) \in C$. De ce fait, d'après la définition 6.10, (e,t) est une NESS-cause directe de (ψ, t_ψ). Par conséquent, d'après le théorème 6.2, nous avons $\Pi(\chi) \vDash direct\_ness(o(e,t), h(\psi, t_\psi))$. Étant donné la règle :

$$(6.18) : ness(o(\text{E1},\text{T1}), h(\text{GD\_L},\text{T2})) :- direct\_ness(o(\text{E1},\text{T1}), h(\text{GD\_L},\text{T2})).$$

nous avons $\Pi(\chi) \vDash ness(o(e,t), h(\psi, t_\psi))$.

Cas récursif : Dans ce cas où $C' \neq C$, nous savons par la définition 6.11 qu'il existe un ensemble non vide d'occurrences d'évènements retirables, $C_R = C \setminus C'$, dont la partition, $C_R(t_1), \ldots, C_R(t_k)$, correspond avec la séquence décroissante $t_1, \ldots, t_k$. Nous savons également qu'il existe un ensemble d'occurrences d'évènements substituantes, $C_S = C' \setminus C$, décomposable en une séquence de sous ensembles, $C_{S_1}, \ldots, C_{S_k}$, qui vérifient :

- $C_S = \bigcup_{i \in \{1,\ldots,k\}} C_{S_i}$.
- $\forall i \in \{1,\ldots,k\}, C_R(t_i) = \varnothing \implies C_{S_i} = \varnothing$.
- $\forall i \in \{1,\ldots,k\}, C_R(t_i) \neq \varnothing \implies C_{S_i}$ ensemble suffisant de NESS-causes de $(tri(C_R(t_i)), t_i)$.

Nous pouvons écrire $C' = C_S \cup (C \cap C')$ et $C = C_R \cup (C \cap C')$. Le cas où $(e,t) \in (C \cap C')$ se prouve de la même façon que le cas base vu précédemment. Nous considérons alors le cas restant où $(e,t) \in C_S$. Plus précisément, étant donné que $C_S = \bigcup_{i \in \{1,\ldots,k\}} C_{S_i}$, nous sommes dans le cas où $C_R(t_i) \neq \varnothing$ et $(e,t) \in C_{S_i}$, et donc (e,t) est une NESS-causes de $(tri(C_R(t_i)), t_i)$. Par conséquent, $\exists (e', t_i) \in C_R(t_i)$ telle que (e,t) est une NESS-cause de $(tri(e'), t_i)$. De plus, comme $(e', t_i) \in C_R(t_i)$ et $C = C_R \cup (C \cap C')$, alors $(e', t_i) \in C$ ce qui veut dire par le cas base que $\Pi(\chi) \vDash ness(o(e', t_i), h(\psi, t_\psi))$. Deux cas peuvent alors être considérés.

Dans le premier cas correspondant au cas base, (e,t) est une NESS-cause directe de $(tri(e'), t_i)$, donc, d'après le théorème 6.2, $\Pi(\chi) \vDash ness(o(e,t), h(tri(e'), t_i))$. Étant donné les règles :

$$(6.18) : ness(o(\text{E1},\text{T1}), h(\text{GD\_L},\text{T3})) :- actual(o(\text{E1},\text{T1}), o(\text{E2},\text{T2})), \\ ness(o(\text{E2},\text{T2}), h(\text{GD\_L},\text{T3})).$$
$$(6.20) : actual(o(\text{E1},\text{T1}), o(\text{E2},\text{T2})) :- ness(o(\text{E1},\text{T1}), h(\text{GD},\text{T2})), \\ happens(\text{E2},\text{GD},\text{T2}), auto(\text{E2},\text{GD},Eff).$$

nous avons $\Pi(\chi) \vDash actual(o(e,t), o(e',t_i))$ et donc $\Pi(\chi) \vDash ness(o(e,t), h(\psi, t_\psi))$.

Le deuxième cas correspond au cas récursif où (e,t) est une NESS-cause de $(tri(e''), t_j)$, avec (e'', t_j) une NESS-cause directe de $(tri(e'), t_i)$. Étant donné que nous sommes dans une formalisation bornée dans le passé, l'application d'un tel raisonnement de manière récursive nous mène inévitablement au premier cas. À ce moment là, nous pouvons déduire que $\Pi(\chi) \vDash actual(o(e,t), o(e'', t_j))$. Par conséquent, nous pouvons chaîner vers l'avant en appliquant les règles (6.18) et (6.20) jusqu'à obtenir $\Pi(\chi) \vDash ness(o(e,t), h(\psi, t_\psi))$.

[$\Longleftarrow$] Nous considérons (e,t) telle que $\Pi(\chi) \vDash \rho$, avec $\rho = ness(o(e,t), h(\psi, t_\psi))$. D'après la proposition 6.7, étant donné que $\exists S \in AS(\Pi(\chi))$ tel que $\rho \in S$, alors $\exists r \in \Pi(\chi)$, $head(r) = \rho$, $body^+(r) \subseteq S$ et $body^-(r) \cap S = \varnothing$.

Nous notons $t_\psi = t + n$, avec $n \in \mathbb{N}^*$. Soit $P(n)$ la proposition suivante : étant donné un cadre causal χ, une formule sous FND, minimale par subsomption et libre de tautologies $\psi \in \mathscr{F}$ telle que $\psi \nvDash \bot$ et $\Pi(\chi) \vDash ness(o(e,t), h(\psi, t+n))$, nous avons $(e,t) \in \mathbb{E} \times \mathbb{T}$ une NESS-cause de $(\psi, t+n)$. Nous faisons une preuve par récursion sur n.

Cas base : Dans le cas où $n = 1$, $(\psi, t+n)$ est $(\psi, t+1)$, donc $\rho = ness(o(e,t), h(\psi, t+1))$. La seule règle r pouvant être utilisée dans ce cas pour obtenir ρ est :

$$(6.18) : ness(o(e,t), h(\psi, t+1)) \ :- \ direct\_ness(o(e,t), h(\psi, t+1)).$$

Cas de la règle (6.18) : comme $body^+(r) \subseteq S$, nous avons $direct\_ness(o(e,t), h(\psi, t+1)) \in S$. De ce fait, d'après le théorème 6.2, $\exists C \subseteq \mathbb{E} \times \mathbb{T}$ tel que $(e,t) \in C$ et $C \rightarrow (\psi, t+1)$. D'après le cas base de la définition 6.11, en prenant $C' = C$, l'ensemble d'occurrences d'évènements C' est un ensemble suffisant de NESS-causes de $(\psi, t+1)$. De plus, comme $(e,t) \in C$ et $C' = C$, alors $(e,t) \in C'$. Par conséquent, (e,t) est une NESS-cause de $(\psi, t+1)$. La proposition $P(1)$ est donc vraie.

Cas récursif : Nous montrons à présent que pour tout $k \geq 1$, si notre hypothèse d'induction $P(k)$ est vraie, alors la proposition $P(k+1)$ l'est également. Les règles (6.18) et (6.19) peuvent être utilisées pour obtenir $\rho = ness(o(e,t), h(\psi, t+k+1))$, étant donné $k = n \in \mathbb{N}^*$.

Cas de la règle (6.18) :

$$(6.18) : ness(o(e,t), h(\psi, t+k+1)) \ :- \ direct\_ness(o(e,t), h(\psi, t+k+1)).$$

comme $body^+(r) \subseteq S$, nous avons $direct\_ness(o(e,t), h(\psi, t+k+1)) \in S$. De ce fait, d'après le théorème 6.2, $\exists C \subseteq \mathbb{E} \times \mathbb{T}$ tel que $(e,t) \in C$ et $C \rightarrow (\psi, t+k+1)$. D'après le cas de base de la définition 6.11, en prenant $C' = C$, l'ensemble d'occurrences d'évènements C' est un ensemble suffisant de NESS-causes de $(\psi, t+k+1)$. De plus, comme $(e,t) \in C$ et $C' = C$, alors $(e,t) \in C'$. Par conséquent, (e,t) est une NESS-cause de $(\psi, t+k+1)$.

Cas de la règle (6.19) :

$$(6.19) : ness(o(e,t), h(\psi, t+k+1)) \ :- \ actual(o(e,t), o(e',t')), ness(o(e',t'), h(\psi, t+k+1)).$$

comme $body^+(r) \subseteq S$, alors $ness(o(e',t'), h(\psi, t+k+1)) \in S$ et $actual(o(e,t), o(e',t')) \in S$. De par notre hypothèse générale que les causes précédent leurs conséquences temporellement, nous savons que $t < t' < t+k+1$. La différence entre t' et $t+k+1$ est d'au plus $t+k$. De ce fait, d'après notre hypothèse de récursion $P(k)$, (e',t') est une NESS-cause de $(\psi, t+k+1))$. La seule règle r pouvant être utilisée pour obtenir $actual(o(e,t), o(e',t'))$ est :

$$(6.20) : actual(o(e,t), o(e',t')) \ :- \ ness(o(e,t), h(\psi', t')), happens(e', \psi', t'), auto(e', \psi', \phi').$$

Comme $body^+(r) \subseteq S$, nous avons $ness(o(e,t), h(\psi', t')) \in S$, ainsi que $happens(e', \psi', t') \in S$ et $auto(e', \psi', \phi') \in S$. Comme précédemment, nous pouvons déduire que (e,t) est une NESS-cause de (ψ', t'), où $e' \in \mathbb{N}$, $tri(e') = \psi'$ et $e' \in E^{\chi}(t')$. Par conséquent, d'après le cas récursif de la définition 6.11, (e,t) est une NESS-cause de $(\psi, t+k+1)$.

Conclusion : Étant donné que le cas base et le cas récursif sont prouvés vrais, par récursion mathématique nous savons que la proposition $P(n)$ est vraie pour tous les $n \in \mathbb{N}^*$. □

6.3.2.4 Causes effectives

Le groupe suivant de règles détermine les causes effectives comme définies dans la définition 6.12. (6.20) indique que $o(e_1, t_1)$ est une cause effective de $o(e_2, t_2)$ si $o(e_1, t_1)$ est une NESS-cause de $h(\psi, t_2)$ où $\psi = tri(e_2)$ et $e_2 \in \mathbb{N}$.

$$actual(o(\text{E1},\text{T1}), o(\text{E2},\text{T2})) :- ness(o(\text{E1},\text{T1}), h(\text{GD},\text{T2})), happens(\text{E2},\text{GD},\text{T2}), auto(\text{E2},\text{GD},Eff). \quad (6.20)$$

Théorème 6.4 [Complétude et correction des causes effectives]**.** *Étant donné un cadre causal χ, une formule sous FND, minimale par subsomption et libre de tautologies $\psi \in \mathcal{F}$ où $\psi = tri(e')$, l'occurrence d'évènements $(e,t) \in \mathbb{E} \times \mathbb{T}$ est une cause effective de l'occurrence d'évènements $(e',t') \in \mathbb{E} \times \mathbb{T}$ ssi $\Pi(\chi) \models$ `actual(o(e,t),o(e',t'))`.*

Démonstration. [$\Longrightarrow$] Soit l'occurrence d'évènement (e,t) une cause effective de (e',t'). Par la définition 6.12, nous savons que (e,t) est alors une NESS-cause de $(tri(e'), t')$. De ce fait, d'après le théorème 6.3, $\Pi(\chi) \models ness(o(e,t), h(tri(e'), t'))$. Comme $e' \in E^{\chi}(t')$ et $e \in \mathbb{N}$, alors $\Pi(\chi) \models happens(e', tri(e'), t')$ et $\Pi(\chi) \models auto(e', tri(e'), eff(e'))$. Étant donné la règle :

$$(6.20) : actual(o(\text{E1},\text{T1}), o(\text{E2},\text{T2})) :- ness(o(\text{E1},\text{T1}), h(\text{GD},\text{T2})), happens(\text{E2},\text{GD},\text{T2}), auto(\text{E2},\text{GD},Eff).$$

nous avons donc $\Pi(\chi) \models actual(o(e,t), o(e',t'))$.

[$\Longleftarrow$] Nous considérons les occurrences d'évènements (e,t) et (e',t') telles que $\Pi(\chi) \models \rho$, où $\rho = actual(o(e,t), o(e',t'))$. D'après la proposition 6.7, étant donné que $\exists S \in AS(\Pi(\chi))$ tel que $\rho \in S$, alors $\exists r \in \Pi(\chi)$, $head(r) = \rho$, $body^+(r) \subseteq S$ et $body^-(r) \cap S = \varnothing$.

La seule règle r qui peut être utilisée pour obtenir ρ est :

$$(6.20) : actual(o(e,t), o(e',t')) :- ness(o(e,t), h(tri(e'), t')), happens(e', tri(e'), t'), auto(e', tri(e'), eff(e')).$$

Comme $body^+(r) \subseteq S$, nous avons $ness(o(e,t), h(tri(e'), t')) \in S$ et donc, d'après le théorème 6.3, (e,t) est une NESS-cause de $(tri(e'), t')$. Par conséquent, d'après la définition 6.12, (e,t) est une cause effective de (e',t'). □

6.4 Discussion sur l'expressivité

Dans cette section nous parlons d'expressivité. Plus précisément, nous commençons par parler d'aspects généraux qui permettent de situer $\mathcal{S}_c$, présenté dans la section 6.1 et implémenté dans la section 6.3.1, par rapport à d'autres langages de description d'action.

Comme nous l'avons détaillé dans le chapitre 3, BATUSOV et SOUTCHANSKI [2018] mentionnent trois éléments qui doivent être présents dans le formalisme de représentation de l'action et du changement choisi pour pouvoir vraiment traiter les cas de causalité effective. Comme l'indique la citation en début de chapitre, ces trois éléments sont la notion de temps, la cooccurrence d'évènements et ce que nous avons appelé les évènements naturels. Le tableau 6.2 compare différents langages de description d'action en fonction de la présence ou non de certains de ces éléments. Deux remarques doivent être faites sur ce tableau. La première est que, le temps étant pris en compte par tous les langages de description d'action, il n'apparaît pas dans le tableau. En l'occurrence, cette distinction serait pertinente si nous ajoutions les équations structurelles. Le deuxième élément est que $\mathscr{S}_c$ peut être considéré comme un point intermédiaire entre PDDL et PDDL+, ou entre $\mathscr{A}_c$ et $\mathscr{C}$. Nous avons choisi de proposer ce point intermédiaire étant donné qu'il nous permettait d'avoir tous les éléments nécessaires selon BATUSOV et SOUTCHANSKI [2018], sans avoir les évènements non déterministes ou duratifs qui soulèvent des questions causales intéressantes, mais au delà du cadre de cette thèse plus axée éthique computationnelle.

	Cooccurrence d'évènements	Évènements naturels ou axiomes	Préconditions disjonctives	Actions non déterministes ou duratives
$\mathscr{A}$**,PDDL**				
$\mathscr{A}_c$	X			
$\mathscr{B}$		X		
$\mathscr{AL}$	X	X		
$\mathscr{S}_c$	X	X	X	
$\mathscr{C}$**,PDDL+**	X	X	X	X

TABLEAU 6.2 – Comparaison de différents langages de description d'action en fonction de la présence ou non de cooccurrence d'évènements, d'évènements naturels, de préconditions disjonctives et d'actions non déterministes et ou duratives.

Nous avons à présent une idée générale de l'expressivité de $\mathscr{S}_c$. Dans la suite de cette section nous abordons plus en détail cette question en nous plaçant dans la perspective du PDDL. En effet, comme montré dans la section 2.2.1, l'état de maturité avancé du PDDL, sa vocation à faciliter l'interchangeabilité et son utilisation par une large communauté, sont autant d'arguments significatifs en faveur de ce formalisme, progressivement étendu par différents fragments. Ce travail comparatif permet d'illustrer précisément ce qu'il est possible de représenter avec $\mathscr{S}_c$.
La section 6.4.1 discute de ce que nous considérons comme étant l'expressivité d'un langage et fait la part entre les éléments qui augmentent réellement l'expressivité et ceux qui ne sont que du sucre syntaxique. Puis, la section 6.4.2 propose une version de $\mathscr{S}_c$ plus expressive et la section 6.4.3 les définitions de causalité positive qui y correspondent. Finalement, la section 6.4.4 établit un pont entre PDDL et $\mathscr{S}_c/\mathscr{S}_c^+$. Celui-ci indique comment traduire un problème en PDDL dans $\mathscr{S}_c/\mathscr{S}_c^+$, permettant ainsi aux utilisateurs de PDDL d'utiliser facilement ce que nous avons proposé dans ce chapitre.

6.4.1 Gain en expressivité ou sucre syntaxique

Dans cette section, nous discutons de ce que nous considérons comme étant l'expressivité d'un langage et faisons la part entre les éléments qui augmentent réellement l'expres-

sivité et ceux qui ne sont que du sucre syntaxique. Dans $\mathscr{S}_c$, la possibilité d'avoir des disjonctions $\psi_1 \vee \cdots \vee \psi_m \in \mathscr{F}$ dans les préconditions et les conditions de déclenchement est ce qui le rend plus *expressif* que STRIPS. Cette affirmation repose sur la définition d'expressivité proposée par NEBEL [2000] selon laquelle l'expressivité est une mesure d'à quel point un problème peut être représenté de façon concise dans un langage : « Expressive power is a measure of how concisely planning domains and plans can be expressed in a particular formalism ».
Regardons d'un peu plus près ce à quoi cela correspond. Pour réaliser cette mesure, NEBEL [2000] introduit la notion de « compilation schemes », des mécanismes qui permettent de reformuler un problème pour passer d'un formalisme à un autre. Par exemple, un évènement avec une précondition disjonctive $\psi_1 \vee \cdots \vee \psi_n$ peut être reformulé en STRIPS en créant une copie de l'évènement pour chacun des éléments disjoints ψ_i qui se retrouvent individuellement comme préconditions d'une des copies. Nous aurions alors n évènements en STRIPS. Pour savoir si la prise en compte des disjonctions permet un gain d'expressivité il faut évaluer si la conversion du problème d'un langage à un autre fait augmenter la taille du domaine exponentiellement par rapport à la représentation plus concise, ou si la longueur d'un même plan ou trace augmente de plus d'un facteur constant. Si une des deux conditions est remplie, alors la prise en compte des disjonctions ou tout autre ajout dans un langage est considéré comme augmentant son expressivité [HASLUM et collab., 2019]. Dans le cas de la disjonction, la transformation proposée n'est possible que si la formule est sous FND. Les formules de $\mathscr{F}$ étant des formules booléennes et toute formule booléenne pouvant être mise sous FND, tout évènement avec préconditions disjonctives peut être représenté sous STRIPS. Toutefois, étant donné que le passage en FND d'une formule booléenne peut faire augmenter sa taille de façon exponentielle, entraînant la même augmentation dans la taille du domaine, la possibilité d'avoir des disjonctions dans les préconditions et les conditions de déclenchement rend bien un langage plus expressif. Par conséquent $\mathscr{S}_c$ est plus expressif que STRIPS.

Ce résultat doit toutefois être relativisé pour l'implémentation. Les théorèmes de complétude et correction 6.2, 6.3 et 6.4 sont prouvés pour ψ étant sous FND. Si $\mathscr{S}_c$ est bien plus expressif que STRIPS, lors de son implémentation pour une utilisation causale, la taille du domaine d'un problème sous $\mathscr{S}_c$ est inférieure ou égale à la taille du problème équivalent reformulé dans STRIPS, mais la différence ne sera pas nécessairement exponentielle. Si nous voulons qu'elle le soit, il faudrait pouvoir prendre une formule ψ sous n'importe qu'elle forme et pouvoir prouver la complétude et la correction de l'implémentation $\pi_{\mathbb{C}}$. Ce travail reste à faire.

La façon dont les formules de $\mathscr{F}$ sont construites dans $\pi_{\mathbb{A}}$ nous permet de gérer d'autres éléments de différentes extensions de PDDL (PDDL [GHALLAB et collab., 1998], PDDL 2.1 [FOX et LONG, 2003], PDDL 2.2 [HOFFMANN et EDELKAMP, 2005]). Pour rappel, la construction des formules de $\mathscr{F}$ dans $\pi_{\mathbb{A}}$ est faite en imbriquant les prédicats `conj(GD)`, `disj(GD)` et `in(GD,GD_L)`. Ce que cela permet de gérer n'augmente pas l'expressivité au sens de NEBEL [2000]. Toutefois, ces éléments de langage restent des « sucres syntaxiques » qui permettent de formaliser un problème plus élégamment [HASLUM et collab., 2019]. Nous verrons des exemples des éléments que nous pouvons gérer dans la section 6.4.4 où nous proposons un pont entre PDDL et $\mathscr{S}_c$.

Mais revenons à $\mathscr{S}_c$ et laissons son implémentation pour plus tard. Il y a bien un élément de langage qui une fois géré permet d'augmenter l'expressivité de celui-ci au sens de NE-

BEL [2000]. Il s'agit des *effets conditionnels*, des effets qui ne sont pas appliqués dès que les préconditions sont satisfaites, il faut en plus que des conditions supplémentaires le soient. Par exemple, nous pourrions vouloir modéliser l'action consistant à lâcher un objet dont les préconditions seraient tenir l'objet et un effet intrinsèque que l'objet tombe jusqu'à la surface en dessous la plus proche. En plus de cela, nous voudrions rajouter que, si l'objet contient un liquide et qu'il se casse lorsqu'il atteint la surface, celle-ci sera mouillée. Nous avons là un nouvel effet intrinsèque de l'action. Toutefois, celui-ci est soumis à une condition qui n'a pas sa place dans les préconditions car nous voulons pouvoir réaliser l'action même si l'objet ne contient pas un liquide ou est suffisamment résistant pour ne pas se casser s'il est lâché. C'est alors qu'apparaît toute l'utilité des effets conditionnels : pouvoir rajouter une condition qui ne s'applique pas à la réalisation de l'action et donc tous ses effets intrinsèques, mais qu'à une partie.
Dans la section 6.4.2 nous présentons les quelques modifications à apporter à $\mathscr{S}_c$ pour obtenir $\mathscr{S}_c^+$, une version plus expressive par l'ajout d'effets conditionnels. Notez cependant que tout problème dans $\mathscr{S}_c^+$ peut être reformulé dans $\mathscr{S}_c$ au moyen d'une méthode pour reformuler le problème très proche de celle présentée pour les disjonctions. Nous présentons ensuite dans la section 6.4.3 les définitions de causalité positive qui gèrent l'ajout des effets conditionnels. Le choix de présenter $\mathscr{S}_c$ et non $\mathscr{S}_c^+$ directement se justifie par l'implémentation. Comme nous le verrons dans la suite, les définitions de causalité sont quelque peu plus complexes. Une première implémentation a été développée en ASP sous le même modèle que $\pi_{\mathbb{C}}$. Toutefois, les preuves de complétude et de correction semblent n'être possibles que si le programme construit les ensembles suffisants de NESS-causes et NESS-causes directes, ce qui n'est pas le cas de $\pi_{\mathbb{C}}$. L'ASP n'étant pas un langage permettant de manipuler facilement les ensembles, une implémentation en Prolog semble plus adéquate. Une première exploration de cette voie a été faite dans le cadre d'un stage de M1 par Ella Dijksman. Ce stage a permis d'obtenir des programmes $\pi_{\mathbb{A}}$' et $\pi_{\mathbb{C}}$' en Prolog reprenant $\mathscr{S}_c$ et ses définitions causales correspondantes, mais en construisant les ensembles suffisants de NESS-causes et NESS-causes directes. Il reste à ajouter à ces programmes les effets conditionnels de $\mathscr{S}_c^+$, adapter les définitions causales et prouver la complétude et correction de $\pi_{\mathbb{C}}$'.

6.4.2 Version plus expressive par l'ajout d'effets conditionnels $\mathscr{S}_c^+$

Dans cette section nous présentons les quelques modifications à apporter à $\mathscr{S}_c$ pour obtenir $\mathscr{S}_c^+$. La première modification consiste à ajouter un autre type de formule que les formules d'état $\mathscr{F}$. Nous notons $\mathscr{E}$ les formules d'effet. Dans $\mathscr{S}_c^+$ les formules d'effet peuvent prendre la forme suivante :

$$\varphi ::= [\psi]l | \varphi_1 \wedge \varphi_2,$$

où $[\psi]l$ est la notation des effets conditionnels indiquant que l est un effet d'un évènement qui se produit si la condition $\psi \in \mathscr{F}$ est satisfaite. Tout effet peut être mis sous cette forme, les effets dits classiques comme ceux dans $\mathscr{S}_c$ se notent $[\top]l$. Les formules d'effet ont donc comme forme générale $\varphi \in \mathscr{E}$, $\varphi = \bigwedge_{i\in\{1,\dots,m\}}[\psi_i]l_i$. Par souci de concision, nous pourrons adopter quand cela sera plus utile une notation ensembliste pour les formules d'effet $\varphi = \{[\psi_i]l_i \in \mathscr{E} | i \in \{1,\dots,m\}\}$.

La deuxième et troisième modification se trouve au niveau des transitions. La fonction qui associe aux évènements leurs effets doit être redéfinie. Formellement, celle-ci est définie comme :

$$\mathit{eff} : \mathbb{E} \to \mathscr{E}.$$

La notion d'évènements interférents repose sur leurs effets. La forme des effets ayant changée, les conditions pour que deux évènements soient considérés comme interférents doit changer également. Dans $\mathscr{S}_c^+$ nous considérons que deux évènements $e, e' \in \mathbb{E}^2$ sont *interférents* si l'ensemble $\{l \in Lit_{\mathbb{F}} | \exists \psi \in \mathscr{F},\ (S \vDash \psi) \wedge ([\psi] l \in eff(e) \cup eff(e'))\}$ n'est pas cohérent au sens de la définition 6.1. Contrairement à la notion d'interférence dans $\mathscr{S}_c$ qui dépendait uniquement des évènements et de leurs effets intrinsèques, ici l'interférence entre deux évènements dépend également de l'état dans lequel ils se produisent.

La quatrième et cinquième modification, les dernières, se trouvent au niveau de la transition entre états. Plus exactement dans la définition des effets effectifs et du STEE. La première définition devient :

Définition 6.14 [effets effectifs $actual\mathrm{E}ff(\mathrm{E},\mathrm{L})$]**.** *Étant donné un état partiel* $\mathrm{L} \subseteq Lit_{\mathbb{F}}$ *et un ensemble d'évènements* $\mathrm{E} \subseteq \mathbb{E}$*, le prédicat* $actual\mathrm{E}ff(\mathrm{E},\mathrm{L})$ *associe à un couple* (E,L) *un état partiel étant les effets effectifs de* E*. Il est défini comme :*

$$actual\mathrm{E}ff(\mathrm{E},\mathrm{L}) = \{l \in Lit_{\mathbb{F}} | \exists e \in \mathrm{E}, ([\psi] l \in eff(e)) \wedge (\mathrm{L} \vDash \psi) \wedge (l \notin \mathrm{L})\}.$$

Modifier $actual\mathrm{E}ff$ modifie directement l'opérateur de mise à jour $\triangleright$. La condition 3 de la définition 6.2 s'écrivant également $\mathrm{S}' = \mathrm{S} \triangleright \mathrm{E}$, celle-ci est également modifiée directement par la définition 6.14. Toutefois, celle-ci a initialement été exprimée sous la forme $\mathrm{S}' = \left\{l \in \mathrm{S} | \forall e \in \mathrm{E}, \overline{l} \notin eff(e)\right\} \cup \{l \in Lit_{\mathbb{F}} | \exists e \in \mathrm{E}, l \in eff(e)\}$. Avec l'ajout des effets conditionnels, la définition de $\mathscr{S}_c^+$ devient :

Définition 6.15 [Système de transition d'états étiqueté $\mathscr{S}_c^+$]**.** *Le système de transition d'états étiqueté* $\mathscr{S}_c^+$ *est un triplet* $\langle 2^{Lit_{\mathbb{F}}}, 2^{\mathbb{E}}, \tau \rangle$ *où* τ *est l'ensemble des relations étiquetées de transition entre états notées* $(\mathrm{S},\mathrm{E},\mathrm{S}')$*. Les triplets de cet ensemble vérifient :*

1. $\mathrm{S} \subseteq Lit_{\mathbb{F}}$ *est un état au sens de la définition 6.1 ;*
2. $\mathrm{E} \subseteq \mathbb{E}$ *vérifie :*
   (a) $\forall e \in \mathrm{E}, \mathrm{S} \vDash pre(e)$ *;*
   (b) $\nexists (e, e') \in \mathrm{E}^2,\ e \succ_{\mathbb{E}} e'$ *;*
   (c) $\forall e \in \mathbb{E}$ *tel que* $\mathrm{S} \vDash tri(e)$, $e \in \mathrm{E}$ *ou* $\exists e' \in \mathrm{E}, e' \succ_{\mathbb{E}} e$ *;*
3. $\mathrm{S}' = \left\{l \in \mathrm{S} | \forall e \in \mathrm{E}, \left([\psi] \overline{l} \notin eff(e)\right) \vee \left(\mathrm{S} \nvDash \psi\right)\right\} \cup \{l \in Lit_{\mathbb{F}} | \exists e \in \mathrm{E}, \left([\psi] l \in eff(e)\right) \wedge \left(\mathrm{S} \vDash \psi\right)\}$.

Pour ce qui est de l'implémentation, pour traduire $\mathscr{S}_c^+$ en ASP il est nécessaire de rajouter à $\pi_{\mathbb{A}}$ le prédicat `when(GD,L)` pour permettre de construire les formules d'effets. En abusant de la notation pour des raisons de compacité, nous permettrons `when(GD,C)` où `C` est une conjonction ce qui veut dire $[\psi] \bigwedge_{l \in \mathrm{C}} l \stackrel{\text{def}}{=} \bigwedge_{l \in \mathrm{C}} [\psi] l$.

Dans la section 6.3.1 nous avions vu que les prédicats `initiated(E,F,T)` et `terminated(E,F,T)` occupaient dans $\pi_{\mathbb{A}}$ le même rôle que $actual\mathrm{E}ff$ dans la sémantique de $\mathscr{S}_c$. La définition de $actual\mathrm{E}ff$ ayant changé par l'ajout des effets conditionnels, il est nécessaire d'ajouter aux règles (6.8) et (6.9) la règle (6.21). En effet, de cette façon les règles déterminant les prédicats `initiated(E,F,T)` et `terminated(E,F,T)` vérifient pour un $e \in \mathrm{E}^{\chi}(t)$ et un $\mathrm{S}^{\chi}(t)$ donnés que les trois conditions de la définition 6.14 sont satisfaites, à savoir (i) $[\psi] l \in eff(e)$, (ii) $\mathrm{S}^{\chi}(t) \vDash \psi$ et (iii) $l \notin \mathrm{S}^{\chi}(t)$. La condition (iii) est commune à $\mathscr{S}_c$ et $\mathscr{S}_c^+$, les conditions (i) et (ii) sont nouvelles. La satisfaction de la condition (i) se fait grâce à la présence de `in(Effect,when(GD,CondEffect))` dans (6.21). Pour la satisfaction de la condition (ii), deux cas sont possibles. Le premier est lorsque $\psi = \top$, donc qu'il s'agit d'un effet

classique. Dans ce cas, la condition (ii) est satisfaite par défaut. Le deuxième est lorsque $\psi \neq \top$, donc un effet conditionnel. Dans ce cas, la condition (ii) est satisfaite par la présence de `holds(GD,T)` dans la règle (6.21) qui indique $\psi \in \mathrm{S}^{\chi}(t)$. Par conséquent, `initiated(E,F,T)` et `terminated(E,F,T)` ensemble traduisent en ASP $actual\mathrm{E}ff(e, \mathrm{S}^{\chi}(t))$, où $e \in \mathrm{E}^{\chi}(t)$.

$$apply(\mathrm{E}, \mathrm{GD\_L}, \mathrm{T}) :- apply(\mathrm{E}, \mathrm{E}ffect, \mathrm{T}), holds(\mathrm{GD}, \mathrm{T}) \\ in(\mathrm{E}ffect, when(\mathrm{GD}, \mathrm{GD\_L})). \tag{6.21}$$

6.4.3 Causalité positive pour $\mathscr{S}_c^+$

Dans cette section nous présentons les modifications à apporter aux définitions causales présentées dans la section 6.2 pour qu'elles conviennent à $\mathscr{S}_c^+$. La définition 6.10 utilisant l'opérateur de mise à jour $\triangleright$, celle-ci est modifiée automatiquement par la définition 6.14. La définition des NESS-causes directes n'a donc pas besoin d'être modifiée.
La première modification arrive dans la définition des NESS-causes. Pour rappel, la définition 6.11 indique que pour construire la chaîne causale il faut s'intéresser aux occurrences d'évènements qui ont causé le déclenchement des NESS-causes. Dans $\mathscr{S}_c$, cela revenait à s'intéresser aux évènements qui ont causé que les NESS-causes aient leurs effets effectifs et donc soient des NESS-causes. Il se trouve qu'avec l'ajout des effets conditionnels cela devient plus complexe. Si nous ne nous intéressons qu'aux évènements qui ont causé le déclenchement des NESS-causes, nous avons une vue partielle des raisons pour lesquelles les NESS-causes ont eu leurs effets effectifs. Pour les effets classiques $[\top]l$ nous continuons à avoir toute l'information. Cependant, pour les effets conditionnels $[\psi]l$, où $\psi \neq \top$, il nous manque une partie de l'information. Dans le cas où l'état $\mathrm{S}^{\chi}(t)$ dans lequel l'évènement se produit ne vérifie pas la condition, $\mathrm{S}^{\chi}(t) \not\models \psi$, rien ne change puisque l n'est pas un effet effectif de l'évènement. Par contre, dans le cas où $\mathrm{S}^{\chi}(t) \models \psi$, si nous n'incluons pas les causes de (ψ, t) dans la construction de la chaîne causale, donc les causes de la satisfaction des conditions de l'effet conditionnel, nous ne serons pas en mesure de construire des ensembles suffisants de NESS-causes. Illustrons cela à l'aide d'un exemple.

Exemple 6.3 [Causer qu'un évènement ait ses effets effectifs]**.** *Soit les littéraux de fluents* $l_0, l_1, l_2, l_3, l_{c_1}, l_{c_3} \in Lit_{\mathbb{F}}^6$*, la formule d'état* $\psi = l_1 \wedge l_2 \wedge l_3$*, les actions* $a, a' \in \mathbb{A}^2$ *avec leurs préconditions* $pre(a) = pre(a') = \top$ *et leurs effets intrinsèques* $eff(a) = l_0$*,* $eff(a') = \left\{[\top]\overline{l_{c_1}}, [\top]l_{c_3}\right\}$*, l'évènement naturel* $e \in \mathbb{N}$ *avec ses conditions de déclenchement* $tri(e) = l_0$ *et ses effets intrinsèques* $eff(e) = \left\{[l_{c_1}]\overline{l_1}, [\top]l_2, [l_{c_3}]l_3\right\}$*. Nous considérons le scénario* $\sigma = \{(a,0),(a',0)\}$*. Voici les traces* τ_{χ}^{e} *et* τ_{χ}^{s} *obtenues étant donné le cadre causal* χ *décrit :*

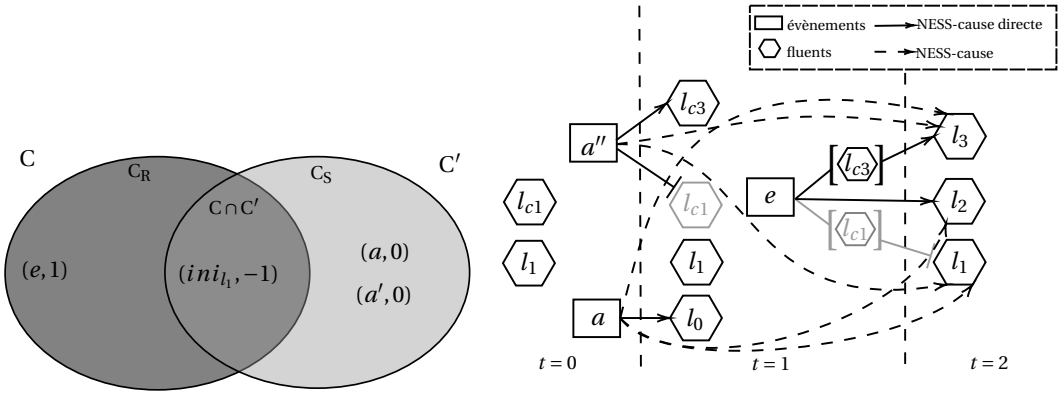
$$\mathrm{S}^{\chi}(-1) = \left\{l_0, \overline{l_1}, l_2, l_3, \overline{l_{c_1}}, l_{c_3}\right\}, \mathrm{E}^{\chi}(-1) = \left\{ini_{\overline{l_0}}, ini_{l_1}, ini_{\overline{l_2}}, ini_{\overline{l_3}}, ini_{l_{c_1}}, ini_{\overline{l_{c_3}}},\right\}$$
$$\mathrm{S}^{\chi}(0) = \left\{\overline{l_0}, l_1, \overline{l_2}, \overline{l_3}, l_{c_1}, \overline{l_{c_3}}\right\}, \mathrm{E}^{\chi}(0) = \{a, a'\}$$
$$\mathrm{S}^{\chi}(1) = \left\{l_0, l_1, \overline{l_2}, \overline{l_3}, \overline{l_{c_1}}, l_{c_3}\right\}, \mathrm{E}^{\chi}(1) = \{e\}$$
$$\mathrm{S}^{\chi}(2) = \left\{l_0, l_1, l_2, l_3, \overline{l_{c_1}}, l_{c_3}\right\}$$

Si nous cherchons les NESS-causes directes de $(\psi, 2)$*, la définition 6.10 nous indique que* $\exists!\mathrm{C},\ \mathrm{C} \underset{\mathrm{W}}{\rightarrow} (\psi, 2)$*. Il s'agit de* $\mathrm{C} = \left\{(e,1), (ini_{l_1}, -1)\right\}$.

Maintenant, que devrions nous obtenir si nous cherchons les NESS-causes de $(\psi,2)$*. D'après le processus pour trouver les NESS-causes décrit dans la section 6.2.2, il faut identifier dans* C *l'ensemble des occurrences d'évènements retirables. Dans l'exemple 6.3, cet ensemble correspond à* $C_R = \{(e,1)\}$*. D'après la définition 6.14, les effets effectifs de l'occurrence d'évènement* $(e,1)$ *sont* $actualEff(\{e\}, S^X(1)) = \{l_2, l_3\}$*. La définition 6.11 adaptée à* $\mathscr{S}_c$ *nous dirait alors qu'il existe deux ensembles suffisants de NESS-causes de* $(\psi,2)$*,* $C = \{(e,1),(ini_{l_1},-1)\}$ *ainsi que* $C' = \{(a,0),(ini_{l_1},-1)\}$*. Un problème se pose alors,* C' *n'est pas un ensemble suffisant.*

Pour réussir à déterminer l'ensemble suffisant, il faut nous intéresser à l'information manquante, les causes de la satisfaction ou non-satisfaction des conditions des effets conditionnels $[l_{c_1}]\overline{l_1}, [l_{c_3}]l_3$*. Dans cet exemple nous retrouvons les deux cas qui peuvent se présenter.*
Le premier est celui concernant l'effet conditionnel $[l_{c_1}]\overline{l_1}$*. Dans ce cas, si la condition avait été satisfaite,* $\overline{l_1}$ *aurait été un effet effectif de* $(e,1)$*, rendant faux* l_1 *étant vrai depuis* $t = 0$*, et donc empêchant* $(\psi,2)$*. Il se trouve que* $(a',0)$ *est une NESS-cause directe de* $(\overline{l_{c_1}},1)$ *; cette occurrence d'évènement a donc permis de « maintenir »* l_1 *ce qui impose qu'elle soit présente dans un ensemble suffisant de NESS-causes qui chercherait à retirer* $(e,1)$*. Notez qu'il serait possible de reformuler cette phrase en disant que* $(a',0)$ *a empêchée* $(e,1)$ *de causer* $(\overline{l_1},2)$*. Toutefois, nous réservons le terme empêché pour parler de causalité négative. Certes, il s'agit ici d'un type de causalité négative, mais ce n'est pas le même type de relation de causalité négative que nous définirons pour* $\mathscr{S}_c$ *dans le chapitre 7. En effet, nous y définirons la relation empêcher comme une relation entre deux occurrences d'évènements, une relation de nature plus proche de celle des causes effectives.*
Le deuxième cas est celui concernant l'effet conditionnel $[l_{c_3}]l_3$*. Dans ce cas, si la condition n'avait pas été satisfaite,* l_3 *n'aurait pas été un effet effectif de* $(e,1)$*, maintenant faux* l_3*, et donc empêchant* $(\psi,2)$*. Il se trouve que* $(a',0)$ *est une NESS-cause directe de* $(l_{c_3},1)$ *; cette occurrence d'évènement a donc permis de « produire »* l_3 *ce qui impose qu'elle soit présente dans un ensemble suffisant de NESS-causes qui chercherait à retirer* $(e,1)$*.*
L'ensemble suffisant de NESS-causes de $(\psi,2)$ *est dans ce cas* $C' = \{(a,0),(a',0),(ini_{l_1},-1)\}$*. Trouver l'information manquante pour construire cet ensemble suffisant de NESS-causes revient à trouver les NESS-causes directes de la formule* $\psi' = \overline{l_{c_1}} \wedge l_{c_3}$*. La figure 6.6a représente de façon ensembliste une partie du raisonnement permettant de trouver les deux ensembles suffisants de NESS-causes pour* $(\psi,2)$*. La figure 6.6b représente les traces d'évènements et d'états auxquelles ont été ajoutées les NESS-causes et les NESS-causes directes de* $(\psi,2)$ *que nous nous attendons à trouver.*

Dans l'exemple 6.3, nous avons vu que la définition de NESS-causes pour $\mathscr{S}_c$ n'est pas adaptée à $\mathscr{S}_c^+$ car il manque de l'information pour construire les ensembles suffisants de NESS-causes. Pour trouver cette information il nous manque de pouvoir déterminer une formule. Cette formule contient les informations reliées aux conditions des effets conditionnels qui par leur véracité ont permis a une occurrence d'avoir ses effets effectifs et donc d'être une cause de la conséquence. Le prédicat $after(E, L_p, L_m)$ nous permet de construire cette formule que nous appellerons formule d'effets conditionnels, il est inspiré des travaux de KHAN et LESPÉRANCE [2021].



(a) Représentation ensembliste du raisonnement permettant de trouver les deux ensembles suffisants de NESS-causes pour $(\psi, 2)$ dans l'exemple 6.3.

(b) Représentation des traces d'évènements et d'états dans l'exemple 6.3 auxquelles ont été ajoutées les NESS-causes et les NESS-causes directes de $(\psi, 2)$.

FIGURE 6.6 – Illustration de l'exemple 6.3 contenant des effets conditionnels, représenté dans $\mathscr{S}_c^+$.

Définition 6.16 [Formule d'effets conditionnels $after(\mathrm{E}, \mathrm{L}_p, \mathrm{L}_m)$]**.** *Soit un cadre causal* χ*, un ensemble d'évènements* $\mathrm{E} \in \mathrm{E}^\chi(t)$ *et des états partiels* $\mathrm{L}_m, \mathrm{L}_p \subseteq Lit_{\mathbb{F}}$ *tels que* $\mathrm{S}^\chi(t) \models \mathrm{L}_m$ *et* $\mathrm{S}^\chi(t) \not\models \mathrm{L}_p$*. Le prédicat* $after(\mathrm{E}, \mathrm{L}_p, \mathrm{L}_m) = \psi$ *avec* $\psi = \bigwedge_{l \in \mathrm{W}_\psi} l$ *où* $\mathrm{W}_\psi \subseteq Lit_{\mathbb{F}}$ *est un état partiel qui vérifie :*

- *Nécessité et minimalité de* W_ψ *:* $\mathrm{W}_\psi \rhd \mathrm{E} \models \mathrm{L}_p \cup \mathrm{L}_m$ *et* $\forall \mathrm{W}'_\psi \subset \mathrm{W}_\psi, \mathrm{W}'_\psi \rhd \mathrm{E} \not\models \mathrm{L}_p \cup \mathrm{L}_m$ *;*
- *Monotonicité de* W_ψ *:* $\forall \mathrm{W}' \supseteq \mathrm{W}_\psi, \mathrm{W}' \rhd \mathrm{E} \models \mathrm{L}_p \cup \mathrm{L}_m$*.*

Étant maintenant en mesure de trouver la formule d'effets conditionnels grâce au prédicat $after(\mathrm{E}, \mathrm{L}_p, \mathrm{L}_m)$, il nous est possible de retrouver les chaînes causales complètes grâce à l'utilisation de ce prédicat dans la définition des NESS-causes dans $\mathscr{S}_c^+$.

Définition 6.17 [NESS-causes]**.** *Étant donné un cadre causal* χ *et la relation* $\mathrm{C} \underset{\mathrm{W}}{\rightarrow} (\psi, t_\psi)$*, l'ensemble d'occurrences d'évènements* $\mathrm{C}' = \{(e,t) | e \in \mathrm{E}^\chi(t), t \in \mathbb{T}\}$ *est un ensemble suffisant de NESS-causes de* (ψ, t_ψ)*, relation que nous notons* $\mathrm{C}' \dashrightarrow (\psi, t_\psi)$*, ssi un des cas suivants est vérifié :*

- *Cas base :* $\mathrm{C}' = \mathrm{C}$*.*
- *Cas récursif : Soit un ensemble non vide d'occurrences d'évènements retirables* $\mathrm{C_R} = \mathrm{C} \setminus \mathrm{C}'$ *et trois partitions associées à la séquence* $t_1, \ldots, t_k$ *:* $\mathrm{W}_1, \ldots, \mathrm{W}_k$ *de* W*,* $\mathrm{C}(t_1), \ldots, \mathrm{C}(t_k)$ *de* C *et* $\mathrm{C_R}(t_1), \ldots, \mathrm{C_R}(t_k)$ *de* $\mathrm{C_R}$*. Il existe une séquence* $\mathrm{C_S}_1, \ldots, \mathrm{C_S}_k$*, non nécessairement monotone en temps, telle que :*
  1. $\mathrm{C_S} = \bigcup_{i \in \{1,\ldots,k\}} \mathrm{C_S}_i$ *;*
  2. $\forall i \in \{1, \ldots, k\}, \mathrm{C_R}(t_i) = \varnothing \implies \mathrm{C_S}_i = \varnothing$ *;*
  3. $\forall i \in \{1, \ldots, k\}, \mathrm{C_R}(t_i) \neq \varnothing \implies \mathrm{C_S}_i \dashrightarrow (\psi'_i, t_i)$*, où :*

$$\begin{aligned} \psi'_i &= tri(\mathrm{C_R}(t_i)) \wedge after(\mathrm{C_R}(t_i), \mathrm{L}_p, \mathrm{L}_m); \\ \mathrm{L}_p &= \mathrm{W}_i \cap actualEff(\mathrm{C_R}(t_i), \mathrm{S}^\chi(t_i)); \\ \mathrm{L}_m &= \big(\mathrm{W}_i \setminus actualEff(\mathrm{C_R}(t_i), \mathrm{S}^\chi(t_i))\big) \cup \mathrm{W}_{i+1} \cup \cdots \cup \mathrm{W}_k. \end{aligned}$$

(e,t) *est une NESS-cause de* (ψ, t_ψ) *ssi* $\exists \mathrm{C}' \subseteq \mathbb{E} \times \mathbb{T}$ *tel que* $(e,t) \in \mathrm{C}'$ *et* $\mathrm{C}' \dashrightarrow (\psi, t_\psi)$*.*

Dans la définition 6.17, la seule différence par rapport à la définition 6.11 est l'ajout de $after(\mathrm{C_R}(t_i),\mathrm{L}_p,\mathrm{L}_m)$ dans la condition 3. Reprenons l'exemple 6.3 pour voir comment nous pouvons retrouver le résultat attendu en appliquant cette définition.

Exemple 6.3 [suite]**.** *Reprenons l'exemple et appliquons le cas récursif. L'ensemble d'occurrences d'évènements retirables est* $\mathrm{C_R} = \{(e,1)\}$ *et sa partition* $\mathrm{C_R}(1) = \{(e,1)\}$ *associée à la séquence* $t_1 = 1$*. Nous sommes donc dans le cas où* $\mathrm{C_R}(t_1) \neq \varnothing$*. Nous devons alors trouver un* $\mathrm{C_{S_1}}$ *tel que* $\mathrm{C_{S_1}} \dashrightarrow (\psi'_1, t_1)$ *où :*

— $\psi'_1 = \big(tri(e,1) \wedge after(\{e\},\mathrm{L}_p,\mathrm{L}_m),1\big)$ *avec* $tri(e,1) = l$*;*

— $\mathrm{L}_p = \mathrm{W}_1 \cap actualEff(\{e\},\mathrm{S}^{\chi}(1)) = \{l_2,l_3\} \cap \{l_2,l_3\} = \{l_2,l_3\}$*;*

— $\mathrm{L}_m = \big(\mathrm{W}_i \setminus actualEff(\{e\},\mathrm{S}^{\chi}(1))\big) \cup \mathrm{W}_2 = \varnothing \cup \{l_1\} = \{l_1\}$.

Nous avons donc $\mathrm{L}_p \cup \mathrm{L}_m = \{l_2,l_3\} \cup \{l_1\} = \{l_1,l_2,l_3\}$ *et d'après la définition 6.16 nous avons alors* $after(\{e\},\mathrm{L}_p,\mathrm{L}_m) = \overline{l_{c_1}} \wedge l_{c_3}$*. Étant donné que* $tri(e,1) = l$ *et* $after(\{e\},\mathrm{L}_p,\mathrm{L}_m) = \overline{l_{c_1}} \wedge l_{c_3}$*, nous cherchons un ensemble suffisant de NESS-causes de* $(l \wedge \overline{l_{c_1}} \wedge l_{c_3},1)$*. D'après la définition 6.10,* $\mathrm{C_{S_1}} = \{(a,0),(a',0)\}$*. Nous retrouvons donc le résultats souhaité, à savoir, qu'à part* C*, l'ensemble suffisant de NESS-causes de* $(\psi,2)$ *est* $\mathrm{C}' = \{(a,0),(a',0),(ini_{l_1},-1)\}$.

Pour ce qui est de l'implémentation, pour traduire la définition des NESS-causes de $\mathscr{S}_c^+$ en ASP il est nécessaire de rajouter à $\pi_{\mathbb{C}}$ (6.22) et (6.23). (6.19) traite le cas où `o(`e_1`,`t_1`)` est une NESS-cause de $tri(\mathrm{C_R}(t_i))$. (6.22) et (6.23) correspondent au cas où `o(`e_1`,`t_1`)` est une NESS-cause de $after(\mathrm{C_R}(t_i),\mathrm{L}_p,\mathrm{L}_m)$, composante de ψ' dans la définition 6.17. (6.22) indique que `o(`e_1`,`t_1`)` est une NESS-cause de `h(l,`t_2`+1)` si `o(`e_2`,`t_2`)` est une NESS-cause directe de `h(l,`t_2`+1)`, $[\psi]l \in eff(e_2)$, $e \in \mathbb{N}$ et `o(`e_1`,`t_1`)` est une NESS-cause directe de `h(`ψ`,`t_2`)`. Cette règle traite le cas où une occurrence d'évènement contribue à ce que (e_2,t_2) produise l, c'est le cas relié à l'état partiel L_p de la définition 6.16. Dans l'exemple 6.3, cette règle concerne `o(a',0)` et le littéral de fluent l_{c3} nécessaire pour que `o(e,1)` ait comme effet effectif l_3. (6.23) indique que `o(`e_1`,`t_1`)` est une NESS-cause de `h(l,`t_2`+1)` s'il y a une relation d'inertie entre `h(l,`t_2`)` et `h(l,`t_2`+1)`, $[\psi]\overline{l} \in eff(e_2)$, $e \in \mathbb{N}$, $e \in \mathrm{E}^{\chi}(t_2)$ et `o(`e_1`,`t_1`)` est une NESS-cause directe de `h(`$\overline{\psi}$`,`t_2`)`. Cette règle traite le cas où une occurrence d'évènement contribue à ce que (e_2,t_2) maintienne l, c'est le cas relié à l'état partiel L_m de la définition 6.16. Dans l'exemple 6.3, cette règle concerne `o(a',0)` et le littéral de fluent l_{c1} dont l'absence est nécessaire pour que `o(e,1)` ne rende pas faux l_1.

$$\begin{aligned} ness(o(\mathrm{E1},\mathrm{T1}),h(\mathrm{L},\mathrm{T2}+1)) :- \; & direct\_ness(o(\mathrm{E1},\mathrm{T1}),h(\mathrm{GD\_L},\mathrm{T2})), \\ & auto(\mathrm{E2},\mathrm{GD2},Eff),in(Eff,when(\mathrm{GD\_L},\mathrm{C})), \\ & in(\mathrm{C},\mathrm{L}),direct\_ness(o(\mathrm{E2},\mathrm{T2}),h(\mathrm{L},\mathrm{T2}+1)). \end{aligned} \tag{6.22}$$

$$\begin{aligned} ness(o(\mathrm{E1},\mathrm{T1}),h(\mathrm{L},\mathrm{T2}+1)) :- \; & direct\_ness(o(\mathrm{E1},\mathrm{T1}),h(\mathrm{GD\_L},\mathrm{T2})), \\ & happens(\mathrm{E2},\mathrm{GD2},\mathrm{T2}),auto(\mathrm{E2},\mathrm{GD2},Eff), \\ & in(Eff,when(\mathrm{NGD\_L},\mathrm{C})),comp(\mathrm{NGD\_L},\mathrm{GD\_L}), \\ & in(\mathrm{C},\mathrm{NL}),comp(\mathrm{NL},\mathrm{L}),inertia(h(\mathrm{L},\mathrm{T2}),h(\mathrm{L},\mathrm{T2}+1)). \end{aligned} \tag{6.23}$$

6.4.4 Un pont entre PDDL et $\mathscr{S}_c/\mathscr{S}_c^+$

Dans cette section nous établissons un pont entre PDDL et $\mathscr{S}_c/\mathscr{S}_c^+$. Il a été construit par la mise en place d'un programme permettant de passer automatiquement d'une représentation d'un problème en PDDL à $\pi_{sce}(\sigma)$ et $\pi_{con}(\kappa_c)$, permettant ainsi aux utilisateurs de PDDL d'utiliser facilement l'implémentation de notre approche $\pi_{\mathbb{A}}$ et $\pi_{\mathbb{C}}$. Nous ne rentrerons pas dans les détails de ce programme. Par contre, à l'aide d'un exemple nous illustrons comment une telle transformation peut être faite. Nous reprenons l'exemple 2.1 dont une représentation en PDDL a été proposée dans la section 2.2.1. Comme montré dans cette section, l'état de maturité avancé du PDDL, sa vocation à faciliter l'interchangeabilité et son utilisation par une large communauté, sont autant d'arguments significatifs en faveur de ce formalisme qui nous ont convaincus d'établir un tel lien.

Le programme conçu peut être décomposé en deux parties. Une première étant un analyseur syntaxique qui a comme fonction d'extraire l'information des programmes *domain.pddl* et *problem.pddl* pour les organiser dans une structure. Une deuxième partie qui consiste à extraire les informations ordonnées dans cette structure pour écrire la description du contexte équivalente en notre formalisme ASP. Pour la première partie, nous utilisons comme base l'analyseur syntaxique PDDL Parser développé par MAGNAGUAGNO et collab. [2020]. Cet analyseur syntaxique gérait uniquement les *requirements* `:strips`, `:negative-preconditions` et `:typing` auxquels nous sommes venus ajouter tous les autres *requirements* mentionnés dans la section 2.2.1. Nous précisons également que notre deuxième partie, celle chargée de la traduction, réalise des vérifications permettant de détecter de possibles erreurs de syntaxe dans la description du problème en PDDL.

Pour faciliter la compréhension, nous adoptons le même code couleur que celui utilisé dans la section 2.2.1. Pour rappel, ce qui identifie les préconditions, les effets et le type d'évènements, est représenté en bleu clair et les opérateurs sont représentés en bleu foncé. Une exception sera faite pour `in(GD,GD_L)`. En effet, celui-ci n'est pas mis en couleur car il n'a pas d'équivalent directe en PDDL, il correspond à l'indentation et les parenthèses propres à la syntaxe de ce langage. Voici un aperçu de ce code couleur.

```
action(action_name,precondition,effect).
auto(auto_name,precondition,effect).
conj(GD).          when(GD_L,GD).
disj(GD).          in(GD,GD_L).
neg(F).
```

`action(accelerate,accelCond0,accelEff0)` est le prédicat définissant l'action `accelerate`. Ce premier évènement permet de voir la structure globale que nous utilisons pour représenter les actions, comment la conjonction et la négation sont traduites et surtout l'utilisation des effets conditionnels.

```
action(accelerate,accelCond0,accelEff0).
conj(accelCond0).
  in(accelCond0,on(vha,main_0)).
conj(accelEff0).
  in(accelEff0,neg(on(vha,main_0))).
  in(accelEff0,on(vha,main_2)).
  in(accelEff0,when(neg(moving),accelCondEff0)).
  conj(accelCondEff0).
    in(accelCondEff0,moving).
```

Nous allons voir maintenant l'évènement naturel `auto(stop,stopCond0,stopEff0)`. Il est utile pour introduire les préconditions quantifiées existentiellement. Ces préconditions quantifiées peuvent être traduites comme une disjonction de chaque terme les composant, avec les instanciations de la variable. Celles-ci sont introduites comme toutes les autres imbrications, par un prédicat `in(GD,GD_exist)`, où `GD_exist` représente la précondition quantifiée, suivi du prédicat indiquant si `GD_exist` représente des éléments disjonctifs ou conjonctifs. Étant dans le cas de préconditions existentielles, le prédicat utilisé sera `disj(GD_exist)`. Après cela, chaque terme $\texttt{GD\_L}_i$ composant la précondition existentielle est représenté par le prédicat $\texttt{in(GD\_exist,GD\_L}_i\texttt{(X))}$ dans lequel il est quantifié. Cette quantification apparaît dès que l'arité du deuxième argument du prédicat `in` est supérieure au premier. Toutefois, dans le cas où la précondition existentielle est à l'intérieur d'une précondition disjonctive, son expression peut être simplifiée en ajoutant simplement chaque instance de la variable à la disjonction en cours. De cette façon, nous pouvons omettre d'introduire le terme `GD_exit` en énonçant directement $\texttt{in(GD,GD\_L}_i\texttt{(X))}$. C'est le cas dans l'évènement naturel `stop`.

```
auto(stop,stopCond0,stopEff0).
conj(stopCond0).
  in(stopCond0,moving).
  in(stopCond0,stopCond1).
  disj(stopCond1).
    in(stopCond1,alone).
    in(stopCond1,stopExistCond0(X)) :- targets(X).
    conj(stopExistCond0(X)) :- targets(X).
      in(stopExistCond0(X),stopExistCond0_1(X)) :- targets(X).
      disj(stopExistCond0_1(X)) :- targets(X).
        in(stopExistCond0_1(X),collide_with(X)) :- targets(X).
      in(stopExistCond0(X),stopExistCond0_2(X)) :- targets(X).
      disj(stopExistCond0_2(X)) :- targets(X).
        in(stopExistCond0_2(X),heavy(X)) :- targets(X).
conj(stopEff0).
  in(stopEff0,neg(moving)).
```

Si l'évènement `stop` nous a servi pour illustrer le cas particulier des préconditions existentiellement quantifiées, `alone(P,L)`,`aloneCond0(P,L)`,`aloneEff0` a le même rôle pour les préconditions universellement quantifiées. À l'opposé des préconditions existentiellement quantifiées, les préconditions quantifiées peuvent être traduites comme une conjonction de chaque terme les composant, avec les instanciations de la variable. Le mécanisme permettant de les traduire est donc le même, il suffit de remplacer dans l'explication précédente les prédicats `disj` par des prédicats `conj`.

```
auto(alone(P,L),aloneCond0(P,L),aloneEff0) :- position(P), occupants(L).
conj(aloneCond0(P,L)) :- auto(alone(P,L),_,_).
  in(aloneCond0(P,L),on(vha,P)) :- auto(alone(P,L),_,_).
  in(aloneCond0(P,L),aloneUnivCond0(P,L,X)) :- auto(alone(P,L),_,_),
                                               targets(X), X!=vha.
  conj(aloneUnivCond0(P,L,X)) :- auto(alone(P,L),_,_), targets(X), X!=vha.
    in(aloneUnivCond0(P,L,X),on(X,P)) :- auto(alone(P,L),_,_), targets(X),
                                                               X!=vha.
    in(aloneUnivCond0(P,L,X),neg(alive(L,X))) :- auto(alone(P,L),_,_),
                                                 targets(X), X!=vha.
conj(aloneEff0).
  in(aloneEff0,alone).
```

Si par une action du véhicule au temps T, celui-ci se retrouve à la même position qu'un autre objet au temps T+1, le déclenchement d'un collision est prioritaire à cet instant par rapport à n'importe quelle action que pourrait faire le véhicule. Cette priorisation est représentée de la façon suivante :

```
priority(Au,A) :- auto(Au,GD,Effect), action(A,GD1,Effect1).
```

Dans le chapitre 5 nous mettons en avant notre volonté de contribuer à la clarification du domaine de la causalité positive. Une de nos contributions dans ce sens a consisté à proposer une typologie permettant de comparer formellement les approches entre elles. Cette comparaison est possible du moment où sont établis les liens nécessaires entre le formalisme utilisé et $\mathscr{S}_s$, ainsi qu'entre les relations causales de l'approche et celles de la section 5.1.2. Par leur sémantique et syntaxe très différente à $\mathscr{S}_s$, les équations structurelles peuvent plus difficilement établir le premier lien. Les approches reposant sur des équations structurelles étant majoritaires aujourd'hui, ne pas pouvoir utiliser la typologie proposée pour comparer ces approches entre elles, ou se comparer avec ces approches serait préjudiciable.
Pour palier à cela, nous proposons un outil permettant de passer d'un programme en PDDL vers un programme en équations structurelles et en Calcul des Situations sous le même modèle que celui dont il a été question dans cette section. L'analyseur syntaxique est commun aux trois programmes, la traduction, elle, est propre à chacun. Une première version de cette traduction a été faite dans le cadre de stages de M1 par Manon Lefèvre et Nadjia Ikhlef. Cette première version a été améliorée et étendue dans le cadre d'un stage de M2 réalisé par Emmanuel Chenuaud. En effet, lors de ce stage, une interface a été créée pour faciliter les comparaisons de résultats pour un même exemple entre l'approche de HALPERN [2016] et celle que nous proposons. Cela inclut l'étape de traduction de l'exemple en PDDL aux formalismes respectifs et l'application de la définition de causalité de chaque approche, puis une dernière étape de comparaison des relations causales.

6.5 Conclusion

Dans ce chapitre nous avons présenté de façon détaillée le cœur de l'approche que nous proposons pour représenter l'action et le changement, puis établir des relations causales complexes nous permettant de mieux comprendre l'évolution du monde. Cette présentation était structurée en trois parties. Dans la première nous avons présenté $\mathscr{S}_c$, un langage de description d'action adapté au raisonnement causal et éthique. Il peut être considéré comme un point intermédiaire entre PDDL et PDDL+, ou entre $\mathscr{A}_c$ et $\mathscr{C}$. Nous avons choisi de proposer ce point intermédiaire étant donné qu'il nous permettait d'avoir tous les éléments nécessaires selon BATUSOV et SOUTCHANSKI [2018] pour traiter la plupart des exemples dans le domaine de la causalité.

Dans la deuxième, nous avons présenté la définition de causalité positive que nous proposons. Cette présentation s'est faite en différentes étapes. Nous avons commencé par introduire les NESS-causes directes, la relation causale à partir de laquelle toutes les autres sont construites. La définition que nous en donnons est une représentation du test NESS de WRIGHT [2011] dans $\mathscr{S}_c$. Puis, nous avons généralisé cette définition et introduit les NESS-causes. Enfin, nous avons défini à partir des NESS-causes la relation classiquement utilisée dans le domaine, les causes effectives. Une fois, ces définitions introduites nous avons

montré leurs propriétés, dont celles par rapport à la typologie de la surdétermination du chapitre 5. Pour cela il a été nécessaire que nous montrions que $\mathscr{S}_c$ pouvait être vu comme une spécification du $\mathscr{S}_s$ introduit dans le chapitre 5.
Ensemble, ces deux premières parties forment la première marche de notre approche permettant de raisonner sur la causalité. Comme montré dans la section 3.1, il existe plusieurs approches permettant un tel raisonnement. Toutefois, comme montré dans la section 3.2, ces approches ne sont pas tout à fait adaptées aux problèmes éthiques qui intéressent l'éthique computationnelle. Celle que nous avons proposée a été conçue pour l'être, aussi bien par les choix de conception dans le langage de description d'action, que par la définition de causalité développée. Il s'agit là de la base de l'approche causale commune pour l'éthique computationnelle que nous proposons.

Ensuite, nous avons discuté de l'expressivité de notre langage de description d'action par rapport à STRIPS. Cette discussion nous a mené à proposer une version plus expressive de $\mathscr{S}_c$, $\mathscr{S}_c^+$. Nous avons adapté nos définitions de causalité à ce langage plus expressif et nous avons donné les éléments à ajouter dans l'implémentation pour qu'elle corresponde à cette version plus expressive.
Finalement, nous avons établi un pont entre PDDL et $\mathscr{S}_c/\mathscr{S}_c^+$. Celui-ci a été construit par la mise en place d'un programme permettant de passer automatiquement d'une représentation d'un problème en PDDL à $\pi_{sce}(\sigma)$ et $\pi_{con}(\kappa_c)$, permettant ainsi aux utilisateurs de PDDL d'utiliser facilement l'implémentation de notre approche, $\pi_{\mathbb{A}}$ et $\pi_{\mathbb{C}}$.

Chapitre 7

Contribution : modélisation et représentation de la négation dans la relation causale et de la transitivité

« *The causal status of omissions is philosophically controversial. While some philosophical theories of causation allow causation by omission, there are reasons to be suspicious of omissions, as of absences more generally.* »

ABRAMS [2022]

Sommaire

La contribution de cette thèse au domaine de l'éthique computationnelle est principalement faite par le biais de la causalité. En effet, dans la section 4.4 il a été montré que la causalité était une pièce fondamentale lorsqu'il est question de formaliser une grande partie des théories morales vues dans le chapitre 1. Dans la section 3.2 il a été montré que les approches en causalité effective existantes ne permettent pas de répondre à un certain nombre d'attentes propres à l'éthique computationnelle. Il a donc été décidé de proposer une nouvelle approche causale commune pour l'éthique computationnelle permettant d'y répondre. Si contrairement au chapitre 4 celui-ci n'a pas été fait en étroite collaboration avec Guillaume Gervois, la contribution qui y est présentée n'aurait pas pu être faite sans la notion de décision développée conjointement et les discussions passionnantes sur la négation dans la causalité.

Ce chapitre est la suite de la présentation détaillée de notre proposition de représentation de l'action, du changement et de la causalité pour l'éthique computationnelle. La présentation de l'ensemble de notre approche causale est distribuée entre le chapitre 6 et ce chapitre. Comme mentionné dans l'introduction du chapitre précèdent, l'ensemble de notre approche peut être vue comme un escalier à quatre marches où chaque marche correspond à un des défis dans le domaine identifiés dans la section 3.2. Dans le chapitre 6, nous avons présenté la première marche qui correspond au traitement de la causalité positive. Une fois construite elle permet de s'interroger sur les causes de la véracité d'une partie de l'état du monde ou sur les causes du fait qu'un évènement se soit produit. Nous avons appelée celle-ci le cœur de l'approche car elle suffit à la causalité effective. En effet, comme nous montrons dans ce nouveau chapitre, l'essentiel des trois défis qu'il reste à traiter sont quelque part en dehors du cadre purement causal, ils sont de l'autre côté de la frontière qui sépare causalité effective et responsabilité. Toutefois, nous montrons que la première marche construite est une base factuelle propice à les traiter.
La deuxième marche traite d'omission. Une fois construite elle permet de s'interroger sur les conséquences qui peuvent être attribuées à la non occurrence d'un évènement. Comme vu dans la section 3.2.2.2, lorsqu'un agent décide de ne pas agir, nous voulons parfois attribuer des conséquences à cette omission. De même, lorsqu'il décide d'agir il peut renoncer à d'autres actions, omissions auxquelles nous voulons également parfois attribuer des conséquences. Ces conséquences ont une importance toute particulière en éthique computationnelle, dans certaines théories morales nous devons être en mesure de pouvoir raisonner sur les décisions alternatives à disposition de l'agent.
La troisième marche traite de la notion d'empêcher. Une fois construite elle permet de s'interroger sur les causes pour lesquelles un évènement ne s'est pas produit. La question de ces causes se pose en éthique computationnelle, nous voulons être en mesure de pouvoir raisonner sur le fait qu'un agent ait agi pour empêcher le mal ou le bien de se produire. Une typologie des cas de surdétermination comme celle présentée dans le chapitre 5 peut aussi être construite pour cette forme de négation dans la causalité. Toutefois, comme nous le verrons par la suite, les relations causales attendues pour un même type de surdétermination ne sont pas nécessairement les mêmes qu'il s'agisse d'un cas de causalité positive ou négative.
La quatrième et dernière marche de notre escalier traite la transitivité. Une fois construite elle permet de s'interroger sur la portée que peut avoir une décision. Dans le chapitre 6 la transitivité de la causalité positive s'arrête du moment où la volition d'un agent intervient dans la chaîne causale. En éthique computationnelle ce choix se justifie. Toutefois, certaines

décisions qu'un agent prend peuvent s'avérer cruciales pour qu'un autre agent puisse agir. S'intéresser à ces relations fait sens en éthique computationnelle, nous voulons être en mesure de pouvoir raisonner sur toutes les conséquences des décisions d'un agent, même si celles-ci dépendent également de la volition d'un autre agent.

		« Conséquence »	
		positive	négative
« Cause »	positive	$(e, t) \rightsquigarrow (e_\psi, t_\psi)$	$(e, t) \rightsquigarrow (\overline{e_\psi}, t_\psi)$
	négative	$(\overline{e}, t) \hookrightarrow (e_\psi, t_\psi)$	$(\overline{e}, t) \hookrightarrow (\overline{e_\psi}, t_\psi)$

TABLEAU 7.1 – Les quatre relations « causales » élémentaires de type causes effectives qu'il peut y avoir dans un STEE.

Le tableau 7.1 illustre les quatre relations élémentaires qu'il peut y avoir dans un STEE. Le cas où aussi bien la cause que la conséquence sont des occurrences d'évènements correspond à la première marche de notre approche. C'est ce que nous avons appelé causalité positive et que nous avons traité dans le chapitre 6. Les deux cas où la cause est une non occurrence d'évènement correspondent à la deuxième marche de notre approche. C'est ce que nous avons appelé *omission*, ou par abus de langage « causes négatives », et que traite la section 7.1. Plus exactement cette section montre que l'omission ne peut avoir un statut causal puisqu'il s'agit purement d'une question de responsabilité. Cette question va donc au-delà du cadre de cette contribution. Le cas où la cause est une occurrence d'évènement et la conséquence une non occurrence d'évènement correspond à la troisième marche de notre approche. C'est ce que nous avons appelé *conséquences négatives*, ou par abus de langage « empêcher », et que traite la section 7.2. Plus exactement cette section montre que la notion d'empêcher peut être décomposée en deux niveaux de raisonnement, un appartenant à la causalité effective, le deuxième étant une question de responsabilité et de transitivité. Pour le premier niveau cette section propose une approche factuelle et l'intègre à l'approche de causalité positive proposée dans le chapitre 6. Le deuxième niveau est au-delà du cadre de cette contribution. Toutefois, la section 7.3 montre que notre approche factuelle est une base pour traiter les questions de responsabilité liés à la notion d'omission et d'empêcher dans toute leur complexité. Cela est fait en proposant plusieurs modélisations possibles de ces notions. Finalement, la section 7.4 montre que la question de la transitivité est également une question de responsabilité. Puis, elle expose les éléments à ajouter à notre approche factuelle afin qu'elle puisse servir de base pour traiter les questions de responsabilité liées à la notion de transitivité.

7.1 Causes négatives, ou omission : une question de responsabilité

Dans cette section nous abordons la problématique des conséquences qui peuvent être attribuées à la non occurrence d'un évènement. La question sur ces conséquences se pose en éthique computationnelle, dans certaines théories morales nous devons être en mesure de pouvoir raisonner sur les décisions alternatives à disposition de l'agent. En effet, un agent peut être considéré responsable de ne pas avoir agi alors qu'il le pouvait. Cette omission de sa part n'est rien d'autre qu'une non occurrence d'un évènement. Nous retrouvons cette intuition dans l'exemple 3.13 sur le jardinier qui omet d'arroser la vigne qui en meurt. C'est ce qu'expliquent BOURGNE et collab. [2021] :

> An agent, broadly speaking, is an entity with the power to act ; exercising this capacity makes agents liable to blame or praise, both in ethics and in the law. Yet this capacity is not just a matter of performed actions : surely we are to blame if we choose not to rescue a drowning child. Responsibility therefore also pertains to the power to *not* act. Whether there is a fundamental moral difference between actions and omissions is an important point of debate within moral philosophy (see BENNETT [1998]; FOOT [2002]).

Avant de passer à la suite, il est important de clarifier les termes utilisés. Le terme omission est couramment utilisé pour se référer à une non occurrence d'évènement particulière, il s'agit de la non occurrence d'une action qu'aurait pu réaliser un agent. Les agents étant au centre des problématiques en éthique et en droit, ce terme est couramment utilisé dans les exemples. Nous l'adopterons dans ce même contexte. Toutefois, les arguments que nous avançons dans cette section sont plus larges que cela, ils s'intéressent à tout type de non occurrence d'évènement car, comme nous le verrons dans la section 7.2, la notion d'empêcher est reliée à la non occurrence d'évènements au sens large.

Le statut causal de l'omission n'est pas aussi clair que sa capacité à entraîner de la responsabilité. Comme l'explique ABRAMS [2022] dans la citation de début de chapitre que nous rappelons, il s'agit d'une question controversée : « The causal status of omissions is philosophically controversial. While some philosophical theories of causation allow causation by omission, there are reasons to be suspicious of omissions, as of absences more generally ». Il existe un courant qui défend l'idée qu'une omission, ou plus généralement une non occurrence d'évènement, ne peut pas être une cause.
Pour mieux comprendre cette position, déterminons formellement ce qu'est une non occurrence d'évènement dans $\mathscr{S}_c$. Commençons par rappeler quelques notations et concepts utilisés précédemment et étendons les pour l'omission. Étant donné χ : une formule vraie dans un état correspond à un couple (ψ, t) qui vérifie $S^{\chi}(t) \vDash \psi$; une occurrence d'évènement correspond à un couple (e, t) qui vérifie $e \in E^{\chi}(t)$. Une *non occurrence d'évènement* correspond à un couple $(\overline{e}, t)$ qui vérifie $e \notin E^{\chi}(t)$. Cela signifie qu'étant donné le contexte χ, l'évènement e ne s'est pas produit au temps t. Dans le cadre d'un STEE, les évènements étant des transitions entre les états, l'absence d'évènement se traduit par l'absence de transition [MILL, 1843] : « From nothing, from a mere negation, no consequences can proceed ». Sans transitions le monde reste tel qu'il est. Puisque de cette absence rien n'est produit, il apparaît impossible de déduire une quelconque relation de causalité effective.
Cette position n'est pas un simple épiphénomène de la modélisation choisie, elle découle de l'existence du principe d'inertie sur lequel repose la solution au problème du décor de tous les formalismes présentés dans le chapitre 2. Pour rappel, cela se traduit dans ces formalismes par le fait que si la valeur des fluents n'est pas affectée par l'effet des évènements de la transition, alors dans l'état d'arrivé leur valeur est la même que dans l'état de départ. De ce fait, sans un évènement qui produit des changements dans le monde, tout se maintient tel que c'est et la seule cause effective de cela est l'inertie. C'est cette vision que nous pouvons percevoir lorsque WRIGHT [1985] écrit : « By definition an omission is a nonevent—something which did not happen—which only rarely will trigger an actual causal sequence that can be directly perceived or traced ». Plus généralement, la position selon laquelle une non occurrence d'évènement n'a pas réellement un statut causal car elle ne produit rien est défendue par ce que ABRAMS [2022] appelle « production theories of causation » [DOWE,

2000; EPSTEIN, 1973; FAIR, 1979; GLENNAN, 2017; MOORE, 2009; SALMON, 1994]. L'hypothèse que nous défendons est que l'intuition consistant à attribuer un statut causal à une non occurrence d'évènement s'explique par la confusion courante entre causalité et responsabilité que dénonce WRIGHT [1985], sujet développé dans la section 3.1.2.2. De fait, derrière toute volonté d'attribuer un statut causal à une non occurrence il y a un raisonnement hypothétique. Une non occurrence ne produit rien, le monde reste inchangé. Mais si un agent avait la possibilité d'agir et donc d'a priori changer le monde, le fait qu'il ne l'ait pas fait le rend dans certaines conditions imputable du fait que le monde soit resté tel qu'il est. Dans l'exemple 3.13 sur le jardinier, en l'absence d'un évènement qui change l'état de la terre de la vigne de sèche à mouillée, la terre reste sèche. Le jardinier ayant la possibilité de changer cela mais ne le faisant pas, il est imputable que la terre reste sèche et donc de ne pas arrêter le processus biochimique entraînant la mort de la plante ; processus qui lui correspond à de la causalité positive. Mais après tout, comme mentionné dans la section 3.2.2.2, pourquoi les autres habitants de l'immeuble ne sont pas eux aussi considérés comme imputables par leur omission au même titre que le jardinier ? La réponse courte est : parce que contrairement aux habitants, nous considérons par la description de l'exemple que c'était le devoir du jardinier et non des habitants. La réponse longue est l'objet de cette section.

Il est important à ce niveau de distinguer deux problématiques distinctes qui apparaissent lorsqu'il est question d'attribuer des conséquences à une non occurrence d'évènement. Chacune des problématiques correspond à un des types de relations causales dans un STEE que nous avons identifiées dans la section 5.1.2. Dans le cas des $\mathscr{F}-causes$ de la définition 5.4, attribuer des conséquences à une non occurrence d'évènement revient à s'intéresser au cas où nous voulons attribuer la conséquence (ψ, t_ψ), où $\psi \in \mathscr{F}$, à la non occurrence d'évènement $(\overline{e}, t)$. Dans le cas des causes effectives de la définition 5.5, attribuer des conséquences à une non occurrence d'évènement revient à s'intéresser au cas où nous voulons attribuer la conséquence (e_ψ, t_ψ) ou $(\overline{e_\psi}, t_\psi)$, à la non occurrence d'évènement $(\overline{e}, t)$. Cette distinction n'est possible que par la clarification apportée par l'utilisation d'un STEE. Comme nous le verrons dans la section 7.3, cette distinction n'est pas anodine. Toutefois, dans cette section nous n'insisterons pas nécessairement sur cette distinction car tout ce qui y est dit s'applique aux deux cas.

Cette section est divisée en trois parties. La section 7.1.1 approfondit l'étude des problèmes reliés à l'omission commencée dans la section 3.2.2.2. Comme pour la causalité positive, les problèmes reliés à l'omission apparaissent dans les cas de surdétermination. Cette étude expose une différence flagrante entre le traitement des cas de causalité positive et les cas d'omission. La section 7.1.2 montre qu'attribuer des conséquences à une non occurrence d'évènement nécessite de passer par un raisonnement hypothétique pour savoir si l'occurrence de l'évènement aurait pu faire la différence. Elle illustre que pouvoir faire la différence n'est pas quelque chose de clairement défini, cela dépend du raisonnement hypothétique qui est réalisé. Finalement, la section 7.1.3 explique que c'est dans le choix des mondes qu'il est possible d'envisager dans le raisonnement hypothétique que repose la décision de quelles conséquences pourront être attribuées à chaque non occurrence. Elle montre que le choix des mondes est intrinsèquement subjectif car dépendant d'un grand nombre de facteurs en dehors de la causalité. Cela montre que l'omission est essentiellement une question de responsabilité.

7.1.1 L'omission et la surdétermination

Dans cette section nous approfondissons l'étude des problèmes reliés à l'omission commencée dans la section 3.2.2.2. Comme pour la causalité positive, les problèmes reliés à l'omission apparaissent dans les cas de surdétermination. Si nous reprenons le cas du jardinier et que nous imaginons que la vigne n'est pas dans un lieu public mais dans un lieu inaccessible à tout être humain à part lui, alors il est probable qu'il devienne plus simple de dire que son omission est une cause de la mort de la vigne. Du moment où nous sortons du cas où il peut exister une surdétermination, la différence entre occurrence d'évènement et non occurrence d'évènement s'estompe. C'est ce qu'à observé MOORE [2019] lorsqu'il a étudié les jugements de nombreux cas en droit pour en extraire les tendances. Dans la plupart des cas de causalité positive où il y avait un unique support W = $\{l_1, \ldots, l_n\}$, les auteurs des occurrences d'évènements ayant chacune respectivement causé un des littéraux $l_1, \ldots, l_n$, sont tous reconnus responsables. Il appelle ces cas « garden variety concurrent cause cases ». Si dans les cas du même type il est question d'une non occurrence d'évènement à la place d'une occurrence d'évènement, les occurrences comme les non occurrences donnent lieu à une responsabilité. Ce résultat ne peut pas directement être transposé à la causalité. Pour rappel, même si une occurrence d'évènement est reconnue comme étant une cause d'un préjudice, cela n'implique pas automatiquement que l'agent l'ayant réalisée sera reconnu responsable. De même, si l'individu est reconnu responsable, cela n'implique pas automatiquement qu'il est à l'origine d'une occurrence d'évènement étant une cause du préjudice, en tout cas dans le sens de la causalité effective. Toutefois, les observations faites par MOORE [2019] dans ces cas réels restent une façon intéressante de sonder les intuitions générales sur la différence entre causalité positive et omission.

Les cas de surdétermination qui impliquent au moins une omission ont fait l'objet de nombreuses études. Comme pour la surdétermination en causalité positive étudiée dans le chapitre 5, les discussions autour de ces cas reposent surtout sur des exemples précis qui font débat, mais à part ces exemples communs il n'y a pas de définitions formelles communes identifiant clairement ce qui est étudié. ABRAMS [2022] a compilé les différentes dénominations qui sont données à ces problèmes : FISCHER [1992] parle de « concurring omissions » ; VYHLIDAL [1999] parle de « concurrent omissions » ; WRIGHT [2001] parle de « over-determined multiple omissions », mais aussi de « overdetermined negative causation » [WRIGHT, 2011] ; MOORE [2012] parle de « omissive overdetermination » ; GREEN [2017] parle de « double omissions ». ABRAMS [2022] présente également plusieurs exemples qui font l'objet de discussions. En voici quelques un.

Exemple 7.1 [Saunders System Birmingham Co. v. Adams]**.** *Un conducteur, distrait, traverse une intersection au volant d'une voiture de location. Il n'aperçoit pas le piéton en train de traverser, il ne freine donc pas et le percute. À l'insu du conducteur, les freins ne fonctionnent pas correctement, un problème que l'agence de location aurait dû détecter et réparer avant de lui louer la voiture, mais elle ne l'a pas fait en raison d'une inspection négligente. La négligence du conducteur à ne pas freiner n'a eu aucune incidence, car les freins n'auraient pas fonctionné de toute façon. De même, la défaillance de l'agence à réparer les freins n'a pas joué de rôle, étant donné que le conducteur n'a pas utilisé les freins en premier lieu. Qui est responsable des blessures du piéton, le conducteur, l'agence, ni l'un ni l'autre, ou les deux ?*

Exemple 7.2 [notice de médicament]**.** *Une entreprise pharmaceutique ou un fabricant de biens omettent de fournir correctement des avertissements sur les risques d'un médicament ou d'un produit. Puis, le médecin ou le consommateur omettent de consulter l'étiquette d'avertissement mal formulée sur le médicament ou le produit.*

Exemple 7.3 [Elayoubi v Zipser]**.** *Un hôpital effectue un accouchement par voie basse sur une patiente ayant précédemment accouché par césarienne, ce qui augmente les risques pour le bébé. Le premier hôpital n'a pas enregistré l'information dans les dossiers médicaux et le deuxième hôpital n'a pas demandé les dossiers.*

Exemple 7.4 [New York Central R.R v. Grimstad]**.** *Le passager d'un navire qui ne dispose pas de gilets de sauvetage tombe par-dessus bord et se noie. Se pose alors la question de savoir s'il y avait eu un gilet de sauvetage, aurait-il réellement été utile ou tout simplement aurait-il été utilisé ?*

Exemple 7.5 [Reynolds v. Texas & Pacific Railway]**.** *Lorsqu'une victime en état d'ébriété est blessée en tombant dans un escalier insuffisamment éclairé, est-ce que l'ajout d'un meilleur éclairage aurait fait une différence ?*

Exemple 7.6 [Weeks v. McNulty]**.** *Lorsqu'une victime en état d'ébriété ne se réveille pas avec le retentissement de l'alarme incendie dans un immeuble en feu sans échappatoire, est-ce que la présence d'une sortie de secours aurait fait une différence ?*

Dans les cas de surdétermination, la différence dans notre intuition entre occurrence d'évènement et non occurrence d'évènement n'est pas négligeable. Comme pour les cas de causalité positive, un simple but-for considère qu'il n'y a pas de causes au préjudice, aucune non occurrence est nécessaire à causer le préjudice. En effet, si l'évènement avait eu lieu le préjudice n'aurait pas été empêché. Le conducteur aurait pu freiner, le médecin aurait pu regarder la notice, le premier hôpital aurait pu enregistrer les informations dans le dossier, et ainsi de suite, le préjudice n'aurait pas été empêché. Comme en causalité positive, dans les cas avec omission l'échec du but-for ne veut pas dire qu'il n'y a aucun responsable. Jusqu'ici, le problème est le même qu'il s'agisse d'une action ou d'une omission.
Si le problème est le même, l'intuition pour le résoudre semble ne pas l'être. C'est ce qui a été constaté par MOORE [2019] après l'étude des jugements de nombreux cas en droit pour en extraire les tendances. Dans la plupart des cas avec de la surdétermination, un agent qui agit et dont l'action contribue à produire une conséquence, est reconnu responsable, même si sans celle-ci le préjudice aurait hypothétiquement quand même eu lieu. Alors que dans la plupart des cas avec de la surdétermination, un agent qui omet d'agir alors que son action aurait hypothétiquement pu empêcher une conséquence, à condition d'ignorer le cas de surdétermination, n'est pas reconnu responsable. De façon plus détaillée :

— Dans la plupart des cas de causalité positive où il y avait plusieurs supports étant des singletons $\mathrm{W}^1 = \{l_1\}, \ldots, \mathrm{W}^n = \{l_n\}$, les auteurs des occurrences d'évènements ayant chacune respectivement causé un des littéraux $l_1, \ldots, l_n$, sont tous reconnus responsables. Il appelle ces cas « symmetrical overdetermination ». Si dans les cas du même type il est question d'une non occurrence d'évènement à la place d'une occurrence d'évènement, les non occurrences ne donnent pas lieu à une responsabilité.

— Dans la plupart des cas de causalité positive où il y avait plusieurs supports comme $\mathrm{W}^1 = \{l_1^1 \wedge \cdots \wedge l_m^1\}, \ldots, \mathrm{W}^n = \{l_1^n \wedge \cdots \wedge l_k^n\}$, les auteurs des occurrences d'évènements

ayant chacune respectivement causé un des littéraux $l_1^1, \ldots, l_m^1, \ldots, l_1^n, \ldots, l_k^n$, sont tous reconnus responsables. Notez que ce cas n'exclut pas la possibilité d'avoir un même littéral dans plusieurs supports. Il appelle ces cas « mixed cases ». Si dans les cas du même type il est question d'une non occurrence d'évènement à la place d'une occurrence d'évènement, les non occurrences ne donnent pas lieu à une responsabilité.

— Dans la plupart des cas de causalité positive où il y avait plusieurs supports comme $\mathrm{W}^1 = \{l_1^1\}, \ldots, \mathrm{W}^n = \left\{l_1^n \wedge \cdots \wedge l_k^n\right\}$ et que la taille de ces supports était très différente, les auteurs des occurrences d'évènements ayant chacune respectivement causé un des littéraux $l_1^1, \ldots, l_1^n, \ldots, l_k^n$, sont tous reconnus responsables. Il appelle ces cas « asymmetrical overdetermination ». Si dans les cas du même type il est question d'une non occurrence d'évènement à la place d'une occurrence d'évènement, les non occurrences ne donnent pas lieu à une responsabilité.

— Dans la plupart des cas de causalité positive où il est question de préemption, seul les auteurs des occurrences d'évènements dont le chemin causal a abouti sont reconnus responsables. Si dans les cas du même type il est question d'une non occurrence d'évènement à la place d'une occurrence d'évènement, les non occurrences ne donnent pas lieu à une responsabilité.

Il est important de garder en tête qu'il ne s'agit là que de tendances, ces résultats n'ont pas été obtenus en appliquant exactement le même raisonnement comme cela devrait l'être pour une enquête causale. Comme l'explique ABRAMS [2022], le jugement de ces cas n'est pas passé par le classique but-for, mais par l'utilisation de raisonnements alternatifs. Cela donne lieu à une inhomogénéité dans les jugements de cas qui d'un point de vue causal semblent être équivalent. Cela témoigne d'une différence de traitement entre les cas de pure causalité positive et ceux contenant des omissions. Cette différence est soulignée par ABRAMS [2022] :

> If cases of omissive overdetermination really were just instances of standard overdetermination, we should expect the law to treat them as such, finding each omitter liable. Yet that is not what courts and commentators have done. In Saunders, for example, the leasing agent's negligence was rendered causally irrelevant, since the driver never attempted to brake. Similar cases include : Rouleau v. Blotner 152 A. 917 (N.H. 1931) (a driver's negligent failure to signal before turning was not cause of an accident if the oncoming driver was not looking) ; Weeks v. McNulty (n 7) (negligent failure to furnish a hotel with a fire escape didn't cause death if the decedent couldn't have used it anyway), as well as Grimstad mentioned above (n 5). On the other hand, there are cases in which courts have found such omissions to be "substantial factors" and hence grounds for liability, e.g Kitchen Krafters Inc. v. Eastside Bank, 780 P.2d 567 (Mont. 1990).

Contrairement aux cas de causalité positive où il existe une structure causale sur laquelle il est possible de s'appuyer, dans les cas d'omission il n'y a pas de structure qui permette cela. Dans les cas de surdétermination préemptive comme définis dans les définitions 5.8 et 5.9, uniquement un des chemins causaux a abouti. Il est donc aisé de relier factuellement les évènements qui appartiennent à ce chemin avec la conséquence. Dans l'exemple 3.8 nous savons que la victime est morte par déshydratation et non par empoisonnement, l'assassin ayant vidé la gourde peut aisément être identifié comme une cause effective de cette mort. Dans les cas de surdétermination duplicative ou symétrique comme définis dans les

définitions 5.10, 5.11 et 5.12, les deux chemins causaux ont abouti. Il est donc aisé de relier factuellement les évènements qui appartiennent aux deux chemins avec la conséquence. Dans l'exemple 3.6 nous savons que la victime est morte par empoisonnement, les deux assassins ayant versé une dose létale de poison dans la boisson de la victime peuvent aisément être identifiés comme des causes effectives de cette mort. Dans les exemples 7.1 à 7.6, il existe des chaînes causales positives menant aux conséquences. En l'occurrence, dans l'exemple 7.1 la voiture doit être en mouvement pour percuter le piéton. Ce mouvement est causé par le conducteur. Il est clair que le conducteur est une cause effective de l'accident. Par contre, aucune chaîne causale existante ne relie une des omissions à la conséquence puisque par définition une non occurrence d'évènement ne peut être à l'origine d'un changement et donc d'un processus. La question qui se pose alors est d'une autre nature : aurait-il pu l'éviter ? C'est à ce moment là que son omission et celle de l'entreprise de location entre dans l'analyse. Cette question appartient à l'enquête sur la proximité de la cause (« proximate-cause inquiry ») introduite dans la section 3.1.2.2. Pour rappel, cette étape propre à déterminer la responsabilité examine d'autres facteurs afin d'évaluer s'ils atténuent ou éliminent la responsabilité juridique du défendeur pour le préjudice. En effet, malgré le fait qu'il soit une cause, s'il n'avait aucun moyen d'éviter le préjudice, il semble contre-intuitif de le tenir pour responsable. La question devient alors d'évaluer s'il est vraiment possible d'affirmer « qu'il n'avait aucun moyen d'éviter le préjudice ». En quelque sorte, pour que l'omission soit pertinente il faut qu'il existe la possibilité de faire la différence par rapport à ce qui s'est réellement passé. Le besoin de devoir faire la différence est approfondi dans la section suivante.

7.1.2 L'omission implique un besoin de pouvoir faire la différence

Dans cette section nous montrons qu'attribuer des conséquences à une non occurrence d'évènement nécessite de passer par un raisonnement hypothétique pour savoir si l'occurrence de l'évènement aurait pu faire la différence. Puis, nous exposons que pouvoir faire la différence n'est pas quelque chose de clairement défini, cela dépend du raisonnement hypothétique qui est réalisé.

Si l'intuition concernant les cas de responsabilité n'est pas la même pour les cas classiques de causalité positive et ceux contenant une omission, c'est principalement du fait que pour que l'omission soit pertinente il faut qu'il existe la possibilité de faire la différence par rapport à ce qui s'est réellement passé.
Comme mentionné précédemment, dans les cas de surdétermination duplicative ou symétrique comme définis dans les définitions 5.10, 5.11 et 5.12, les deux chemins causaux ont abouti. Il est donc aisé de relier factuellement les évènements qui appartiennent aux deux chemins avec la conséquence. Dans l'exemple 3.6 nous savons que la victime est morte par empoisonnement, les deux assassins ayant versé une dose létale de poison dans la boisson de la victime peuvent aisément être identifiés comme des causes effectives de cette mort. Pourtant, aucune des deux actions n'a fait réellement une différence, il suffit qu'elles aient contribué à travers un chemin causal. Cela correspond bien avec les résultats observés par Moore [2019] lorsqu'il a étudié les jugements de nombreux cas en droit.
La situation est différente lorsqu'il est question d'omission. En faisant la synthèse de ce qu'observe Moore [2019], nous constatons que dans la plupart des cas, du moment où il y a surdétermination et donc que les éléments ne font plus la différence individuellement, l'omission n'est plus sujette à entraîner de la responsabilité. C'est ce qu'explique Abrams

[2022] lorsqu'il écrit « With omissions, both conceptually, and as a matter of law, the very meaning of attributing causal status to an omission is to attribute some difference-making feature to the omission : the omitted act would have had a relevant impact [...] ». Pour appuyer ces propos il cite l'exemple du cas Piqua v. Morris où la négligence dans l'entretien d'un barrage a été considérée causalement sans importance étant donné qu'au vu de l'inondation massive, l'entretien n'aurait de toute façon pas fait de différence. BRAHAM et VAN HEES [2012] présentent une expérience de pensée qui après modification permet de faire ressortir cette intuition.

Exemple 7.7 [exemple de Frankfurt modifié]**.** *Un agent omet de faire une action, disons sauver un enfant qui se noie. Toutefois, s'il avait décidé de la faire, l'activation d'un dispositif implanté dans son cerveau par un autre agent l'en aurait empêché. Omission de sa part ou pas, le résultat est le même dans tous les cas. Considérer qu'il peut en être responsable est moins intuitif par rapport au cas où il n'est pas fait mention de ce dispositif.*

Pour savoir si une conséquence peut être attribuée à une non occurrence, il est donc nécessaire d'imaginer que l'occurrence avait lieu et essayer de trouver une chaîne causale qui empêche cette conséquence d'avoir lieu. Dans la description de cette procédure WRIGHT [1985] fait clairement apparaître que le raisonnement réalisé est un raisonnement hypothétique : « Thus, in order to determine whether an omission was a cause of an injury, it is necessary to conduct a hypothetical inquiry. The omitted act must be hypothetically supplied, and a hypothetical causal sequence [...] must be constructed and traced to determine whether it would have prevented the occurrence of the injury ».
Si l'idée de la procédure générale à suivre est très claire, l'application du raisonnement hypothétique qu'elle requiert ne l'est pas. En effet, celle-ci semble dépendre de facteurs en dehors de la causalité. C'est ce que montre ABRAMS [2022] en présentant des variantes de l'exemple 7.1. Imaginons une première variante où juste avant de s'engager dans l'intersection, le conducteur réalise que les freins ne fonctionnent pas, raison pour laquelle il ne prend pas la peine d'essayer de freiner lorsqu'il s'aperçoit qu'il va percuter le piéton. Il semble alors moins intuitif d'attribuer à son omission de freiner la conséquence, et plus intuitif de l'attribuer à l'omission de l'entreprise de location. Maintenant, considérons le cas où le conducteur avait pour intention de percuter le piéton car il voulait se venger, raison pour laquelle il ne prend pas la peine d'essayer de freiner lorsqu'il s'aperçoit qu'il va percuter le piéton. Il semble alors plus intuitif d'attribuer à son omission de freiner la conséquence, et moins intuitif de l'attribuer à l'omission de l'entreprise de location. Le fait qu'il sache que les freins étaient défectueux ou qu'il ait l'intention de percuter le piéton sont des états mentaux de l'agent, ceux-ci ne changent rien à l'information factuelle du problème et donc aux lois causales qui ont été à l'œuvre. Comme nous allons le voir dans la section suivante, si ces informations modifient l'intuition que l'agent pouvait faire la différence c'est parce qu'elles ont un impact sur le raisonnement hypothétique que nous nous autorisons à réaliser.

7.1.3 Tout raisonnement hypothétique est normatif

Dans cette section nous montrons que c'est dans le choix des mondes qu'il est possible d'envisager dans le raisonnement hypothétique que repose la décision de quelles conséquences pourront être attribuées à chaque non occurrence. S'agissant d'un choix subjectif dépendant d'un grand nombre de facteurs en dehors de la causalité, nous montrons que l'omission est essentiellement une question de responsabilité.

Tout raisonnement hypothétique concernant l'omission a un aspect normatif. Commençons par montrer cela pour le cas simple où a priori il n'y a pas de surdétermination. Nous parlons ici des cas comme celui du maître nageur qui omet de sauver une personne qui se noie ou celui du jardinier qui omet d'arroser la vigne dans l'exemple 3.13. Des considérations normatives façonnent l'intuition derrière l'envie d'attribuer ou non des conséquences à une non occurrence d'évènement. Si c'est aux omissions du maître nageur et du jardinier que nous voulons attribuer les conséquences, et non pas aux autres agents dans la situation, c'est parce que par la description du cas nous considérons que c'était leur devoir et non pas celui des autres individus. De ce fait, les raisonnements hypothétiques consistant à se demander ce qui se serait passé si le maître nageur ou le jardinier étaient intervenus est accepté. En revanche, le raisonnement hypothétique consistant à se demander ce qui se serait passé si un passant quelconque était intervenu est plus difficilement justifiable. Dans le premier cas, le raisonnement hypothétique consiste à se placer dans un monde où personne d'autre que le maître nageur ou le jardinier n'agit. Étant donné leur fonction, nous supposons que leurs capacités leur permettent de sauver l'individu en détresse et la vigne respectivement. Leur omission a donc fait la différence, nous voulons donc leur attribuer la conséquence. Dans le deuxième cas, le raisonnement hypothétique consiste à se placer dans un monde où personne d'autre que le passant quelconque n'agit. Cette option est plus difficile à accepter, pourquoi lui et pas un autre ? En fin de compte, si nous le considérons lui il faudrait considérer tous les individus « quelconques ». Ce cas là nous mène à une situation où la responsabilité se dilue entre de nombreux acteurs. À moins de pouvoir en démarquer certains, leur omission individuelle n'a pas réellement fait la différence. Nous ne voulons donc pas attribuer la conséquence à ces non occurrences. Supposons qu'arbitrairement nous le prenions lui. Qu'est ce qui nous dit qu'il a les capacités d'empêcher la conséquence ? Il se peut qu'il ne sache pas nager par exemple. Cela nous mène dans un monde trop incertain pour pouvoir affirmer que son omission a réellement fait la différence. Nous ne voulons donc pas attribuer la conséquence à cette non occurrence. Nous avons vu dans la section 3.2.2.2 et dans ce chapitre que les connaissances sur le monde, comme par exemple les états mentaux des agents et leurs capacités, ont une influence sur la volonté d'attribuer ou non des conséquences à une non occurrence. Les exemples ci-dessus montrent que ces aspects influencent notre opinion sur cette attribution du fait qu'ils influencent les mondes que nous nous permettons d'explorer dans le raisonnement hypothétique nécessaire pour traiter ces cas. C'est ce qu'explique très clairement ABRAMS [2022] :

> The appropriate question is not "what did the gardener's not watering the plant cause?" The answer to that question is : nothing. Rather, we ask : had the gardener performed his duty (i.e., watered the plant) what would have happened (what would have been caused) ? The plaintiff has a right to be in that world (in that position). When she sues the gardener, she is vindicating a right to be where, had he fulfilled his duty, she (causally) would be. The law determines, normatively, what world Plaintiff has a right to be in, and determines that, had Defendant not breached, he would have brought about (caused her to be in) that world. This explains why liability attaches to the gardener and not to third parties who also failed to water the plant. The gardener's liability in this case is not, strictly speaking, causal (it is not for what he caused) but for what he had a duty to bring about.

Ce même mécanisme est à l'œuvre dans les cas où il y a surdétermination. Dans la mesure où toute la complexité de ces cas repose sur le fait qu'ils font disparaître la nécessité, les choix sur les mondes envisageables dans le raisonnement hypothétique sont plus compliqués et leur importance devient plus évidente. Nous parlons ici des cas comme ceux des exemples 7.1 à 7.6 où il peut être considéré qu'au moins deux agents ont failli à leur devoir. Prenons l'exemple 7.1 pour montrer cela. Deux omission sont considérées : celle du conducteur qui n'a pas freiné et celle de l'entreprise de location qui n'a pas fait les vérifications et la maintenance nécessaire. La procédure classique pour savoir si une des ces omissions peut se voir attribuer le préjudice causé au piéton percuté est d'imaginer un monde où ne pas avoir fait l'omission aurait abouti à un résultat différent. Le monde dans lequel se placer n'est pas simple à déterminer. Contrairement au cas du jardinier où il était acceptable étant donné son devoir de considérer le cas ou seul lui agissait, il n'est pas évident, tel quel, de dire qu'il est possible d'évaluer l'action de freiner en supposant que les freins fonctionnent, ou d'évaluer l'action de faire la maintenance en supposant que le conducteur freine. Pourquoi pourrions-nous choisir un monde plutôt qu'un autre ? Il semblerait que dans la forme classique de l'exemple, soit aucune des omissions ne peut se voir imputée le préjudice, soit les deux le peuvent. Considérons maintenant une première variante où juste avant de s'engager dans l'intersection, le conducteur réalise que les freins ne fonctionnent pas, raison pour laquelle il ne prend pas la peine d'essayer de freiner lorsqu'il s'aperçoit qu'il va percuter le piéton. Nous avons mentionné dans le section 7.1.2 qu'il semblait alors moins intuitif d'attribuer à son omission de freiner la conséquence, et plus intuitif de l'attribuer à l'omission de l'entreprise de location. En effet, dans ce cas il pourrait être défendu qu'il existe une raison pour se placer dans des mondes où le conducteur freine et de comparer deux mondes hypothétiques : celui où la maintenance est faite et celui où elle ne l'est pas. Dans le premier le préjudice ne devrait pas avoir lieu, alors que dans le deuxième oui. L'omission de l'agence de location était en mesure de faire la différence et donc il semble acceptable de lui imputer le préjudice. Maintenant, considérons la variante où le conducteur avait pour intention de percuter le piéton car il voulait se venger, raison pour laquelle il ne prend pas la peine d'essayer de freiner lorsqu'il s'aperçoit qu'il va percuter le piéton. Nous avons mentionné dans le section 7.1.2 qu'il semblait alors plus intuitif d'attribuer à son omission de freiner la conséquence, et moins intuitif de l'attribuer à l'omission de l'entreprise de location. En effet, dans ce cas il pourrait être défendu qu'il existe une raison pour se placer dans des mondes où la maintenance est faite et de comparer deux mondes hypothétiques : celui où le conducteur freine et celui où il ne freine pas. Dans le premier le préjudice ne devrait pas avoir lieu, alors que dans le deuxième oui. L'omission du conducteur était en mesure de faire la différence et donc il semble acceptable de lui imputer le préjudice. Voilà comment les états mentaux de l'agent, influencent le raisonnement.
Il en est de même pour tous les autres exemples de surdétermination. La question qui se pose est dans quel monde il est possible de se placer pour évaluer si une omission avait pu faire la différence. Faut-il comparer le monde tel qu'il a été avec le monde où la seule chose qui change est l'omission qui est évaluée ? Cela correspondrait au cas où seul l'agent évalué accomplit son devoir. Dans l'exemple 7.1, si nous évaluons l'omission du conducteur, cela reviendrait à savoir si son omission fait la différence en prenant uniquement des mondes où la maintenance n'a pas été faite. Faut-il comparer le monde tel qu'il a été avec le monde où toutes les omissions en surdétermination changent ? Cela pourrait correspondre à deux cas. Le premier est celui où tous les agents accomplissent leur devoir. Dans l'exemple 7.1,

si nous évaluons l'omission du conducteur, cela reviendrait à savoir si son omission fait la différence en prenant des mondes où la maintenance a été faite. Le deuxième est celui où toutes les omissions changent, même celles des agents qui n'en sont pas tenus par leur devoir. Dans l'exemple avec le maître nageur, si nous évaluons son omission, cela reviendrait à savoir si son omission fait la différence en prenant des mondes où toutes les personnes ayant la possibilité d'agir pour empêcher la noyade le font. Faut-il comparer le monde tel qu'il a été avec un autre monde que ceux évoqués précédemment, un point intermédiaire ? Cette décision est purement normative. Pour l'omission il n'est pas possible de s'appuyer uniquement sur des aspects factuels comme dans la causalité positive. L'omission est une question de responsabilité.

Supposons que nous acceptions de sortir du cadre purement causal, et donc factuel, pour essayer de trouver une définition commune de ce qu'est attribuer une conséquence à une non occurrence, même si celle-ci fait intervenir des notions propres à la responsabilité. Pouvons nous trouver un point d'entente ? À première vue cela semble compliqué. Reprenons le cas du maître nageur qui omet de sauver l'individu en détresse, cas qui pour rappel est considéré comme simple puisqu'il n'est pas considéré comme un cas de surdétermination dans sa version classique. Dans cette version, il semble communément admis qu'étant donné sa position, c'était son devoir de sauver l'individu. Faisons varier ce cas, sans pour autant en faire un cas de surdétermination. Premièrement, imaginons que nous découvrions que le maître nageur ne sait pas nager. Certes, il est quelque part blâmable d'avoir menti lors de son embauche, ou en tout cas d'avoir omis de préciser cette information ; mais pour ce qui nous concerne, s'il n'avait pas les capacités de sauver l'individu, le lien entre le préjudice et son omission d'être allé sauver l'individu apparaît plus faible. À ce moment là, une nouvelle omission semble devoir être prise en compte, celle du gérant de la piscine qui ne s'est pas assuré que le maître nageur qu'il embauchait savait nager. Puis, il est également possible de considérer que la personne dont le devoir était de sauver la personne n'étant pas en capacité de le faire, les agents assistant à la scène héritent d'une partie de ce devoir. Sans information supplémentaire, il est difficile de faire une distinction entre ces agents. C'est là où intervient la deuxième variante. Imaginons maintenant que tous les autres agents présents savent nager, mais que l'un deux est la mère de l'individu qui se noie, individu qui s'avère être un mineur. Dans ce cas là, il serait possible de défendre l'idée qu'étant donné sa position, elle était plus contrainte par le devoir que les autres. Et ainsi de suite.
Nous pouvons faire de même avec le cas où A décide de frapper B et que C intervient et bloque le coup. Nous avions mentionné qu'il n'est pas évident de dire que si C n'était pas intervenu alors il serait responsable du fait que B ait été frappé. Mais cela dépend à nouveau fortement de l'information que nous avons. Préciser que C est le garde du corps de A peut modifier notre avis, tout comme préciser que A et B sont des enfants et que C est le père de A. Bref, de nombreuses considérations supplémentaires pourraient venir modifier notre intuition.

L'élément qui s'avère décisif semble être le devoir des agents. Malheureusement, déterminer cela n'est possible qu'en adoptant un point de vue spécifique. En l'occurrence, ABRAMS [2022] indique qu'en droit le devoir de ne pas produire de mal par l'action est prédominant par rapport au devoir de ne pas produire de mal par l'omission, lui même prédominant sur le devoir d'agir pour produire le bien ou pour empêcher le mal. Toutefois, il s'agit là d'une vision anglo-saxonne, cela peut varier en fonction de chaque pays.
Même s'il existait un consensus en droit, cela n'est pas le cas en éthique. Chaque théorie

morale a une vision particulière de quelles actions doivent être faites par les agents dans ces cas. Comme nous l'avons expliqué dans le chapitre 1, le but même d'une théorie morale est de déterminer ce qui est requis, optionnel ou interdit. Autrement dit, chaque théorie morale vise à déterminer le statut déontique des actions et donc la façon dont l'agent doit agir, son devoir. Supposer qu'il est possible de trouver une vision commune suppose que les théories morales sont en fin de compte toutes d'accord. Il n'est pas nécessaire de rentrer dans le détail du fait que cela est absurde, les seuls exemples utilisés dans le chapitre 1 et 4 suffisent à montrer que ce n'est pas le cas. ABRAMS [2022] apporte une autre preuve que selon la famille de théories morales la conception du devoir peut changer lorsqu'il dit : « My analysis here differs from MOORE [2009], who solves omission liability by appealing to a general background consequentialist obligation, which entails a (weaker, non-deontic) duty to prevent harm ». Il n'est donc pas possible de trouver une vision qui convienne à toutes les théories morales, même si nous acceptions de sortir du cadre purement causal en faisant un pas vers la responsabilité. Pour cette raison, proposer une définition de comment prendre en compte l'omission est en dehors du cadre de cette contribution.

La conclusion de cette section est la même que celle que fait ABRAMS [2022], nous adoptons l'idée que l'omission, et plus généralement la non occurrence d'évènements, n'a pas de statut causal : « I'll be following a tradition in philosophy that denies the causal status of omissions, and in legal philosophy that takes the act/omission distinction seriously in understanding causation in the law. This tradition has been defended recently, most notably by Michael Moore ». Attribuer des conséquences à une non occurrence d'évènement étant une question purement normative, si les non occurrences d'évènements avaient un statut causal, la causalité serait normative et donc contraire à ce que nous avons défendu dans la section 3.1.2.2. De plus, comme nous l'avons montré dans le chapitre 5 et 6, la causalité positive peut être déterminée de façon tout à fait factuelle, sans intervention d'aspects subjectifs propres à la responsabilité. Qui plus est, comme nous l'avons précisé pour quelques exemples, la causalité positive est toujours présente dans les exemples et elle est toujours aussi claire. Dans les exemples 7.1 à 7.6, il existe des chaînes causales positives menant aux conséquences. En l'occurrence, dans l'exemple 7.1 la voiture doit être en mouvement pour percuter le piéton. Ce mouvement est causé par le conducteur. Il est clair que le conducteur est une cause effective de l'accident. La prise en compte des non occurrences d'évènements n'affecte en rien l'enquête causale classique, elle n'est en réalité qu'une considération supplémentaire qui peut jouer un rôle dans les autres étapes permettant de déterminer la responsabilité. Pour rappel ces étapes ont été décrites par WRIGHT [1985] et ont été présentées dans la section 3.1.2.2. L'omission de l'agence de location peut venir atténuer ou éliminer la responsabilité du conducteur, elle ne change en rien la causalité effective du problème.

7.2 Conséquences négatives, ou empêcher : à mi chemin entre causalité effective et responsabilité

Dans cette section nous abordons la problématique de déterminer les causes pour lesquelles un évènement ne s'est pas produit. S'intéresser à ces causes fait sens en éthique computationnelle, dans certaines théories morales nous voulons être en mesure de pouvoir raisonner sur le fait qu'un agent ait agi pour empêcher le mal ou le bien de se produire. Comme montré dans la section 3.2.2.1, la notion d'empêcher est étudiée dans le domaine de la causalité également. Il est important à ce niveau de distinguer deux problématiques dis-

tinctes qui apparaissent lorsqu'il est question de conséquences négatives. Chacune des problématiques correspond à un des types de relations causales dans dans un STEE que nous avons identifiées dans la section 5.1.2. Dans le cas des $\mathscr{F} - causes$ de la définition 5.4, s'intéresser aux conséquences négatives revient à s'intéresser au cas où l'occurrence d'évènement (e,t) est la cause et $(\neg f, t_\psi)$ la conséquence, où $f \in \mathbb{F}$. Dans le cas des causes effectives de la définition 5.5, s'intéresser aux conséquences négatives revient à s'intéresser au cas où l'occurrence d'évènement (e,t) est la cause et la non occurrence d'évènement $(\overline{e_\psi}, t_\psi)$ est la conséquence, où $e_\psi \in \mathbb{E}$. Cette distinction n'est possible que par la clarification apportée par l'utilisation d'un STEE. Cette distinction n'est pas anodine.
Dès les travaux de MILL [1843], il est considéré que la causalité ne peut être suffisante que si, en plus des causes positives, les conditions qui n'étaient pas vraies et dont l'absence était une condition nécessaire à la survenue du résultat sont prises en compte également. Ainsi, les évènements étant des causes de leur absence sont également des causes du résultat. En l'occurrence, dans l'exemple 3.1 sur la pollution d'un lac repris et représenté dans $\mathscr{S}_c$ dans l'exemple 6.1, les conditions de déclenchement pour le déversement des eaux usées qui entraîne que le seuil soit atteint sont $tri(dev_o) = o_e \vee (o_m \wedge se_{hs})$, où se_{hs} indique que la station d'épuration est hors service. Dans ces conditions se_{hs}, pourrait très bien être remplacé par $\neg se_{es}$, où se_{es} indiquerait que la station d'épuration est en service. Il se trouve que $\neg se_{es}$ est ce dont MILL [1843] parle lorsqu'il mentionne les conditions qui n'étaient pas vraies et dont l'absence était une condition nécessaire à la survenue du résultat. En travaillant sur des littéraux de fluents, notre approche de causalité positive présentée dans le chapitre 6 gère déjà le cas relatif aux $\mathscr{F} - causes$ de la définition 5.4.
Le deuxième cas est celui relatif aux causes effectives de la définition 5.5, i.e. lorsque la conséquence est la non occurrence d'évènement $(\overline{e_\psi}, t_\psi)$. Dans cette section nous étudions ce deuxième cas. Comme nous l'avons montré dans le section 3.2.2.1, il n'est pas si simple à traiter que cela.

Contrairement à l'omission, et plus généralement aux cas avec des causes négatives, les cas qui impliquent des conséquences négatives peuvent être traités factuellement. L'approche proposée dans le chapitre 6 en est la preuve pour le cas des $\mathscr{F} - causes$ de la définition 5.4. Si nous étendons ce raisonnement, les causes de $(\overline{e_\psi}, t_\psi)$ sont les occurrences d'évènements étant des causes : soit de la présence d'une des conditions dont l'absence était nécessaire à la suffisance d'un support, soit de l'absence d'une des conditions dont la présence était nécessaire à la suffisance d'un support. Une typologie des cas de surdétermination comme celle présentée dans le chapitre 5 peut être construite pour cette forme de négation dans la conséquence. Cette analyse détaillée montre qu'il existe une différence entre les cas de causalité positive et cette forme de négation dans la conséquence. Des types de surdétermination possibles en causalité positive, ne le sont pas dans le deuxième cas.

Malgré l'existence de cette solution factuelle, celle-ci ne correspond pas tout à fait à la notion d'empêcher qui vient intuitivement à l'esprit, en tout cas dans un type de cas en particulier. L'hypothèse que nous défendons est qu'à nouveau, le conflit peut s'expliquer par la confusion courante entre causalité et responsabilité que dénonce WRIGHT [1985]. Les notions de responsabilité sont introduites par le fait que dans ces cas, nous souhaitons attribuer des conséquences à une non occurrence d'évènement. Prenons l'exemple 3.11 où A intercepte une balle en direction de B avant qu'elle n'atteigne une fenêtre, surpassant la réactivité de B et évitant ainsi que la fenêtre soit brisée. Deux niveaux de raisonnement peuvent être identifiés : le premier consiste à déterminer ce qui a empêché la balle de poursuivre son

chemin, alors que le deuxième consiste à déterminer ce qui a empêché que la fenêtre soit brisée. Le premier raisonnement est simple et factuel, nous savons que c'est A qui a intercepté la balle. Le deuxième est plus compliqué, il requiert un raisonnement hypothétique car il demande de relier le fait que la balle n'ait pas pu continuer son chemin avec le fait que la fenêtre n'ait pas été brisée. Dit autrement, il demande d'attribuer des conséquences à une non occurrence d'évènement. Il s'agit là du cas traité dans la section précédente qui nous l'avons vu est intrinsèquement normatif.

Cette section est divisée en trois parties. La section 7.2.1 représente dans $\mathscr{S}_c$ l'approche factuelle pour traiter la non occurrence d'un évènement et étudie sa sensibilité à la surdétermination, comme pour la causalité positive dans le section 6.2.4. Pour cela, elle modélise dans $\mathscr{S}_s$ la surdétermination dans le cas de conséquences négatives et établit une typologie en suivant la même procédure que dans le chapitre 5. La section 7.2.2 montre que pour certains cas, l'approche factuelle donne des résultats qui ne correspondent pas à ce qui semble être compris avec la notion d'empêcher. Finalement, la section 7.2.3 explique que la notion d'empêcher peut être décomposée en deux niveaux de raisonnement, un factuel et un normatif. Dans les cas qui incombent au premier niveau, l'approche factuelle que nous proposons est satisfaisante. Les cas qui posent problème font intervenir le deuxième. Cette section établit que l'approche causale factuelle n'est pas satisfaisante due au fait que ces problèmes sont en dehors du cadre causal, ce sont des problèmes de responsabilité qui demandent d'attribuer des conséquences à la non occurrence d'évènements.

7.2.1 L'approche factuelle pour traiter les conséquences négatives

Dans cette section nous représentons dans $\mathscr{S}_c$ l'approche factuelle pour traiter la non occurrence d'un évènement et nous étudions sa sensibilité à la surdétermination, comme pour la causalité positive dans le section 6.2.4. Pour cela, nous modélisons la surdétermination dans le cas de conséquences négatives et établissons une typologie en suivant la même procédure que dans le chapitre 5. Pour faciliter la distinction entre cette surdétermination et celle déjà étudiée, nous parlons ici de surdétermination négative.

La définition permettant de traiter factuellement le cas où l'occurrence d'évènement (e,t) est la cause et la non occurrence d'évènement $(\overline{e_\psi}, t_\psi)$ est la conséquence découle des définitions proposées dans le chapitre 6. Plus exactement, la relation causale entre (e,t) et $(\overline{e_\psi}, t_\psi)$ est définie en s'appuyant sur les NESS-causes. L'occurrence d'un premier évènement est considérée comme une cause effective de la non occurrence d'un second si l'occurrence du premier est une NESS-cause de la négation des conditions de déclenchement du deuxième.

Définition 7.1 [*Causes effectives de non occurrence*]**.** *Étant donné un cadre causal* χ*, une occurrence d'évènement* (e,t) *et une non occurrence d'évènement* $(\overline{e_\psi}, t_\psi)$*,* (e,t) *est une cause effective de* $(\overline{e_\psi}, t_\psi)$*, relation que nous notons* $(e,t) \rightsquigarrow (\overline{e_\psi}, t_\psi)$*, ssi* (e,t) *est une NESS-cause de* $(\neg tri(e_\psi), t_\psi)$*.*

Pour être en mesure de construire la typologie pour la surdétermination négative, il est nécessaire que nous proposions une définition formelle de ce qu'est la surdétermination négative et des chemins causaux. Comme dans le chapitre 5, pour que cette contribution puisse bénéficier au plus grand nombre, nous nous plaçons dans $\mathscr{S}_s$.
Commençons par définir ce qu'est la surdétermination négative. Comme mentionné dans

la section 5.1, pour parler de surdétermination il faut pouvoir raisonner de façon contrefactuelle. Pour ce faire, il est nécessaire d'imaginer des traces différents dans $\mathscr{S}_s$. Nous imaginons des traces alternatives en modifiant la politique π_s. Ces politiques alternatives peuvent être vues comme des politiques contrefactuelles. Pour rappel, pour construire ces politiques contrefactuelles nous avons défini l'opération :

$$\pi_s \setminus \mathrm{E} \stackrel{\text{def}}{=} \forall \mathrm{S}, \forall t,\ \pi_s(\mathrm{S}, t) = \pi_s(\mathrm{S}, t) \setminus \mathrm{E}.$$

Définition 7.2 [*Surdétermination négative*]**.** *Soit $\chi = (\kappa_s, \pi_s)$ le cadre causal, $(a^1, t^1), (a^2, t^2)$ deux occurrences d'évènements et $(\overline{e_\psi}, t_\psi)$ une non occurrence d'évènement. Étant donné trois cadres causaux contrefactuels :*

$$\chi_{\mathrm{I}}^1 = (\kappa_s, \pi_s \setminus \{a^2\}), \qquad \chi_{\mathrm{I}}^2 = (\kappa_s, \pi_s \setminus \{a^1\}), \qquad \chi_- = (\kappa_s, \pi_s \setminus \{a^1, a^2\}),$$

où χ_{I}^1 et χ_{I}^2 sont dits des cadres causaux Individuels, nous sommes dans un cas de surdétermination négative entre a^1 et a^2 dans le cadre causal χ si sont vérifiées :

- $e_\psi \notin \mathrm{E}^{\chi}(t_\psi)$;
- $e_\psi \in \mathrm{E}^{\chi_-}(t_\psi)$;
- $e_\psi \notin \mathrm{E}^{\chi_{\mathrm{I}}^1}(t_\psi)$;
- $e_\psi \notin \mathrm{E}^{\chi_{\mathrm{I}}^2}(t_\psi)$.

Dans la section 3.1.2.1, nous avons vu que les travaux existants font tous référence d'une façon ou d'une autre au concept de « chemin causal » lorsqu'ils parlent des différents cas de surdétermination. Ce concept paraît donc être central pour définir ces cas. En vue de clarifier les différents cas de surdétermination négative, nous définissons ce qu'est un chemin causal vers une non occurrence dans $\mathscr{S}_s$.

Définition 7.3 [*Chemin causal vers une non occurrence $\overline{\omega}$*]**.** *Étant donné un cadre causal χ et un évènement $e_\psi \notin \mathrm{E}^{\chi}(t_\psi)$, avec $\psi = pre(e_\psi)$, la séquence d'occurrences $\overline{\omega} = (e_n, t_n), \ldots, (e_1, t_1)$ est un chemin causal reliant (e_n, t_n) à $(\overline{e_\psi}, t_\psi)$ ssi :*

- *(e_n, t_n) est une $\mathscr{F}$-cause de $(pre(e_{n-1}), t_{n-1})$;*
- *…;*
- *(e_2, t_2) est une $\mathscr{F}$-cause de $(pre(e_1), t_1)$;*
- *(e_1, t_1) est une $\mathscr{F}$-cause de $(\neg\psi, t_\psi)$.*

Dit autrement, un chemin causal entre une occurrence d'évènement et une non occurrence d'évènement est une séquence d'occurrences d'évènements où chaque évènement est une $\mathscr{F}$-cause du suivant, si bien qu'il contribue au déclenchement du suivant, et le dernier est une $\mathscr{F}$-cause de la négation des conditions de déclenchement de l'évènement qui ne se produit pas.

Ce que nous appelons surdétermination négative et chemin causal vers une non occurrence étant maintenant définis formellement, nous passons à notre proposition de typologie des cas de surdétermination négative. Le tableau 7.2 indique le type de cas de surdétermination qui correspond à certaines conditions causales, données par l'entête des colonnes, auxquelles nous ajoutons certaines conditions temporelles, données par l'entête des lignes. Il s'agit là d'une typologie qui permet de classifier clairement les cas de surdétermination négative dans cinq catégories distinctes, une de moins que pour le cas de surdétermination positive. Cette typologie montre qu'il existe bien une différence entre la

causalité positive et ce cas de négation dans la conséquence.
Le processus menant à cette structure est le même que celui du chapitre 5. Nous ne détaillerons que les grandes lignes du raisonnement dans lesquelles nous utilisons les nouvelles notations correspondant à notre cas avec négation.

Algèbre des intervalles d'Allen		$\overline{\Omega}^1 = \{\overline{\omega}^1\}, \overline{\Omega}^2 = \emptyset$		$\overline{\Omega}^1 = \{\overline{\omega}^1\}, \overline{\Omega}^2 = \{\overline{\omega}^2\}$	
		$(e_i^1, t_i^1) \rightarrow \left(\neg pre(e_j^2), t_j^2\right)$	$(\overline{e_\psi}, t_\psi) \rightarrow \left(\neg pre(e_j^2), t_j^2\right)$	$W^1 \neq W^2$	$W^1 = W^2$
$\overline{\omega}^1$ égal à $\overline{\omega}^2$		Préemptive précoce	**	Duplicative synchrone	Symétrique
$\overline{\omega}^1$ termine $\overline{\omega}^2$		Préemptive précoce	**	Duplicative synchrone	Symétrique
$\overline{\omega}^1$ terminé par $\overline{\omega}^2$		Préemptive précoce	**	Duplicative synchrone	Symétrique
$\overline{\omega}^1$ chevauche $\overline{\omega}^2$		Préemptive précoce	**	Duplicative asynchrone	Imitative
$\overline{\omega}^1$ chevauché par $\overline{\omega}^2$		Préemptive précoce	**	***	***
$\overline{\omega}^1$ démarre $\overline{\omega}^2$		Préemptive précoce	**	Duplicative asynchrone	Imitative
$\overline{\omega}^1$ démarré par $\overline{\omega}^2$		Préemptive précoce	**	***	***
$\overline{\omega}^1$ se déroule pendant $\overline{\omega}^2$		Préemptive précoce	**	Duplicative asynchrone	Imitative
$\overline{\omega}^1$ englobe le déroulé de $\overline{\omega}^2$		Préemptive précoce	**	***	***
$\overline{\omega}^1$ rencontre $\overline{\omega}^2$		Préemptive précoce	**	Duplicative asynchrone	Imitative
$\overline{\omega}^1$ est rencontré par $\overline{\omega}^2$		*	**	***	***
$\overline{\omega}^1$ se déroule avant $\overline{\omega}^2$		Préemptive précoce	**	Duplicative asynchrone	Imitative
$\overline{\omega}^1$ se déroule après $\overline{\omega}^2$		*	**	***	***

TABLEAU 7.2 – Typologie formelle des cas de surdétermination négative prenant en compte toutes les relations temporelles possibles entre deux chemins causaux. (*) Incohérence entre la relation causale à l'origine de l'interruption de $\overline{\omega}^2$ et la relation temporelle entre les intervalles. (**) Incohérence dans la relation causale, une telle relation ne peut pas exister. (***) Incohérence entre l'hypothèse que $\overline{\omega}^1$ est toujours le premier chemin causal à aboutir et la relation temporelle entre les intervalles.

Nous nous plaçons dans le cadre causal χ et nous nous considérons dans un cas de surdétermination. D'après la définition 7.2, nous avons donc un évènement $e_\psi \notin E^\chi(t_\psi)$ avec $pre(e_\psi) = \psi$, deux occurrences d'évènements $(e_m^1, t_m^1), (e_n^2, t_n^2)$ et trois cadres causaux contrefactuels $\chi_I^1 = (\kappa_s, \pi_s \setminus \{e_n^2\})$, $\chi_I^2 = (\kappa_s, \pi_s \setminus \{e_m^1\})$ et $\chi_- = (\kappa_s, \pi_s \setminus \{e_m^1, e_n^2\})$. Étant donné un couple d'indices $(i,k) \in \{(1,m),(2,n)\}$, l'ensemble de tous les chemins causaux qui relient (e_k^i, t_k^i) à $(\overline{e_\psi}, t_\psi)$ dans χ_I^i est noté $\overline{\Omega}_I^i$. Si $i = 1$, nous sommes dans le cadre χ_I^1 où nous supposons que (e_n^2, t_n^2) ne s'est pas produit puisqu'il est retiré de la politique, et donc nous étudions les chemins causaux partant de (e_m^1, t_m^1) individuellement. Si $i = 2$, nous sommes dans le cadre χ_I^2 où nous supposons que (e_m^1, t_m^1) ne s'est pas produit puisqu'il est retiré de la politique, et donc nous étudions les chemins causaux partant de (e_n^2, t_n^2) individuellement. De la même façon, l'ensemble de tous les chemins causaux qui relient (e_k^i, t_k^i) à $(\overline{e_\psi}, t_\psi)$ dans χ est noté $\overline{\Omega}^i$. Dans ce cas là, nous sommes dans le cadre χ où nous avons aussi bien (e_m^1, t_m^1), que (e_n^2, t_n^2).
Par souci de concision et de clarté, nous restreignons le cadre de notre analyse en faisant les deux mêmes hypothèses que dans le chapitre 5 : il existe un unique chemin causal reliant une occurrence d'évènement et une non occurrence d'évènement, i.e. $|\overline{\Omega}_I^1| = |\overline{\Omega}_I^2| = 1$, et aucun chemin causal n'est créé de l'interaction entre chemins dans $\overline{\Omega}_I^1$ et $\overline{\Omega}_I^2$, i.e. $\overline{\Omega}^i \setminus \overline{\Omega}_I^i = \emptyset$. Ces hypothèses posées, nous savons que $\overline{\Omega}_I^1 = \overline{\Omega}^1 = \{\overline{\omega}^1\}$, $\overline{\Omega}_I^2 = \{\overline{\omega}^2\}$ et soit $\overline{\Omega}^2 = \{\overline{\omega}^2\}$, ou $\overline{\Omega}^2 = \emptyset$. Cette distinction entre les deux cas possibles pour $\overline{\Omega}^2$ est cruciale car elle permet de différencier les cas de surdétermination préemptive des autres. La typologie que nous présentons pouvant être organisée comme un tableau à double entrée, chacun des deux cas possibles pour $\overline{\Omega}^2$ correspond dans notre typologie à une première différenciation en colonnes.

Une étude exhaustive des relations temporelles entre chemins causaux est indispensable. Pour la réaliser nous utilisons l'algèbre des intervalles d'ALLEN [1983]. Les chemins causaux $\overline{\omega}^i$ sont représentés comme des intervalles temporels qui commencent à t_k^i et qui finissent à t_1^i, temps correspondant respectivement à la première occurrence (e_k^i, t_k^i) et à la dernière occurrence (e_1^i, t_1^i) du chemin causal. Quelles que soient les combinaisons de $\overline{\Omega}^1$ et $\overline{\Omega}^2$ testées, nous les confrontons toutes aux treize relations possibles entre deux intervalles, comme définies par l'algèbre des intervalles d'Allen, et regardons si cela à une quelconque influence sur le type de cas de surdétermination négative. Dans la représentation de notre typologie dans le tableau 7.2, chacune des treize relations d'Allen correspond à une ligne du tableau. Comme dans le chapitre 5, les noms des intervalles d'Allen ne doivent pas être compris causalement. Les informations causales sont données par les entêtes de colonne.

La différence entre causalité positive et causalité dans ce cas de négation dans la conséquence apparaît dans la première grande colonne du tableau 7.2. Il s'agit du cas où $\overline{\Omega}^2 = \varnothing$. Dans cette colonne nous pouvons déduire que l'occurrence (e_m^1, t_m^1) a quelque chose à voir avec l'interruption de $\overline{\omega}^2$. Nous savons que dans le cadre causal χ_I^2, $\overline{\omega}^2$ est bien un chemin causal, tandis que dans χ il ne l'est plus. La seule différence entre χ_I^2 et χ étant la présence dans le deuxième de (e_m^1, t_m^1), il est possible de déduire que cette occurrence a quelque chose à voir avec l'interruption de $\overline{\omega}^2$. Interrompre un chemin causal revient à empêcher le déclenchement d'une des occurrences composant ce chemin.
Dans la typologie des cas de surdétermination positive présentée par le tableau 5.1 et 5.2, deux cas sont possibles. Soit ce qui a interrompu ω^2 est l'occurrence de (e_ψ, t_ψ), due au fait que ω^1 a abouti en premier, soit ce qui a interrompu ω^2 est une des occurrences composant le chemin causal ω^1. Le premier cas de figure s'exprime formellement :

$$\exists (e_j^2, t_j^2) \in \omega^2, (e_\psi, t_\psi) \rightarrow \left(\neg pre(e_j^2), t_j^2\right).$$

Le deuxième s'exprime formellement :

$$\exists \left((e_i^1, t_i^1) \in \omega^1, (e_j^2, t_j^2) \in \omega^2\right), \ (e_i^1, t_i^1) \rightarrow \left(\neg pre(e_j^2), t_j^2\right).$$

La première grande colonne du tableau 5.1 et 5.2 peut alors être divisée en deux sous-colonnes en fonction des ces deux cas de figure.
Dans la typologie des cas de surdétermination négative présentée par le tableau 7.2 et 7.3, seul un cas des deux est possible. Ce qui a interrompu $\overline{\omega}^2$ est nécessairement une des occurrences composant le chemin causal $\overline{\omega}^1$. Ce cas s'exprime formellement :

$$\exists \left((e_i^1, t_i^1) \in \overline{\omega}^1, (e_j^2, t_j^2) \in \overline{\omega}^2\right), \ (e_i^1, t_i^1) \rightarrow \left(\neg pre(e_j^2), t_j^2\right).$$

L'autre cas n'est pas possible puisqu'il reviendrait à dire qu'une non occurrence a eu un effet, ce qui nous l'avons vu dans la section 7.1 n'est pas possible dans un STEE. De fait, il reviendrait à dire que ce qui aurait interrompu $\overline{\omega}^2$ serait la non occurrence $(\overline{e_\psi}, t_\psi)$, due au fait que $\overline{\omega}^1$ a abouti en premier. Formellement, cela correspondrait à :

$$\exists (e_j^2, t_j^2) \in \overline{\omega}^2, (\overline{e_\psi}, t_\psi) \rightarrow \left(\neg pre(e_j^2), t_j^2\right).$$

Ayant vu dans la section 7.1 que la non occurrence d'un évènement n'a pas de statut causal, la relation $(\overline{e_\psi}, t_\psi) \rightarrow \left(\neg pre(e_j^2), t_j^2\right)$ ne peut être possible. La première grande colonne

du tableau 7.2 et 7.3 ne peut pas être divisée en deux sous-colonnes comme pour la surdétermination positive. Les cas de surdétermination de type préemptive tardive n'existent pas pour la surdétermination négative. À part ce type correspondant à la définition 5.9, les définitions 5.8 à 5.13 correspondant aux différents types de surdétermination positive peuvent toutes être appliquées telles quelles à la surdétermination négative.

	$\overline{\Omega}^1 = \{\overline{\omega}^1\}, \overline{\Omega}^2 = \emptyset$		$\overline{\Omega}^1 = \{\overline{\omega}^1\}, \overline{\Omega}^2 = \{\overline{\omega}^2\}$	
	$(e_i^1, t_i^1) \rightarrow (\neg pre(e_j^2), t_j^2)$	$(\overline{e_\psi}, t_\psi) \rightarrow (\neg pre(e_j^2), t_j^2)$	$W^1 \neq W^2$	$W^1 = W^2$
$t_1^1 < t_1^2$	Préemptive précoce	**	Duplicative asynchrone	Imitative
$t_1^1 = t_1^2$	Préemptive précoce	**	Duplicative synchrone	Symétrique
$t_1^1 > t_1^2$	Préemptive précoce	**	***	***

TABLEAU 7.3 – Typologie formelle concise des cas de surdétermination négative prenant en compte les relations temporelles pertinentes entre deux chemins causaux. (**) Incohérence dans la relation causale, une telle relation ne peut pas exister. (***) Incohérence entre l'hypothèse que $\overline{\omega}^1$ est toujours le premier chemin causal à aboutir et la relation temporelle entre les intervalles.

Le tableau 7.3 présente une version de la typologie obtenue en remplaçant les treize relations de Allen par le paramètre temporel identifié comme discriminant. Dans le chapitre 5, l'équivalent de cette version avait le défaut de perdre la distinction entre préemption tardive et préemption tardive durative. La préemption tardive n'existant pas en surdétermination négative, cette version plus concise n'a pas ce désavantage.

Comme pour la surdétermination positive, grâce à la typologie présentée ci-dessus nous proposons d'aller au delà d'exemples individuels et de généraliser la comparaison entre approches. Ayant des définitions claires des différents types de cas de surdétermination, il est possible d'établir des propriétés qui caractérisent la façon dont une définition de causalité va considérer un type de surdétermination. Pour reprendre une formulation précédente, si pour un premier exemple une théorie trouve que (e_m^1, t_m^1) est une cause et pas (e_n^2, t_n^2), alors, pour un autre exemple classé comme étant du même type que le premier par notre typologie, cette théorie devrait donner la même réponse. Plutôt que de confronter les différentes approches à de multiples exemples pas nécessairement représentatifs de l'ensemble de cas possibles, nous pouvons à présent prouver quelles seront les occurrences d'évènements considérées par l'approche comme des causes, et cela pour tous les exemples d'un même type. Faire ce travail pour les cinq catégories proposées permet de comparer plus facilement les approches entre elles. Voici un exemple de comment il serait possible de caractériser les différentes approches :

Définition 7.4 [Sensibilité des approches causales]**.** *Une approche en causalité effective est sensible à la surdétermination négative :*

- *préemptive : si dans ces cas elle considère* $(e_m^1, t_m^1) \in \overline{\omega^1}$ *une cause effective de* $(\overline{e_\psi}, t_\psi)$*, contrairement à* $(e_n^2, t_n^2) \in \overline{\omega^2}$*.*
- *duplicative et symétrique : si dans ces cas elle considère aussi bien* $(e_m^1, t_m^1) \in \overline{\omega^1}$ *que* $(e_n^2, t_n^2) \in \overline{\omega^2}$ *comme des causes effectives de* $(\overline{e_\psi}, t_\psi)$*.*

Pour finir la présentation de l'approche factuelle que nous proposons d'intégrer à notre approche causale, nous déterminons les propriétés de notre définition de causes effectives de non occurrence par rapport à la typologie de surdétermination négative que nous venons de proposer.

Théorème 7.1. *Notre approche de causalité effective à laquelle nous intégrons la définition 7.1 est sensible à la surdétermination négative préemptive, duplicative et symétrique au sens de la définition 7.4.*

Démonstration. Nous nous plaçons dans le cadre causal χ et dans un cas de surdétermination négative. Par la définition 7.2, nous avons alors $\overline{e_\psi} \in E^\chi(t_\psi)$, $(e_m^1, t_m^1), (e_n^2, t_n^2)$ deux occurrences d'évènements, et trois cadres causaux contrefactuels :

$$\chi_I^1 = (\pi^\sigma \setminus \{e_n^2\}, \kappa_c), \qquad \chi_I^2 = (\pi^\sigma \setminus \{e_m^1\}, \kappa_c), \qquad \chi_- = (\pi^\sigma \setminus \{e_m^1, e_n^2\}, \kappa_c).$$

De plus, nous considérons deux chemins causaux vers une non occurrence d'évènement : $\overline{\omega}^1$ dans χ_I^1 qui relie (e_m^1, t_m^1) à $(\overline{e_\psi}, t_\psi)$ et $\overline{\omega}^2$ dans χ_I^2 qui relie (e_n^2, t_n^2) à $(\overline{e_\psi}, t_\psi)$. Pour rappel, la typologie est construite avec les hypothèses que $|\overline{\Omega}_I^1| = |\overline{\Omega}_I^2| = 1$, i.e. qu'il y a un unique chemin causal par ensemble de chemins causaux individuels, $\overline{\Omega}^i \setminus \overline{\Omega}_I^i = \varnothing$, i.e. qu'aucun chemin causal n'est créé de l'interaction entre $\overline{\omega}^1$ et $\overline{\omega}^2$ dans χ. De ce fait, $\overline{\Omega}_I^1 = \overline{\Omega}^1 = \{\overline{\omega}^1\}$ et $\overline{\Omega}_I^2 = \{\overline{\omega}^2\}$. Deux cas sont alors possibles : soit $\overline{\Omega}^2 = \{\overline{\omega}^2\}$, ou $\overline{\Omega}^2 = \varnothing$.

[Cas surdétermination négative symétrique et duplicative] : nous sommes dans le cas où $\overline{\Omega}^2 = \{\overline{\omega}^2\}$. Nous voulons prouver que notre approche de causalité effective est sensible à la surdétermination négative symétrique et duplicative. D'après la définition 7.4, cela veut dire que dans ces deux cas de surdétermination, notre approche causale considère aussi bien (e_m^1, t_m^1) que (e_n^2, t_n^2) comme des causes effectives de $(\overline{e_\psi}, t_\psi)$.

D'après le lemme 6.2, étant donné que $\overline{\omega}^1$ et $\overline{\omega}^2$ sont tous deux des chemins causaux dans χ, les occurrences d'évènements (e_m^1, t_m^1) et (e_n^2, t_n^2) sont considérées des NESS-causes de $(\neg\psi, t_\psi)$. Par conséquent, d'après la définition 7.1, aussi bien (e_m^1, t_m^1) que (e_n^2, t_n^2) sont considérées comme des causes effectives de $(\overline{e_\psi}, t_\psi)$.

Compte tenu des relations causales $(e_1^1, t_1^1) \underset{W^1}{\rightarrow} (\neg\psi, t_\psi)$ et $(e_1^2, t_1^2) \underset{W^2}{\rightarrow} (\neg\psi, t_\psi)$, ce résultat est vrai que ce soit dans le cas où $W^1 = W^2$ ou dans le cas où $W^1 \neq W^2$, et cela indépendamment de la relation temporelle entre t_1^1 et t_1^2. De ce fait, notre approche causale est sensible aux cas de surdétermination négative duplicative asynchrone, duplicative synchrone et symétrique.

[Cas surdétermination négative préemptive] : nous sommes dans le cas où $\overline{\Omega}^2 = \varnothing$. Nous voulons prouver que notre approche de causalité effective est sensible à la surdétermination négative préemptive. D'après la définition 7.4, cela veut dire que dans ces cas de surdétermination, notre approche causale considère (e_m^1, t_m^1) comme une cause effective de $(\overline{e_\psi}, t_\psi)$, mais pas (e_n^2, t_n^2).

D'après le lemme 6.2, étant donné que $\overline{\omega}^1$ est un chemin causal dans χ, l'occurrence d'évènement (e_m^1, t_m^1) est considérée une NESS-cause de $(\neg\psi, t_\psi)$. Par conséquent, d'après la définition 7.1, (e_m^1, t_m^1) est considérée comme une cause effective de $(\overline{e_\psi}, t_\psi)$.

Considérons maintenant le cas de (e_n^2, t_n^2). Étant donné que $\overline{\omega}^2$ n'est pas un chemin causal dans χ, nous savons que : soit $\exists (e_i^2, t_i^2) \in \overline{\omega}^2$ tel que, (e_1^2, t_1^2) n'est pas une NESS-cause directe de $(\neg\psi, t_\psi)$, soit (e_i^2, t_i^2) n'est pas une NESS-cause directe de $(tri(e_{i-1}^2), t_{i-1}^2)$. En reprenant le raisonnement utilisé dans la preuve du lemme 6.2, nous pouvons déduire qu'étant donné nos hypothèses, $|\overline{\Omega}_I^2| = 1$ et $\overline{\Omega}^2 \setminus \overline{\Omega}_I^2 = \varnothing$, dans aucun des deux cas (e_n^2, t_n^2) ne peut être considérée une NESS-cause de $(\neg\psi, t_\psi)$.
Dans le premier cas, les NESS-causes étant construites en remontant le temps à partir d'un ensemble suffisant de NESS-causes directes, l'absence d'un tel ensemble rend impossible

le cas base et le cas récursif dans la définition 6.11.
Dans le deuxième cas, comme (e_i^2, t_i^2) n'est pas une NESS-cause directe de $(tri(e_{i-1}^2), t_{i-1}^2)$, il n'y a pas de moyen que (e_i^2, t_i^2) soit une NESS-cause de $(tri(e_{i-1}^2), t_{i-1}^2)$. En effet, le même raisonnement que pour le cas précédent peut être appliqué avec $(tri(e_{i-1}^2), t_{i-1}^2)$ étant le nouveau $(\neg\psi, t_\psi)$. Par conséquent, d'après la définition 7.1, comme (e_n^2, t_n^2) n'est pas une NESS-cause de $(\neg\psi, t_\psi)$, (e_n^2, t_n^2) n'est pas non plus une cause effective de $(\overline{e_\psi}, t_\psi)$.

Ces résultats étant applicables dans le cas où :

$$\exists\left((e_i^1, t_i^1) \in \overline{\omega}^1, (e_j^2, t_j^2) \in \overline{\omega}^2\right), (e_i^1, t_i^1) \rightarrow (\neg tri(e_j^2), t_j^2),$$

notre approche causale est sensible aux cas de surdétermination préemptive précoce. □

7.2.2 L'approche factuelle face au besoin de faire la différence

Dans cette section nous montrons que pour certains cas, l'approche factuelle donne des résultats qui ne correspondent pas à ce qui semble être compris avec la notion d'empêcher. Commençons pour cela par reprendre l'exemple 3.11 sur la balle et la fenêtre et représentons le dans $\mathscr{S}_c$.

Exemple 7.8 [la balle et la fenêtre]**.** *Nous notons $mv \in Lit_{\mathbb{F}}$ le littéral indiquant que la balle est en mouvement, $pos_0, pos_1, pos_2, pos_3, pos_4, pos_5 \in Lit_{\mathbb{F}}^6$ les littéraux indiquant la position dans laquelle se trouve la balle et $br \in Lit_{\mathbb{F}}$ le littéral indiquant que la fenêtre est brisée. L'ensemble de temps de simulation est $\mathbb{T} = \{0,1,2,3,4\}$.*

L'évènement naturel simulant le mouvement de la balle est noté $nxt_t \in \mathbb{N}$, où $t \in \mathbb{T}$. Il est déclenché lorsque la balle est en mouvement mv et est en position pos_t, et a pour effet d'avancer d'une position la balle qui ne sera plus en pos_t mais en pos_{t+1}.
L'évènement naturel simulant que la fenêtre est impactée est noté $imp \in \mathbb{N}$. Il est déclenché lorsque la balle est dans la position pos_4 et est en mouvement mv, et a pour effet que la fenêtre est brisée br et que la balle est arrêtée $\overline{mv}$.
Les actions correspondant aux interventions de A *et* B *sont respectivement $int_a, int_b \in \mathbb{A}^2$.* A *qui est en pos_1 peut attraper la balle s'il intervient lorsque celle-ci est à la position pos_0 et* B *qui est en pos_3 si celle-ci est à la position pos_2. Ces deux actions ont pour effet d'arrêter la balle $\overline{mv}$. Toute l'information donnée ci-dessus peut être représentée ainsi :*

$$pre(int_a) = pos_0,\ eff(int_a) = \overline{mv};$$
$$pre(int_b) = pos_2,\ eff(int_b) = \overline{mv};$$
$$tri(imp) = mv \wedge pos_4, eff(imp) = \overline{mv} \wedge br;$$
$$tri(nxt_t) = mv \wedge pos_t,\ eff(nxt_t) = \overline{pos_t} \wedge pos_{t+1}.$$

L'information donnée jusqu'ici sur l'exemple correspond au contexte κ_c. Nous allons maintenant préciser un scénario σ possible auquel nous allons nous intéresser par la suite. Il s'agit du cas de surdétermination négative préemptive précoce décrit dans l'exemple 3.11 où A *attrape la balle avant* B. *Le scénario correspondant s'écrit $\sigma = \{(int_a, 0), (int_b, 2)\}$. Voici les traces τ_χ^e et τ_χ^s qui correspondent au cadre causal χ :*

$$S^\chi(-1) = \left\{\overline{pos_0}, pos_1, pos_2, pos_3, pos_4, pos_5, \overline{mv}, br\right\},\ E^\chi(-1) = \left\{ini_l | l \in S^\chi(0)\right\};$$
$$S^\chi(0) = \left\{pos_0, \overline{pos_1}, \overline{pos_2}, \overline{pos_3}, \overline{pos_4}, \overline{pos_5}, mv, \overline{br}\right\}, E^\chi(0) = \{int_a, nxt_0\};$$

$$S^{\chi}(1) = \left\{\overline{pos_0}, pos_1, \overline{pos_2}, \overline{pos_3}, \overline{pos_4}, \overline{pos_5}, \overline{mv}, \overline{br}\right\}, E^{\chi}(1) = \emptyset\ ;$$

$$S^{\chi}(2) = S^{\chi}(3) = S^{\chi}(4) = S^{\chi}(5) = S^{\chi}(1), \qquad E^{\chi}(2) = E^{\chi}(3) = E^{\chi}(4) = E^{\chi}(1).$$

Pour cet exemple, notre approche étant sensible à la surdétermination négative préemptive d'après le théorème 7.1, celle-ci considère que $(int_a, 0)$ est une cause effective de $(\overline{nxt_1}, 1)$, mais pas $(int_b, 2)$. Cela s'explique par le fait que $(int_a, 0)$ est une NESS-cause de $(\overline{mv}, 1)$ et donc de $(\neg tri(nxt_1), 1)$ également. La balle n'étant plus en mouvement, elle ne peut pas atteindre la position où se trouve B et donc l'action int_b ne peut pas avoir lieu. Intuitivement, si nous adoptons un point de vue purement factuel comme pour la causalité positive, nous avons envie de dire que c'est uniquement A qui a empêché la fenêtre d'être brisée, c'est son processus causal qui a abouti, celui de B peut être vu comme préempté.

Toutefois, cette vision n'est pas acceptée par tous. Comme tout cas de surdétermination, aucune des deux actions n'était nécessaire à la conséquence, ici empêcher la fenêtre d'être brisée. En effet, si A n'avait pas attrapé le ballon, B l'aurait fait et vice versa. Le fait que B s'apprête à attraper la balle a quelque part rendu la contribution de A superflue, la balle n'aurait pas pu atteindre la fenêtre. D'après cette vision, il est possible de considérer que A n'a pas réellement empêché la fenêtre d'être brisée mais uniquement B. Une variante de ce problème présentée dans l'exemple 3.12 est souvent utilisée pour faire ressortir cette intuition.

Exemple 7.9 [la balle, la fenêtre et le mur]**.** *La représentation de cet exemple est quasiment identique. La seule modification consiste à remplacer l'action int_b qui devient l'évènement naturel simulant l'impact de la balle avec le mur, que nous notons $mur \in \mathbb{N}$. Il est déclenché lorsque la balle est dans la position pos_2 et a pour effet d'arrêter la balle $\overline{mv}$. Cette information est représentée ainsi :*

$$tri(mur) = pos_2,\ \mathit{eff}(mur) = \overline{mv}.$$

Le scénario correspondant s'écrit $\sigma = \{(int_a, 0)\}$. *Les traces dans les deux cas sont exactement les mêmes.*

Pour cet exemple, le résultat causal donné par notre approche est le même. Il n'y a pas de raison qu'il ne le soit pas, les traces sont exactement les mêmes. Intuitivement, si nous adoptons un point de vue purement factuel comme pour la causalité positive, nous avons envie de dire que c'est uniquement A qui a empêché la fenêtre d'être brisée, c'est son processus causal qui a abouti, celui de B peut être vu comme préempté. Par contre, la notion d'empêcher semble dans cet exemple faire intervenir plus fortement la notion de nécessité que dans sa version précédente, ou que dans la causalité positive. Dans l'exemple 7.9 l'intervention de A nous semble vraiment superflue, elle ne fait pas de différence car en fin de compte la fenêtre n'avait aucun risque d'être brisée de par la présence du mur. Nous retrouvons l'idée d'un besoin de pouvoir faire la différence comme dans le cas de l'omission traité dans la section 7.1. Ce besoin explique que dans certains cas, l'application du test NESS ne semble pas satisfaisante pour les résoudre.
Cette intuition ne s'applique pas uniquement à cet exemple, elle semble pouvoir expliquer la majorité des exemples débattus dans le domaine qui font intervenir la notion d'empêcher. Est-il juste de dire que nous empêchons un train d'arriver à une gare si nous le dévions, mais qu'en fin de compte il y avait une boucle qui ramène le train à la station ? Est-il

juste de dire que nous empêchons la mort par empoisonnement d'une personne si lorsque nous avons mis l'antidote dans sa boisson elle n'était pas empoisonnée et que c'est parce que nous avons mis l'antidote que l'autre agent a décidé de mettre le poison? Dans l'extension que nous avons proposée de l'exemple 3.8, il n'est finalement pas si évident de savoir quelle action relier avec le fait que le voyageur ne meurt pas lors de son voyage. Certes, avoir remplacé la fiole a fait que l'eau dans la gourde n'était pas empoisonnée, mais ayant des réserves d'eau rien ne dit que dans le cas où la gourde aurait été empoisonnée le voyageur serait mort. De même, avoir pris des réserves d'eau a fait que le voyageur ne puisse pas mourir déshydraté, mais ayant remplacé la fiole empoisonnée avec une fiole avec un produit inoffensif, rien ne dit que dans le cas où son garde du corps n'aurait pas pris les réserves d'eau le voyageur serait mort.
Plus généralement, en faisant la synthèse de ce qu'a observé MOORE [2019] lorsqu'il a étudié les jugements de nombreux cas en droit, deux conclusions peuvent être tirées : dans la plupart des cas, (i) du moment où il y a un unique chemin causal qui abouti seul l'évènement appartenant à ce chemin se voit attribuer une responsabilité; (ii) si deux chemins causaux aboutissent alors aucun des deux évènements ne se voit attribuer de responsabilité. L'approche factuelle présentée dans la section 7.2.1 semble satisfaire le premier cas, mais pas le deuxième. Toutefois, cela n'est pas un élément pouvant servir de preuve. Pour rappel, ces observations ne correspondent pas à des règles strictes, elles s'appliquent « dans la plupart des cas », des exceptions existent. De plus, rappelons que MOORE [2019] s'intéresse à la responsabilité, ce qui implique d'autres éléments en plus de la causalité. Par contre, ces résultats confortent l'idée que si l'approche factuelle ne satisfait pas tous les cas, c'est parce que la notion d'empêcher semble être corrélée au besoin de faire la différence.

7.2.3 Empêcher, une notion à deux niveaux dont un normatif

Dans cette section nous montrons que la notion d'empêcher peut être décomposée en deux niveaux de raisonnement, un factuel et un normatif. Dans les cas qui incombent au premier niveau, l'approche factuelle que nous proposons est satisfaisante. Les cas qui posent problème font intervenir le deuxième. La raison pour laquelle l'approche causale factuelle n'est pas satisfaisante est parce que ces problèmes sont en dehors du cadre causal; ce sont des problèmes de responsabilité qui demandent d'attribuer des conséquences à la non occurrence d'évènements.

Commençons par mettre en évidence l'existence des deux niveaux de raisonnement. Reprenons la discussion soulevée dans la section 3.2.2.1. Le raisonnement pour déterminer les causes pour lesquelles une condition s'est produite est différent de celui pour déterminer les causes pour lesquelles une condition ne s'est pas produite. Le second repose sur l'étude d'un processus causal qui n'a pas abouti. Comme nous l'avons vu dans la section 7.2.1, pour qu'un tel échec ait lieu, il faut qu'au moins une condition nécessaire dans chaque ensemble de conditions suffisantes ne soit pas vérifiée. La problématique est de savoir quels éléments sont à l'origine de l'échec de chacun de ces cas. Dans l'exemple 7.8, l'approche factuelle considère que $(int_a, 0)$ est une cause effective de $(\overline{nxt_1}, 1)$, mais pas $(int_b, 2)$ car $(int_a, 0)$ est une NESS-cause de $(\overline{mv}, 1)$ et donc de $(\neg tri(nxt_1), 1)$. La balle n'étant plus en mouvement, elle ne peut pas atteindre la position où se trouve B et donc l'action int_b ne peut pas avoir lieu et donc n'a aucun effet. Ce premier niveau de raisonnement plus élémentaire est ce que WRIGHT [2011] appelle « simple prevention », il intervient lorsque l'action que nous souhaitons évaluer empêche quelque chose de se produire, il est la cause d'une absence dans

le monde. Ce niveau de raisonnement ne pose pas de problèmes. Si la question posée est qu'est-ce qui a empêché la balle d'être en mouvement ou de continuer son chemin ? Il n'est pas abusif de considérer qu'il y a unanimité pour dire que c'est l'intervention de A.
Mais ce premier niveau ne semble pas être suffisant lorsque nous pensons à la relation empêcher ; quelque part ce qui nous intéresse réellement dans ce cas est de déterminer les raisons pour lesquelles la fenêtre n'a pas été brisée. WRIGHT [2011] parle alors de « double prevention », un raisonnement qui intervient lorsque l'action que nous souhaitons évaluer interrompt un processus causal qui, s'il n'avait pas été interrompu, aurait eu des conséquences sur le monde. Dans ce cas, nous avons envie de dire que l'action que nous évaluons est la cause que ces conséquences n'aient pas eu lieu. En l'occurrence, dans l'exemple 7.8 et 7.9, nous avons envie de relier l'action qui interrompt le mouvement de la balle $(int_a, 0)$, au fait que la fenêtre ne se brise pas $(\overline{imp}, 4)$. C'est à ce niveau de raisonnement que la présence de B ou du mur vient contredire l'aspect factuel.

Si ce deuxième niveau pose des problèmes à l'approche factuelle c'est parce qu'au delà du premier niveau nous ne sommes plus dans le terrain de la causalité effective, mais dans celui de la responsabilité. En effet, pour pouvoir relier le tout, il est nécessaire d'attribuer des conséquences à la non occurrence d'évènements. Une fois que nous avons déterminé que $(int_a, 0)$ est une cause effective de $(\overline{nxt_1}, 1)$, nous voulons relier le fait que la balle ne soit plus en mouvement $(\overline{nxt_1}, 1)$ au fait que la fenêtre ne se brise pas $(\overline{imp}, 4)$. Comme nous l'avons montré dans la section 7.1.2, $(\overline{nxt_1}, 1)$ étant une non occurrence d'évènement cela nécessite de passer par un raisonnement hypothétique. De ce fait, comme nous l'avons vu dans la section 7.1.3, cela passe nécessairement par l'introduction d'aspects normatifs pour choisir les mondes hypothétiques, nous faisant ainsi sortir du cadre purement causal. Si l'exemple 7.9 paraît demander un résultat différent à celui de l'exemple 7.8, c'est parce qu'il est plus compliqué d'imaginer le monde hypothétique où le mur n'est pas là, que le monde hypothétique où B n'essaye pas d'intercepter le balle ou n'y parvient juste pas.
Le fait que ces cas soient des cas de responsabilité explique que des aspects a priori non causaux modifient l'intuition du résultat attendu. Reprenons le cas du voyageur dans le désert cible de deux assassins, et ses variantes. Imaginons que l'agent commanditaire du meurtre du voyageur paye l'assassin A pour l'empoisonner, puis qu'il paye l'assassin B pour tuer l'assassin A. Même si factuellement il a contribué à empêcher que l'autre assassin puisse empoisonner la gourde du voyageur, il semble contre intuitif de dire qu'il est la cause que le voyageur soit encore en vie. Dans la variante où l'assassin A aurait été embauché par le même agent qui embauche ensuite B, mais aussi par un autre agent, il devient un peu plus acceptable de faire ce lien. Celui-ci l'est encore plus si l'agent qui embauche l'assassin A et l'assassin B sont distincts. Reprenons l'exemple proposé par HITCHCOCK [2007] que nous avions mentionné dans le chapitre 3. Il s'agit du cas où un garde du corps verse un antidote inoffensif pour la santé humaine dans le café de l'agent qu'il protège. Puis, Buddy verse un poison létal dans le café, mais les effets de celui-ci se voient annulés par la présence de l'antidote. On considère dans cet exemple, pour des raisons qui lui appartiennent, que Buddy n'aurait pas empoisonné le café si le garde du corps n'avait pas en premier versé l'antidote. L'agent boit son café et ne meurt pas. À nouveau, nous sommes dans une situation où factuellement avoir versé l'antidote a empêché le poison de faire effet et l'agent à l'origine de cette action est également à l'origine de la menace. Mais ici la situation est plus complexe car ce n'est pas juste le même agent qui est à l'origine du danger et de son absence, c'est la même action. En effet, l'exemple précise que c'est le fait de verser l'antidote qui fait que

Buddy verse le poison.

La notion d'empêcher faisant intervenir des aspects normatifs, il peut exister autant de points de vue que de personnes qui se penchent sur le problème. Mentionnons en quelques uns. Une première vision pourrait être de considérer que dans les cas de surdétermination négative duplicative, les occurrences en jeu peuvent plus être considérées comme empêchant que dans les cas de surdétermination négative symétrique ou imitative car en attaquant deux supports différents, il sera plus difficile par la suite de causer $e \in \mathrm{E}^{\mathrm{X}}(t)$. Dans l'exemple 6.2 du peloton d'exécution par chaise électrique, s'attaquer uniquement aux interrupteurs peut être considéré moins robuste qu'ouvrir les deux interrupteurs et éteindre le générateur. Dans les cas de surdétermination négative symétrique ou imitative l'évènement qui arrive après ne semble servir à rien. Cette vision semble être celle de HITCHCOCK [2007] qui considère que lorsque deux chemins causaux sont en compétition, celui ayant abouti en premier est le seul à avoir empêché : « In the case where both bodyguards administer antidotes, it seems that only the first to put her antidote into the coffee causes the coffee to be neutralized ». Pour justifier cette vision COLLINS [2000] introduit une forme de préemption différente que celle où un des chemins interrompt le deuxième. Pour lui, un chemin causal peut être préempté par le simple fait qu'un autre chemin causal existe et donc rende le premier non nécessaire. Pour insérer une asymétrie dans cette relation qui lui permette de choisir quel chemin causal est celui responsable d'avoir empêché, il fait allusion à une comparaison de mondes hypothétiques :

> [...] the counterfactual assumption of the absence of the pure dependence preventer in his story is more far-fetched than the corresponding assumption of absence in mine. It does not require much of a stretch to suppose that I simply get my timing slightly wrong, so that when I leap, I do so at not quite the right moment to be ready to take the catch. It is more far-fetched, on the other hand, to suppose that the brick wall be absent, or that the ball would miraculously pass straight though it.
> So let us adjust the would-be dependence analysis accordingly : A causal chain is a chain of true propositions of occurrence which is a chain of counterfactual dependence or would be were some true proposition of occurrence false in some not too far-fetched way.

Une autre vision possible peut être celle de WRIGHT [2011] qui défend que l'information du premier niveau de raisonnement, celui factuel, doit prévaloir sur le deuxième. Pour lui, faire autrement serait confondre des notions distinctes : « confuses strong necessity or lawful strong suffficiency (guaranteeing an outcome), respectively, with causal strong sufficiency (actually causing the outcome) ».

Finalement, une autre vision possible est celle de BRAHAM et VAN HEES [2012] qui considère que pouvoir faire la différence n'est pas si important. Pour lui ce qui compte c'est de ne pas prendre part à la chaîne causale. Le fait de ne pas être une cause pourrait être suffisant sans avoir besoin d'empêcher : « [...] on our account the important consideration is not whether a person could have realized a different outcome but whether he could have avoided making a causal contribution to it ». Nous pourrions retrouver ce raisonnement lorsqu'un agent refuse d'acheter un téléphone volé ou de manger des produits de provenance animale, alors même qu'il est convaincu de ne pas pouvoir changer les choses. Le seul fait de ne pas participer à la chaîne causale a une importance, même si c'est a posteriori de l'acte qu'ils jugent moralement injuste.

7.3 Causalité positive, adaptée pour représenter toutes les formes de négation dans la relation causale

Dans cette section nous montrons que l'approche de causalité proposée est une base appropriée pour représenter toutes les formes de négation dans la relation causale, malgré leur aspect normatif. Pour rappel, notre premier défi consiste à proposer une approche de causalité positive commune adaptée à l'éthique computationnelle, ce qui nous l'avons montré passe par une définition adaptée qui sépare clairement causalité et responsabilité. En effet, chaque théorie morale pouvant avoir une vision propre des aspects à considérer pour déterminer quelles conséquences d'une action doivent être prises en compte dans son évaluation, l'approche que nous nous sommes fixés de proposer doit être le plus factuelle possible pour ne véhiculer aucune vision de responsabilité. Nous avons montré dans la section précédente qu'attribuer des conséquences à une non occurrence d'évènement est nécessairement une question de responsabilité et qu'aussi bien la notion d'omission que la notion d'empêcher demandent qu'une telle attribution soit faite. De plus, nous avons montré qu'il n'est pas possible de trouver une vision qui convienne à toutes les théories morales même si nous acceptions de sortir du cadre purement causal en faisant un pas vers la responsabilité. Pour cette raison, proposer une définition de comment prendre en compte la notion d'omission et d'empêcher est au-delà du cadre de cette contribution. Toutefois, nous montrons dans cette section comment l'approche de causalité positive proposée dans le chapitre 6 peut être utilisée pour modéliser différentes visions de quand attribuer des conséquences à une non occurrence. Par conséquent, nous montrons que notre approche peut être une base pour modéliser différentes visions de la notion d'omission et d'empêcher.

Cette section est divisée en trois parties. La section 7.3.1 montre comment intégrer la notion de « décisions » introduite dans le chapitre 4 à notre raisonnement causal. La section 7.3.2 propose plusieurs définitions d'omission. Chacune représente un point de vue différent sur comment évaluer si un agent aurait pu faire la différence, mais a omis de le faire. Cela est fait pour les deux types de relations causales dans un STEE que nous avons identifiés dans la section 5.1.2, les $\mathscr{F} - causes$ de la définition 5.4 et les causes effectives de la définition 5.5. Finalement, la section 7.3.3 discute des similarités et des différences entre ces propositions de responsabilité et les définitions de causalité de type Halpern.

7.3.1 Intégration de la notion de décision à $\mathscr{S}_c$

Dans cette section nous montrons comment intégrer la notion de « décisions » introduite dans le chapitre 4 à notre raisonnement causal. Nous y avons montré que la notion d'omission ne peut être clairement définie qu'en utilisant la notion de décision.

Pour rappel, dans la définition 4.1 les décisions ont été définies de la façon suivante : étant donné un ensemble d'actions $\mathbb{A}$ et un ensemble de contraintes C concernant la faisabilité des décisions, l'ensemble de décisions $\mathbb{D}$ est défini comme :

$$\mathbb{D} = \left\{ i : \mathbb{A} \to \{1,0\} \mid [\mathrm{C}]^i = 1 \right\}.$$

Il se trouve que dans $\mathscr{S}_c$, un des exemples de contraintes possibles serait d'interdire la présence d'évènements interférents. Pour rappel, nous considérons que deux évènements $e, e' \in \mathbb{E}^2$ sont interférents si l'ensemble $\left\{ l \in Lit_{\mathbb{F}} \mid l \in eff(e) \cup eff(e') \right\}$ n'est pas cohérent au

sens de la définition 6.1.
Pour être en mesure de raisonner sur des ensembles, nous avons introduit une notation ensembliste pour une décision, $d = \{a|i(a) = 1\} \cup \{\overline{a}|i(a) = 0\}$. Nous avons fait référence à la décision consistant à ne réaliser aucune action comme l'omission d'agir. Cette notion ne doit pas être confondue avec l'omission d'une décision, qui elle consiste à réaliser une alternative parmi celles existantes dans l'ensemble de décisions. Étant donné une décision $d \in \mathbb{D}$, omettre de réaliser d correspond à réaliser une décision $d' \in \overline{d}$, où $\overline{d} = \mathbb{D} \setminus \{d\}$ est l'ensemble des alternatives.

Reprenons l'exemple 7.8 sur la balle qui se dirige vers la fenêtre et les deux agents qui peuvent l'intercepter afin d'illustrer la notion de décision.

Exemple 7.10 [la balle et la fenêtre en omission]**.** *Nous reprenons le même contexte* κ_c *de l'exemple 7.8. L'ensemble d'actions dans notre exemple est* $\mathbb{A} = \{int_a, int_b\}$*. Formellement, l'ensemble de décisions réalisables* $\mathbb{D}$ *dans cet exemple où il n'y a pas de contraintes est :*

$$\left\{\{int_a, int_b\}, \left\{int_a, \overline{int_b}\right\}, \left\{\overline{int_a}, int_b\right\}, \left\{\overline{int_a}, \overline{int_b}\right\}\right\}.$$

Dans notre exemple l'omission d'agir correspond à la décision $d = \left\{\overline{int_a}, \overline{int_b}\right\}$*. Il s'agit du cas où ni* A*, ni* B *n'attrapent la balle.*

Dans le chapitre 4 nous avions soulevé que parler d'omission d'une action est ambigu. Quelle décision devrait être envisagée si nous voulions évaluer si l'omission de l'agent A aurait pu faire la différence en interceptant la balle, i.e. en omettant d'omettre d'intercepter la balle ? La décision obtenue en retirant uniquement $\overline{int_a}$ de d qui serait $d' = \left\{int_a, \overline{int_b}\right\}$? La décision obtenue si aucun agent n'avait omis d'agir $d'' = \{int_a, int_b\}$? N'importe quelle décision parmi celles contenant int_a, i.e. d' ou d'' ? Ou toutes les décisions parmi celles contenant int_a, i.e. d' et d'' ?
Le choix de l'ensemble de décisions à considérer correspond à ce que nous avons appelé dans la section 7.1 le choix du ou des mondes où il est possible de se placer pour évaluer si une omission aurait pu faire la différence. Nous avions spécifiquement évoqué deux cas pour traiter l'exemple 7.1 portant sur l'omission de freiner du conducteur au volant d'une voiture de location dont les freins étaient défectueux. Faut-il comparer le monde tel qu'il a été avec le monde où la seule chose qui change est l'omission qui est évaluée ? Cela correspondrait au cas où seul l'agent évalué accomplit son devoir. Dans l'exemple 7.1, si nous évaluons l'omission du conducteur, cela reviendrait à savoir si son omission fait la différence en prenant uniquement des mondes où la maintenance n'a pas été faite. Dans l'exemple 7.10, cela correspond au cas où nous nous intéressons à $d' = \left\{int_a, \overline{int_b}\right\}$. Sinon, faut-il comparer le monde tel qu'il a été avec le monde où tous les agents accomplissent leur devoir ? Dans l'exemple 7.1, si nous évaluons l'omission du conducteur, cela reviendrait à savoir si son omission fait la différence en prenant des mondes où la maintenance a été faite. Dans l'exemple 7.10, cela correspond au cas où nous nous intéressons à $d'' = \{int_a, int_b\}$. Comme nous l'avons montré, cette décision est purement normative. Pour l'omission il n'est pas possible de s'appuyer uniquement sur des aspects factuels comme dans la causalité positive. L'omission est une question de responsabilité.

Pour pouvoir intégrer les décisions dans $\mathscr{S}_c$, il est nécessaire que nous puissions relier la notion de décision avec celle de scénario σ définie dans la définition 6.6 et que nous utilisons pour définir le cadre causal χ. Pour rappel, un scénario noté σ est un ensemble

d'actions couplées à un temps $\sigma \subseteq \mathbb{A} \times \mathbb{T}$. Il représente la volition des agents et associe les actions à un temps. En combinant les deux notions nous pouvons définir l'ensemble des scénarios possibles pour un contexte κ_c donné.

Définition 7.5 [*Ensemble de scénarios* $\sigma^{\mathbb{D}}$]**.** *Étant donné un contexte* κ_c*, dans lequel nous avons un ensemble d'actions* $\mathbb{A}$ *et de temps* $\mathbb{T}$*, et un ensemble de contraintes* C *concernant la faisabilité des décisions, l'ensemble des scénarios envisageables dans* κ_c *que nous notons* $\sigma^{\mathbb{D}}$ *est défini comme :*

$$\sigma^{\mathbb{D}} = \left\{ i : \mathbb{A} \times \mathbb{T} \to \{1,0\} \mid [\mathrm{C}]^i = 1 \right\}.$$

Un scénario comme défini dans la définition 6.6 est une interprétation dans $\sigma^{\mathbb{D}}$, à chaque décision dans $\mathbb{D}$ correspond un unique scénario dans $\sigma^{\mathbb{D}}$. Si la décision évaluée est d, nous noterons le scénario lui correspondant σ^d. Pour être en mesure de raisonner sur des ensembles, nous étendons aux scénarios la notation ensembliste proposée pour une décision de sorte à ce qu'elle coïncide avec la notation définie pour une non occurrence d'évènement : $\sigma^d = \{(a,t) \mid i(a,t) = 1\} \cup \left\{(\overline{a},t) \mid i(a,t) = 0\right\}$.

Exemple 7.10 [suite]**.** *Pour notre exemple simple, l'ensemble de scénarios dans notre exemple est déjà très complexe. Pour rester concis, nous nous restreindrons au cas où l'action* int_a *ne peut être envisagée qu'au temps* $t = 0$ *et où l'action* int_b *qu'au temps* $t = 2$*. L'ensemble de scénarios que nous obtenons est alors :*

$$\sigma^{\mathbb{D}} = \left\{ \{(int_a,0),(int_b,2)\}, \left\{(int_a,0),(\overline{int_b},2)\right\}, \left\{(\overline{int_a},0),(int_b,2)\right\}, \left\{(\overline{int_a},0),(\overline{int_b},2)\right\} \right\}.$$

Pour rappel, dans notre exemple l'omission d'agir correspond à la décision $d = \left\{\overline{int_a}, \overline{int_b}\right\}$*. Il s'agit du cas où ni* A*, ni* B *n'attrapent la balle. Le scénario correspondant est dans ce cas* $\sigma^d = \left\{(\overline{int_a},0),(\overline{int_b},2)\right\}$*. Voici les traces* τ^e_χ *et* τ^s_χ *qui correspondent au nouveau cadre causal* χ *:*

$$\mathrm{S}^\chi(-1) = \left\{\overline{pos_0}, pos_1, pos_2, pos_3, pos_4, pos_5, \overline{mv}, br\right\},\ \mathrm{E}^\chi(-1) = \left\{ini_l \mid l \in \mathrm{S}^\chi(0)\right\};$$

$$\mathrm{S}^\chi(0) = \left\{pos_0, \overline{pos_1}, \overline{pos_2}, \overline{pos_3}, \overline{pos_4}, \overline{pos_5}, mv, \overline{br}\right\}, \mathrm{E}^\chi(0) = \{nxt_0\};$$

$$\mathrm{S}^\chi(1) = \left\{\overline{pos_0}, pos_1, \overline{pos_2}, \overline{pos_3}, \overline{pos_4}, \overline{pos_5}, mv, \overline{br}\right\}, \mathrm{E}^\chi(1) = \{nxt_1\};$$

$$\mathrm{S}^\chi(2) = \left\{\overline{pos_0}, \overline{pos_1}, pos_2, \overline{pos_3}, \overline{pos_4}, \overline{pos_5}, mv, \overline{br}\right\}, \mathrm{E}^\chi(2) = \{nxt_2\};$$

$$\mathrm{S}^\chi(3) = \left\{\overline{pos_0}, \overline{pos_1}, \overline{pos_2}, pos_3, \overline{pos_4}, \overline{pos_5}, mv, \overline{br}\right\}, \mathrm{E}^\chi(3) = \{nxt_3\};$$

$$\mathrm{S}^\chi(4) = \left\{\overline{pos_0}, \overline{pos_1}, \overline{pos_2}, \overline{pos_3}, pos_4, \overline{pos_5}, mv, \overline{br}\right\}, \mathrm{E}^\chi(4) = \left\{nxt_4, imp\right\};$$

$$\mathrm{S}^\chi(5) = \left\{\overline{pos_0}, \overline{pos_1}, \overline{pos_2}, \overline{pos_3}, \overline{pos_4}, pos_5, \overline{mv}, br\right\}, \mathrm{E}^\chi(5) = \emptyset.$$

7.3.2 Modélisation de différents points de vue sur la responsabilité

Maintenant que nous avons notre ensemble $\sigma^{\mathbb{D}}$, nous pouvons dans cette section donner quelques définition correspondant à différents points de vue sur l'omission. Plus précisément, nous donnerons une définition pour chacun des quatre points de vue possibles dans notre exemple sur comment évaluer si un des agents aurait pu faire la différence mais

a omis de la faire, aussi bien sur l'occurrence ou la non occurrence d'un évènement, que sur la véracité ou non d'une formule $\psi \in \mathscr{F}$. Nous couvrons ainsi les deux types de relations causales dans un STEE que nous avons identifiées dans la section 5.1.2, les $\mathscr{F} - causes$ de la définition 5.4 et les causes effectives de la définition 5.5.
Contrairement à la causalité positive, dans le cas de l'omission, il n'y a pas de lien entre les relations de type $\mathscr{F} - causes$ et les relations de type causes effectives. La connaissance de relations d'un type ne permet pas de déduire des relations l'autre. Cela s'explique par le fait que cette relation n'est pas nécessairement transitive comme nous le verrons par la suite. Si nous considérons que $(\overline{int_a}, 0)$ est responsables de $(imp, 4)$, $(\overline{int_a}, 0)$ n'est pas nécessairement responsable de $(br, 5)$. En effet, il se peut que dans les mondes hypothétiques explorables, l'omission de A puisse faire la différence par rapport à l'impact avec la fenêtre, mais pas avec le fait que la fenêtre soit brisée.

Dans quelques unes de ces définitions nous ferons référence au cardinal de la différence symétrique entre deux ensembles, qui pour rappel donne le nombre d'éléments différents entre ces deux ensembles et qui est défini pour deux ensembles A et B comme :

$$|\mathrm{A} \triangle \mathrm{B}| = |(\mathrm{A} \cup \mathrm{B}) \setminus (\mathrm{A} \cap \mathrm{B})|.$$

Les relations que nous allons présenter ci-dessous n'étant pas purement causales, nous introduisons une notation différente à celles utilisées jusqu'ici pour les relations causales. Nous noterons $(\overline{e}, t) \hookrightarrow (\overline{e_\psi}, t_\psi)$ le fait qu'une non occurrence d'évènement $(\overline{e}, t)$ soit responsable de la non occurrence d'évènement $(\overline{e_\psi}, t_\psi)$. Cette relation n'est pas uniquement entre deux non occurrences d'évènements, comme nous le verrons par la suite, d'autres combinaisons sont possibles.

Considérons en premier le cas où seul l'agent évalué accomplit son devoir. Si nous évaluons l'agir de A avec la définition 7.6, le monde hypothétique envisagé est celui correspondant à $d' = \left\{int_a, \overline{int_b}\right\}$ où nous prenons le scénario obtenu en retirant uniquement $\overline{int_a}$ de d. Dans ce cas, appliquer ce raisonnement pour A ou pour B donne qu'aussi bien $(\overline{int_a}, 0)$ que $(\overline{int_b}, 2)$ sont responsables de $(imp, 4)$.

Définition 7.6 [*Responsabilité individuelle par omission*]**.** *Soit le cadre causal* $\chi = (\kappa_c, \sigma^d)$*, deux non occurrences d'évènements* $(\overline{e}, t)$, $(\overline{e_\psi}, t_\psi)$ *telles que* $t < t_\psi$*, une occurrence d'évènement* (e_ϕ, t_ϕ) *telle que* $t < t_\phi$*, et* (ψ, t_ψ) *tel que* $\psi \in \mathscr{F}$ *et* $t < t_\psi$*. Étant donné un ensemble de décisions* $\mathrm{D} = \left\{d' \in \mathbb{D} | i(e,t) = 1\right\}$ *correspondant à toutes les décisions contenant* (e,t) *:*

- *Nous considérons* $(\overline{e}, t)$ *responsable de* (e_ϕ, t_ϕ)*, que nous notons* $(\overline{e}, t) \hookrightarrow (e_\phi, t_\phi)$*, ssi* $\mathrm{S}^\chi(t) \vDash pre(e)$ *et* (e,t) *est une cause effective de* $(\overline{e_\phi}, t_\phi)$ *dans le cadre causal* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in \mathrm{D}$ *est la décision telle que* $\forall d'' \in \mathrm{D}, \left|d \triangle d'\right| \leq \left|d \triangle d''\right|$*;*
- *Nous considérons* $(\overline{e}, t)$ *responsable de* $(\overline{e_\psi}, t_\psi)$*, que nous notons* $(\overline{e}, t) \hookrightarrow (\overline{e_\psi}, t_\psi)$*, ssi* $\mathrm{S}^\chi(t) \vDash pre(e)$ *et* (e,t) *est une cause effective de* (e_ψ, t_ψ) *dans le cadre causal* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in \mathrm{D}$ *est la décision telle que* $\forall d'' \in \mathrm{D}, \left|d \triangle d'\right| \leq \left|d \triangle d''\right|$*;*
- *Nous considérons* $(\overline{e}, t)$ *responsable de* (ψ, t_ψ)*, que nous notons* $(\overline{e}, t) \rightharpoonup (\psi, t_\psi)$*, ssi* $\mathrm{S}^\chi(t) \vDash pre(e)$ *et* (e,t) *est une NESS-cause de* $(\neg\psi, t_\psi)$ *dans le cadre causal* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in \mathrm{D}$ *est la décision telle que* $\forall d'' \in \mathrm{D}, \left|d \triangle d'\right| \leq \left|d \triangle d''\right|$*.*

Considérons maintenant le cas où tous les agents accomplissent leur devoir. Si nous évaluons l'agir de A avec la définition 7.7, le monde hypothétique envisagé est celui correspondant à $d'' = \{int_a, int_b\}$ où nous prenons le scénario obtenu en retirant aussi bien $\overline{int_a}$

que $\overline{int_b}$ de d. Dans ce cas, appliquer ce raisonnement pour A ou pour B donne que $(\overline{int_a}, 0)$ est responsable de $(imp, 4)$, alors que $(\overline{int_b}, 2)$ ne l'est pas.

Définition 7.7 [*Responsabilité collective par omission*]**.** *Soit le cadre causal* $\chi = (\kappa_c, \sigma^d)$*, deux non occurrences d'évènements* $(\overline{e}, t)$*,* $(\overline{e_\psi}, t_\psi)$ *telles que* $t < t_\psi$*, une occurrence d'évènement* (e_ϕ, t_ϕ) *telle que* $t < t_\phi$*, et* (ψ, t_ψ) *tel que* $\psi \in \mathscr{F}$ *et* $t < t_\psi$*. Étant donné un ensemble de décisions* $\mathrm{D} = \left\{d' \in \mathbb{D} | i(e,t) = 1\right\}$ *correspondant à toutes les décisions contenant* (e,t) *:*

- *Nous considérons* $(\overline{e}, t)$ *responsable de* (e_ϕ, t_ϕ)*, que nous notons* $(\overline{e}, t) \rightsquigarrow (e_\phi, t_\phi)$*, ssi* $S^\chi(t) \vDash pre(e)$ *et* (e,t) *est une cause effective de* $(\overline{e_\phi}, t_\phi)$ *dans le cadre causal* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in \mathrm{D}$ *est la décision telle que* $\forall d'' \in \mathrm{D}, \left|d \triangle d'\right| \geq \left|d \triangle d''\right|$*;*
- *Nous considérons* $(\overline{e}, t)$ *responsable de* $(\overline{e_\psi}, t_\psi)$*, que nous notons* $(\overline{e}, t) \rightsquigarrow (\overline{e_\psi}, t_\psi)$*, ssi* $S^\chi(t) \vDash pre(e)$ *et* (e,t) *est une cause effective de* (e_ψ, t_ψ) *dans le cadre causal* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in \mathrm{D}$ *est la décision telle que* $\forall d'' \in \mathrm{D}, \left|d \triangle d'\right| \geq \left|d \triangle d''\right|$*;*
- *Nous considérons* $(\overline{e}, t)$ *responsable de* (ψ, t_ψ)*, que nous notons* $(\overline{e}, t) \rightharpoonup (\psi, t_\psi)$*, ssi* $S^\chi(t) \vDash pre(e)$ *et* (e,t) *est une NESS-cause de* $(\neg\psi, t_\psi)$ *dans le cadre causal* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in \mathrm{D}$ *est la décision telle que* $\forall d'' \in \mathrm{D}, \left|d \triangle d'\right| \geq \left|d \triangle d''\right|$*.*

Considérons maintenant le cas que nous pourrions qualifier de crédule. Si nous évaluons l'agir de A avec la définition 7.8, n'importe quel monde hypothétique parmi ceux contenant int_a peut être envisagé. Nous pouvons donc envisager d' ou d''. Dans ce cas, appliquer ce raisonnement pour A ou pour B donne qu'aussi bien $(\overline{int_a}, 0)$ que $(\overline{int_b}, 2)$ sont responsables de $(imp, 4)$.

Définition 7.8 [*Responsabilité crédule par omission*]**.** *Soit le cadre causal* $\chi = (\kappa_c, \sigma^d)$*, deux non occurrences d'évènements* $(\overline{e}, t)$*,* $(\overline{e_\psi}, t_\psi)$ *telles que* $t < t_\psi$*, une occurrence d'évènement* (e_ϕ, t_ϕ) *telle que* $t < t_\phi$*, et* (ψ, t_ψ) *tel que* $\psi \in \mathscr{F}$ *et* $t < t_\psi$*. Étant donné un ensemble de décisions* $\mathrm{D} = \left\{d' \in \mathbb{D} | i(e,t) = 1\right\}$ *correspondant à toutes les décisions contenant* (e,t) *:*

- *Nous considérons* $(\overline{e}, t)$ *responsable de* (e_ϕ, t_ϕ)*, que nous notons* $(\overline{e}, t) \rightsquigarrow (e_\phi, t_\phi)$*, ssi* $S^\chi(t) \vDash pre(e)$ *et* (e,t) *est une cause effective de* $(\overline{e_\phi}, t_\phi)$ *dans n'importe quel cadre causal* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in \mathrm{D}$*;*
- *Nous considérons* $(\overline{e}, t)$ *responsable de* $(\overline{e_\psi}, t_\psi)$*, que nous notons* $(\overline{e}, t) \rightsquigarrow (\overline{e_\psi}, t_\psi)$*, ssi* $S^\chi(t) \vDash pre(e)$ *et* (e,t) *est une cause effective de* (e_ψ, t_ψ) *dans n'importe quel cadre causal* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in \mathrm{D}$*;*
- *Nous considérons* $(\overline{e}, t)$ *responsable de* (ψ, t_ψ)*, que nous notons* $(\overline{e}, t) \rightharpoonup (\psi, t_\psi)$*, ssi* $S^\chi(t) \vDash pre(e)$ *et* (e,t) *est une NESS-cause de* $(\neg\psi, t_\psi)$ *dans n'importe quel cadre causal* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in \mathrm{D}$*.*

Considérons maintenant le cas que nous pourrions qualifier de sceptique. Si nous évaluons l'agir de A avec la définition 7.9, tous les mondes hypothétiques parmi ceux contenant int_a doivent être envisagés. Nous devons donc envisager d' et d''. Dans ce cas, appliquer ce raisonnement pour A ou pour B donne que $(\overline{int_a}, 0)$ est responsable de $(imp, 4)$, alors que $(\overline{int_b}, 2)$ ne l'est pas.

Définition 7.9 [*Responsabilité sceptique par omission*]**.** *Soit le cadre causal* $\chi = (\kappa_c, \sigma^d)$*, deux non occurrences d'évènements* $(\overline{e}, t)$*,* $(\overline{e_\psi}, t_\psi)$ *telles que* $t < t_\psi$*, une occurrence d'évènement* (e_ϕ, t_ϕ) *telle que* $t < t_\phi$*, et* (ψ, t_ψ) *tel que* $\psi \in \mathscr{F}$ *et* $t < t_\psi$*. Étant donné un ensemble de décisions* $\mathrm{D} = \left\{d' \in \mathbb{D} | i(e,t) = 1\right\}$ *correspondant à toutes les décisions contenant* (e,t) *:*

— *Nous considérons* $(\overline{e}, t)$ *responsable de* (e_ϕ, t_ϕ)*, que nous notons* $(\overline{e}, t) \rightsquigarrow (e_\phi, t_\phi)$*, ssi* $S^\chi(t) \vDash pre(e)$ *et* (e, t) *est une cause effective de* $(\overline{e_\phi}, t_\phi)$ *dans tous les cadres causaux* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in D$*;*

— *Nous considérons* $(\overline{e}, t)$ *responsable de* $(\overline{e_\psi}, t_\psi)$*, que nous notons* $(\overline{e}, t) \rightsquigarrow (\overline{e_\psi}, t_\psi)$*, ssi* $S^\chi(t) \vDash pre(e)$ *et* (e, t) *est une cause effective de* (e_ψ, t_ψ) *dans tous les cadres causaux* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in D$*;*

— *Nous considérons* $(\overline{e}, t)$ *responsable de* (ψ, t_ψ)*, que nous notons* $(\overline{e}, t) \rightharpoonup (\psi, t_\psi)$*, ssi* $S^\chi(t) \vDash pre(e)$ *et* (e, t) *est une NESS-cause de* $(\neg\psi, t_\psi)$ *dans tous les cadres causaux* $\chi' = (\kappa_c, \sigma^{d'})$ *où* $d' \in D$.

Les propriétés suivantes montrent quelques liens qui existent entre les définitions ci-dessus. La démonstration de cette proposition est triviale.

Proposition 7.1 [Liens visions de responsabilité]**.** *Soit un cadre causal* χ*, deux non occurrences d'évènements* $(\overline{e}, t)$*,* $(\overline{e_\psi}, t_\psi)$*, une occurrence d'évènement* (e_ϕ, t_ϕ)*, et* (ψ, t_ψ) *tel que* $\psi \in \mathscr{F}$.

— *Si* $(\overline{e}, t)$ *est responsable de* (e_ϕ, t_ϕ)*, de* $(\overline{e_\psi}, t_\psi)$ *ou de* (ψ, t_ψ) *d'après la définition 7.6 ou 7.7, alors elle l'est également d'après la définition 7.8;*

— *Si* $(\overline{e}, t)$ *est responsable de* (e_ϕ, t_ϕ)*, de* $(\overline{e_\psi}, t_\psi)$ *ou de* (ψ, t_ψ) *d'après la définition 7.9, alors elle l'est également d'après la définition 7.6 et 7.7;*

— *Si* $(\overline{e}, t)$ *n'est ni responsable de* (e_ϕ, t_ϕ)*, ni de* $(\overline{e_\psi}, t_\psi)$*, ni de* (ψ, t_ψ) *d'après la définition 7.6 ou 7.7, alors elle ne l'est pas non plus d'après la définition 7.9;*

— *Si* $(\overline{e}, t)$ *n'est ni responsable de* (e_ϕ, t_ϕ)*, ni de* $(\overline{e_\psi}, t_\psi)$*, ni de* (ψ, t_ψ) *d'après la définition 7.8, alors elle ne l'est pas non plus d'après la définition 7.6 et 7.7.*

7.3.3 Discussion sur l'aspect contrefactuel de ces points de vue

Dans cette section nous discutons des similarités et des différences entre les propositions de responsabilité que nous avons faites et les définitions de causalité de type Halpern. Il est important de remarquer que le raisonnement contrefactuel qui est fait dans les définitions 7.6 à 7.9 ne fait pas de ces définitions des approches contrefactuelles au même titre que celles de type Halpern, présentées dans la section 3.1.1.2. Elles ont en commun ce que nous avons appelé le troisième élément dans les définitions de type Halpern : la notion de « contrefactuelle » qui va représenter la condition nécessaire pour qu'il y ait causalité. Pour rappel, cette conditions implique que, si c est une cause effective de e dans un scénario, alors c doit pouvoir prendre une autre valeur dans un scénario contrefactuel où, avec cette autre valeur, c n'est pas suffisant causalement pour e. Par contre, le processus pour y parvenir n'est pas le même.
D'un côté, les approches contrefactuelles classiques utilisent l'interventionnisme pour relier dépendance causale et causalité. Pour rappel, dans les définitions type Halpern, c est une cause effective de e ssi e est dépendant causalement de c étant donné une intervention. Les définitions de type Halpern reposent sur deux raisonnements hypothétiques cumulés. Le premier est celui consistant à chercher s'il existe une dépendance causale entre la cause et la conséquence. Cela passe par un but-for test où, pour rappel, il faut imaginer ce qui serait arrivé si la cause n'avait pas eu lieu. Ce premier raisonnement hypothétique semble être

proche du raisonnement de la définition 7.6 que nous avons identifié comme de responsabilité individuelle. Dans ce raisonnement, le monde exploré est celui où seul l'action évaluée est modifiée. Mais ce monde hypothétique n'est pas nécessairement obtenu depuis le cas effectif directement, il peut être fait depuis un monde hypothétique obtenu après avoir réalisé une intervention. Pour rappel, une intervention est une opération permettant de fixer arbitrairement la valeur d'un ensemble d'évènements autres que c et e dans le scénario en respectant un ensemble de conditions. Cette opération est le deuxième raisonnement hypothétique du processus. Il semble être proche du raisonnement de la définition 7.8, qualifié de responsabilité crédule, car après avoir imposé certaines conditions, il suffit qu'il existe une intervention qui mène vers un monde où il y a dépendance causale. Il est important de préciser que dans les définitions de type Halpern ces conditions n'exigent pas de devoir respecter la cohérence des équations. Par exemple, il est possible de fixer la valeur d'un nœud enfant x dans des scénarios contrefactuels où la valeur de ses nœuds parents sont censés donner une autre valeur à x que celle fixée.
D'un autre côté, les définitions 7.6 à 7.9 ne considèrent pas que la causalité est contrefactuelle et n'utilisent pas la notion d'interventionnisme. Le seul raisonnement hypothétique est celui qui mène aux mondes à considérer ; une fois placés dans ce cadre causal la relation de causalité reste une relation de causalité positive factuelle. ABRAMS [2022] insiste sur la différence qui existe entre considérer que la causalité est contrefactuelle, comme le font les approches de type Halpern, et établir des relations en utilisant un raisonnement contrefactuel sur des relations de causalité :

> Notice that this counterfactual ("would have caused") is a counterfactual about causation, rather than an analysis of causation as counterfactual. The biochemical process that would have killed the plant would have caused the plant to die. The gardener, by watering the plant, interacts with that process to prevent that causing.

En plus de cette première différence, les définitions 7.6 à 7.9 n'utilisent pas la notion d'interventionnisme. Toutes les traces qui sont explorées avec les différents scénarios respectent la dynamique du monde décrite par le contexte κ_c.

Nous voulons en profiter pour souligner l'importance d'utiliser un langage suffisamment riche pour représenter la causalité effective. Dans le chapitre 5, nous avons vu que les STEE permettent de clarifier ce qu'est la surdétermination, mais aussi de distinguer les désaccords causaux fondamentaux qui restent dans le domaine, des désaccords causaux non fondamentaux. Dans le chapitre 6 et celui-ci, si nous avons pu décortiquer la causalité en différentes notions pour pouvoir les étudier le plus indépendamment possible, c'est grâce à l'utilisation des STEE. En effet, ceux-ci nous offrent une sémantique permettant de distinguer une occurrence d'évènement d'une non occurrence d'évènement, ou une occurrence d'un évènement de la véracité d'une formule. Comme nous l'avons vu dans les différents chapitres traitant de causalité, cette dernière est une notion complexe. En effet, elle est composée de nombreuses autres notions et partage une frontière très fine avec la responsabilité. Ne pas pouvoir séparer clairement les différentes notions qui la compose rend la tâche de formalisation encore plus complexe.

Nous avons fait des propositions pour différents points de vue de ce qu'est la responsabilité en omission. La similarité entre ces propositions et les approches contrefactuelles de type Halpern, aussi bien dans le besoin de raisonnement contrefactuel que dans la non transitivité des relations que nous discuterons dans la section suivante, nous fait nous in-

terroger. Ces approches sont pour la plupart représentées en SEF où il n'existe aucune différence sémantique entre l'occurrence d'un évènement et sa non occurrence. Serait-il possible que les intuitions sur lesquelles ces définitions reposent, dont la place centrale qu'elles accordent à la nécessité, soient influencées par le besoin de traiter la causalité négative et positive de la même façon ? Répondre à cette question est en dehors du cadre de cette thèse. Il s'agit toutefois d'une perspective intéressante car elle montrerait que le domaine de la causalité effective peut également s'enrichir en dialoguant avec le domaine qui s'intéresse à la représentation de l'action et du changement.

Avant de clore cette section, il est important de préciser que les définitions proposées dans cette section n'ont pour but que de montrer que l'approche de causalité proposée est une base appropriée pour représenter toutes les formes de négation dans la relation causale, malgré leur aspect normatif. Leur fonctionnement a été montré pour un cas simple avec un nombre d'actions limité. Pour prétendre proposer une définition satisfaisante de responsabilité pour ces cas de négation dans la relation causale, il faut approfondir la discussion à des exemples avec plus d'actions possibles et des contraintes de faisabilité de ces actions. L'ensemble de décisions peut devenir très complexe, il est nécessaire pour chaque vision de développer des règles claires qui permettent de traiter ces cas dans toute leur complexité.

7.4 Modélisation et représentation de la volition comme facteur de transitivité

Dans cette section nous traitons la dernière marche de notre escalier, la transitivité. Une fois construite elle permet de relier les actions à leurs conséquences indirectes tout en évitant le problème de ramification. Toutefois, comme nous l'avons vu dans la section 3.2.3, savoir si la causalité est transitive suscite de nombreux débats dans la communauté [MENZIES et BEEBEE, 2020]. Comme pour les autres notions traitées dans ce chapitre, les débats autour de la transitivité semblent être reliés à des questions de responsabilité. Nous allons séparer cette discussion sur la transitivité en deux parties : la transitivité des relations causales d'un côté, puis celle des relations de responsabilité de l'autre. Nous traiterons la première dans la section 7.4.1 et la deuxième dans la section 7.4.2.

7.4.1 La transitivité des relations causales

La volition des agents semble jouer un rôle important dans l'intuition que nous avons de la transitivité des relations causales. La volition faisant référence à l'état mental des agents, elle est liée à des aspects normatifs. Définir formellement une approche gérant tous les cas de transitivité qui convienne à toutes les théories morales est donc impossible. Toutefois, nous proposons d'introduire une nuance permettant à notre approche factuelle de pouvoir servir comme base pour modéliser les différents points de vue.

Commençons par préciser ce que nous entendons par transitivité des relations causales. Nous avons proposé des définitions pour toutes les relations apparaissant sur le tableau 7.1. Il existe sept cas de transitivité de base possibles en combinant ces différentes relations. Seul un de ces cas correspond à ce dont nous parlons dans cette section, le seul cas entièrement factuel. Il s'agit du cas où une occurrence d'évènement (e, t) est une cause effective d'une autre occurrence d'évènement (e_φ, t_φ), qui à son tour est cause effective d'une non occurrence d'évènement $(\overline{e_\psi}, t_\psi)$. Cela s'exprime formellement :

$$(e, t) \rightsquigarrow (e_\varphi, t_\varphi) \rightsquigarrow (\overline{e_\psi}, t_\psi).$$

Nous avons ici un cas lié à la notion d'empêcher, mais uniquement au premier niveau de raisonnement, celui factuel. Notez que ce cas est le seul en supposant que nous combinons entre elles les relations causales du tableau 7.1, sans prendre deux fois la même. Toutefois, lorsque nous parlons de transitivité des relations causales, nous pensons également au cas où une occurrence d'évènement (e, t) est une cause effective d'une autre occurrence d'évènement (e_φ, t_φ), qui à son tour est cause effective d'une occurrence d'évènement (e_ψ, t_ψ). Cela s'exprime formellement :

$$(e, t) \rightsquigarrow (e_\varphi, t_\varphi) \rightsquigarrow (e_\psi, t_\psi).$$

La question à laquelle nous allons essayer de répondre est : pouvons nous déduire dans le premier cas que $(e, t) \rightsquigarrow (\overline{e_\psi}, t_\psi)$ et dans le deuxième cas que $(e, t) \rightsquigarrow (e_\psi, t_\psi)$? Tout d'abord, montrons le lien qui existe entre transitivité et volition des agents. Pour cela, intéressons nous au problème suivant discuté par HITCHCOCK [2007] :

Exemple 7.11 [assassin et garde du corps]**.** *Un assassin met du poison dans le café d'une cible. Son garde du corps réagit en mettant un antidote dans le café, ce qui neutralise le poison. L'antidote seul ne représente pas de danger pour la santé. La cible boit le café et survit. Si l'assassin n'avait pas empoisonné le café, la garde du corps n'aurait pas administré l'antidote. Si elle n'avait pas administré l'antidote, la victime serait morte du poison.*

Dans cet exemple, malgré l'existence d'une chaîne causale tout à fait factuelle entre l'assassin qui met du poison et la survie de la cible, HITCHCOCK [2007] défend que la plupart des individus n'ont pas envie de dire que l'assassin est une cause de la survie de la cible : « Assassin's poisoning the coffee caused Bodyguard to administer the antidote, and Bodyguard's administering the antidote caused Victim to survive, but most people judge that Assassin's poisoning the coffee is not a token cause of Victim's survival ». Cette intuition qui semble contredire la transitivité de la causalité est liée à la notion de volition. Dans la description qui est faite de l'exemple, il est indiqué que « Si l'assassin n'avait pas empoisonné le café, la garde du corps n'aurait pas administré l'antidote ». Ce raisonnement hypothétique qui met en évidence une supposée dépendance causale est ce qui permet à HITCHCOCK [2007] d'établir une relation de causalité entre l'assassin qui empoisonne le café et la garde du corps qui verse le poison. Mais ce raisonnement hypothétique, comme tout raisonnement hypothétique est normatif et véhicule une vision subjective de la situation. Étant donné le raisonnement hypothétique qui est fait ici, il serait plus judicieux de représenter la garde du corps comme un évènement naturel plutôt que comme un agent qui peut décider de réaliser une action. En effet, celui-ci semble simplement réagir aux stimulis externes, comme un évènement naturel qui se déclenche dès que ses conditions de déclenchement sont satisfaites. Il pourrait tout à fait être remplacé par un robot qui avec un capteur détecte si la boisson de son propriétaire est empoisonnée. La relation entre la cible et la garde du corps peut expliquer cette vision. Par sa position et le contrat qui l'engage, la garde du corps a un devoir de protéger la cible. Que se passerait-il si, au lieu de faire intervenir une garde du corps, l'exemple faisait référence à l'infirmier de la cible qui verse le traitement quotidien de la cible qui s'avère être, par pur hasard, l'antidote ? Certes l'exemple dans sa globalité deviendrait plus difficilement crédible, mais cette variante permet de mettre en évidence

l'aspect normatif d'établir une relation causale entre l'assassin qui empoisonne le café et la garde du corps qui verse le poison. En modifiant le métier de l'individu et ses intentions, deux informations qui n'ont rien de causales, le lien de causalité établit semble ne plus être possible. Qui plus est, la garde du corps peut tout à fait mettre l'antidote sans que la boisson ne soit empoisonnée ou ne pas mettre l'antidote et laisser mourir la cible ; elle et l'infirmier sont des agents dotés d'une volition.
L'essentiel de l'approche causale que nous avons proposée a été présenté dans la section 6.2 et 7.2.1. Dans celle-ci, la transitivité de la causalité s'arrête du moment où la volition d'un agent intervient dans la chaîne causale. Dans son état actuel, notre approche n'établira pas de relation de causalité entre l'assassin qui empoisonne le café et la garde du corps qui verse le poison. Nous avons adopté l'idée que, bien qu'elle puisse être sujette à des influences extérieures, du moment où il est question de volition, il y a une part de décision propre à l'agent qui est imprévisible. Celle-ci ne peut pas être causée. Pour rappel, nous avons adopté la vision de REVAZ [2009] qui défend que l'occurrence d'un évènement naturel « [...] peut être expliquée par des lois alors que l'agir humain ne peut être que compris, c'est-à-dire interprété ». Notre approche étant conçue pour l'éthique computationnelle, ce choix se justifie. En effet, sans une forme de volition avec une part d'imprévisible, il ne ferait aucun sens de parler d'éthique puisque de toute façon les agents ne décident de rien et tout est le simple résultat de facteurs externes.
Toutefois, certaines décisions qu'un agent prend peuvent s'avérer cruciales pour qu'un autre agent puisse agir. Si cette information n'appartient pas au domaine de la causalité effective, elle peut être pertinente pour certaines théories morales. Les exemples mentionnés dans la section 3.2.3 et que nous rappelons montrent que cette intuition est communément admise en droit. Lorsqu'un meurtre est commis par un tueur à gage, le processus judiciaire ne se limite pas à juger le tueur à gage, le commanditaire est également recherché, même si ce n'est pas lui qui a commis le meurtre directement. La personne qui a vendu l'arme avec laquelle le meurtre a été commis peut également être recherchée. Dans l'exemple 7.11, bien qu'il soit incorrect de dire que l'assassin a causé l'action du garde du corps, il a tout de même eu une influence sur la décision du garde du corps. Selon la théorie morale formalisée, cette relation peut avoir une importance en éthique computationnelle. Nous voulons être en mesure de pouvoir raisonner sur toutes les conséquences des décisions d'un agent, même si celles-ci dépendent également de la volition d'un autre agent.

Pour proposer une approche causale qui puisse être une base commune aux différentes théories morales, nous adoptons l'idée d'intégrer dans notre approche la notion de « permettre ». Cette notion doit être comprise dans le sens de rendre possible quelque chose ; elle correspond en anglais à « enables » et a fait l'objet de travaux dans différents domaines [BERREBY et collab., 2018; CHOI et FARA, 2021; LEWIS, 1997; MARTIN, 1994; SLOMAN et collab., 2009]. Plus spécifiquement, nous adoptons la modélisation de cette notion proposée par BERREBY et collab. [2018] qui défendent que rendre vraies les conditions de déclenchement d'un évènement naturel ne peut pas être mis sur un pied d'égalité avec le fait de rendre vraies les préconditions d'une action dont l'occurrence dépend également du choix indépendant de l'agent de la réaliser ou non. Si le premier cas donne lieu à la relation de causalité effective à laquelle nous nous sommes intéressés jusqu'à présent, le deuxième correspond à la relation permettre.
Cette solution semble résoudre les cas problématiques discutés dans le domaine. Prenons une version de l'exemple 7.11 où la garde du corps est un agent avec une volition. Plus exac-

tement, nous intégrons l'idée que la garde du corps décide si verser l'antidote mais qu'il ne peut le faire que si l'assassin empoisonne le café. Dans ce cas, cette solution dira que l'assassin permet au garde du corps de verser l'antidote, que verser l'antidote sauve la cible, et donc que l'assassin permet au garde du corps de sauver la cible. Prenons maintenant l'exemple 3.14 du cambriolage où un agent ouvre la porte de chez lui, condition pour que le voleur puisse entrer et voler ses biens. Dans ce cas, cette solution dira que l'habitant permet au voleur de commettre son délit. Prenons l'exemple 3.15 où un individu poignarde un autre et qu'en raison de ses convictions religieuses la victime refuse un traitement médical et meurt. Dans ce cas, cette solution dira que le criminel permet la mort de la victime. Finalement, prenons l'exemple de la bioéthique appliquée où la différence entre causer et permettre un évènement semble cruciale et bien conceptualisée. Cela est particulièrement vrai lorsqu'il est question de fin de vie avec la différence entre l'euthanasie et le suicide assisté. La distinction cruciale entre tuer et laisser mourir repose sur l'hypothèse d'un type différent de responsabilité morale lorsqu'il s'agit de causer ou de permettre qu'un évènement se produise.

Pour intégrer la notion de permettre à notre approche de causalité il suffit de remplacer la fonction tri par la fonction pre dans les définitions 6.12 et 7.1.

Définition 7.10 [*Permet*]**.** *Étant donné un cadre causal* χ *et deux occurrences d'évènements* (e,t) *et* (e_ψ, t_ψ)*,* (e,t) *permet* (e_ψ, t_ψ)*, ssi* (e,t) *est une NESS-cause de* $(pre(e_\psi), t_\psi)$*.*

Définition 7.11 [*Force une non occurrence*]**.** *Étant donné un cadre causal* χ*, une occurrence d'évènement* (e,t) *et une non occurrence d'évènement* $(\overline{e_\psi}, t_\psi)$*,* (e,t) *force la non occurrence* $(\overline{e_\psi}, t_\psi)$*, ssi* (e,t) *est une NESS-cause de* $(\neg pre(e_\psi), t_\psi)$*.*

Dans la distinction, Berreby et collab. [2018] intègrent l'idée que cette nouvelle relation est moins forte que celle de cause effective. De fait, bien que nous puissions être pleinement responsables de nos actions et des évènements naturels que nous déclenchons, nous ne pouvons pas être pleinement responsables des choix des autres. Si nous partageons pleinement ce point de vue, il s'agit là d'un parti pris que nous ne pouvons pas intégrer dans notre approche car cela pourrait ne pas correspondre à toutes les théories morales. Si nous reprenons l'exemple 3.15 et que nous ajoutons le fait que le criminel connaissait les convictions religieuses de la victime et donc savait qu'elle refuserait tout traitement, alors son acte peut apparaître pour certaines théories morales comme plus grave. Si par contre nous ajoutons qu'il ignorait cela et qu'en plus la blessure infligée ne représentait aucun danger sous condition d'accepter quelques points de suture, alors c'est l'inverse. Prendre en compte les connaissances de l'agent semble influencer l'intuition que nous avons sur le problème. Il faut garder à l'esprit que cette différenciation n'est pas une notion de causalité effective mais de responsabilité.
En intégrant la nuance à notre approche causale nous faisons en sorte que différents points de vue puissent être formalisés à partir de celle-ci. Si la théorie morale considère qu'il n'y a aucune différence entre causer et permettre, il suffit de considérer les relations de permettre trouvées par notre approche comme des relations de causalité effectives. Si la théorie morale considère que les deux notions doivent être prises en compte différemment car il s'agit de relations avec des forces différentes, l'approche différencie les deux par défaut. Finalement, si la théorie morale considère que la notion de permettre ne doit en aucun cas être prise en compte, il suffit de considérer uniquement les relations de causalité effective.

7.4.2 La transitivité des relations de responsabilité

Pour rappel, sept cas de transitivité possibles peuvent être obtenus en combinant les différentes relations apparaissant sur le tableau 7.1. Nous avons traité un de ces cas dans la section précédente. Dans cette section nous allons parler des six cas restants. Les six relations qui y correspondent sont toutes des relations qui font intervenir de la responsabilité.

Commençons par parler de la relation associée au deuxième niveau de raisonnement de la notion d'empêcher. Il s'agit du cas où une occurrence d'évènement (e,t) est une cause effective d'une non occurrence d'évènement $(\overline{e_\varphi}, t_\varphi)$, qui à son tour est responsable d'une non occurrence d'évènement $(\overline{e_\psi}, t_\psi)$. Cela s'exprime formellement :

$$(e,t) \rightsquigarrow (\overline{e_\varphi}, t_\varphi) \hookrightarrow (\overline{e_\psi}, t_\psi).$$

Dans ce cas là, il semble correct de pouvoir dire que $(e,t) \hookrightarrow (\overline{e_\psi}, t_\psi)$, puisque (e,t) cause une non occurrence responsable de $(\overline{e_\psi}, t_\psi)$. Ce cas est celui de l'exemple 7.8 où puisque l'agent A arrête la balle, $(int_a, 0) \rightsquigarrow (\overline{nxt_1}, 1)$, et que le fait que la balle ne soit pas en mouvement est responsable qu'elle n'impacte pas la fenêtre, $(\overline{nxt_1}, 1) \hookrightarrow (\overline{imp}, 4)$, alors nous voulons conclure que l'agent A est responsable du fait que la fenêtre ne soit pas impactée par la balle, $(int_a, 0) \hookrightarrow (\overline{imp}, 4)$.

Passons au cas suivant. Il s'agit du cas où une occurrence d'évènement (e,t) est une cause effective d'une non occurrence d'évènement $(\overline{e_\varphi}, t_\varphi)$, qui à son tour est responsable d'une occurrence d'évènement (e_ψ, t_ψ). Cela s'exprime formellement :

$$(e,t) \rightsquigarrow (\overline{e_\varphi}, t_\varphi) \hookrightarrow (e_\psi, t_\psi).$$

Dans ce cas là, il semble correct de pouvoir dire que $(e,t) \hookrightarrow (e_\psi, t_\psi)$, puisque (e,t) cause une non occurrence responsable de (e_ψ, t_ψ). Ce cas est celui d'un agent qui empêcherait un maître nageur d'aller sauver une personne en détresse $(e,t) \rightsquigarrow (\overline{e_\varphi}, t_\varphi)$, et comme la non intervention du maître nageur est responsable de la mort de l'individu $(\overline{e_\varphi}, t_\varphi) \hookrightarrow (e_\psi, t_\psi)$, alors nous voulons conclure que l'agent est responsable de la mort de l'individu $(e,t) \hookrightarrow (e_\psi, t_\psi)$.

Nous avons traité tous les cas où le premier élément de la chaîne était une occurrence d'évènement. Passons aux cas où le premier élément est une omission. Le premier cas est celui communément appelé « fail to prevent » où une non occurrence d'évènement $(\overline{e}, t)$ est responsable d'une occurrence d'évènement (e_φ, t_φ), qui à son tour est une cause d'une occurrence d'évènement (e_ψ, t_ψ). Cela s'exprime formellement :

$$(\overline{e}, t) \hookrightarrow (e_\varphi, t_\varphi) \rightsquigarrow (e_\psi, t_\psi).$$

Dans ce cas là, il semble correct de pouvoir dire que $(\overline{e}, t) \hookrightarrow (e_\psi, t_\psi)$, puisque $(\overline{e}, t)$ est responsable d'une occurrence qui cause (e_ψ, t_ψ). Cela veut dire que $(\overline{e}, t)$ aurait pu faire la différence, elle aurait pu empêcher (e_φ, t_φ). Ce cas est celui d'un agent qui malgré avoir tous les éléments pour empêcher un incendie de se déclarer ne le fait pas, ce qui le rend responsable du déclenchement de l'incendie $(\overline{e}, t) \hookrightarrow (e_\varphi, t_\varphi)$. Comme l'incendie cause la perte d'une grande partie de la biodiversité de la région $(e_\varphi, t_\varphi) \rightsquigarrow (e_\psi, t_\psi)$, alors nous voulons conclure que l'agent est responsable de cette perte de biodiversité $(\overline{e}, t) \hookrightarrow (e_\psi, t_\psi)$.

Le cas suivant est celui communément appelé « fail to cause » où une non occurrence d'évènement $(\overline{e}, t)$ est responsable d'une occurrence d'évènement (e_φ, t_φ), qui à son tour est une cause d'une non occurrence d'évènement $(\overline{e_\psi}, t_\psi)$. Cela s'exprime formellement :

$$(\overline{e}, t) \hookrightarrow (e_\varphi, t_\varphi) \rightsquigarrow (\overline{e_\psi}, t_\psi).$$

Dans ce cas là, il semble correct de pouvoir dire que $(\overline{e}, t) \hookrightarrow (\overline{e_\psi}, t_\psi)$, puisque $(\overline{e}, t)$ est responsable d'une occurrence qui cause $(\overline{e_\psi}, t_\psi)$. Cela veut dire que $(\overline{e}, t)$ aurait pu faire la différence, elle aurait pu empêcher (e_φ, t_φ). Ce cas est celui d'un agent qui, malgré avoir pu empêcher une affaire de corruption en dénonçant son chef de parti politique aux autorités, ne le fait pas, ce qui le rend responsable d'une perte d'argent public $(\overline{e}, t) \hookrightarrow (e_\varphi, t_\varphi)$. Comme cette perte cause la non création de nouveaux postes d'enseignants $(e_\varphi, t_\varphi) \rightsquigarrow (\overline{e_\psi}, t_\psi)$, alors nous voulons conclure que l'agent est responsable de cette absence de création de postes dans l'enseignement public $(\overline{e}, t) \hookrightarrow (\overline{e_\psi}, t_\psi)$.

Nous avons traité tous les cas où il y avait au moins un élément factuel. Passons aux cas où il est uniquement question de responsabilité. Dans les deux cas restants, il est question d'une non occurrence d'évènement $(\overline{e}, t)$ qui est responsable d'une non occurrence d'évènement $(\overline{e_\varphi}, t_\varphi)$, qui à son tour est soit responsable d'une occurrence d'évènement (e_ψ, t_ψ), soit d'une non occurrence d'évènement $(\overline{e_\psi}, t_\psi)$. Ces deux cas s'expriment respectivement :

$$(\overline{e}, t) \hookrightarrow (\overline{e_\varphi}, t_\varphi) \hookrightarrow (e_\psi, t_\psi),$$

$$(\overline{e}, t) \hookrightarrow (\overline{e_\varphi}, t_\varphi) \hookrightarrow (\overline{e_\psi}, t_\psi).$$

Dans ces cas là, cela ne semble pas correct de pouvoir dire que $(\overline{e}, t) \hookrightarrow (e_\psi, t_\psi)$ ou que $(\overline{e}, t) \hookrightarrow (\overline{e_\psi}, t_\psi)$. En effet, la relation de responsabilité telle qu'elle a été définie ne semble pas transitive, ce n'est pas parce qu'une omission $(\overline{e}, t)$ aurait pu faire une différence par rapport à une non occurrence d'évènement $(\overline{e_\varphi}, t_\varphi)$ et que celle-ci aurait pu faire une différence par rapport à une occurrence (e_ψ, t_ψ) ou une non occurrence $(\overline{e_\psi}, t_\psi)$, que $(\overline{e}, t)$ aurait pu faire une différence par rapport à (e_ψ, t_ψ) ou $(\overline{e_\psi}, t_\psi)$. Prenons une variante de l'exemple 7.10 où la balle est en fait un dispositif pour éteindre un incendie qui s'est déclaré dans la maison. Il se trouve qu'une fois passé la fenêtre, le dispositif allait se déclencher grâce à la chaleur et éteindre le feu. Le fait que la relation de responsabilité ne soit pas transitive ne veut pas dire que l'agent qui arrête la balle ne peut pas être responsable du fait que l'incendie ne soit pas éteint. Cependant, cela ne peut pas être déterminé sans se poser la question si l'agent qui arrête la balle aurait pu faire la différence par rapport à l'incendie directement. Cela est tout à fait possible puisque ce raisonnement hypothétique demande de s'interroger sur un cas de causalité positive pour lequel la transitivité existe.
Les définition 7.6 à 7.9 ne s'appliquent pas nécessairement qu'à des éléments proches temporellement. En effet, il est possible d'être responsable d'une occurrence ou une non occurrence d'évènement, même si d'autres occurrences d'évènements nous séparent de celles-ci. Toutefois, selon la vision de responsabilité qui est adoptée, cela ne semble pas toujours pouvoir être fait par transitivité. Cette observation est une piste intéressante à explorer pour expliquer pourquoi les approches contrefactuelles ont tendance à refuser la transitivité dans leur approche causale.

7.5 Conclusion

Dans ce chapitre nous avons esquissé les trois dernières marches composant notre approche causale commune pour l'éthique computationnelle. Plus exactement, nous avons vu que la première marche qui concerne la causalité positive est le cœur de l'approche, la seul des marches pouvant être commune entièrement à toutes les théories morales. Les trois marches dont nous avons parlé dans ce chapitre vont au delà de l'aspect purement causal. Pour être entièrement construites, elles demandent d'étudier pour chaque théorie morale quels choix liés à la responsabilité correspondent à chaque théorie. Toutefois, nous avons montré que l'approche que nous proposons peut servir comme base pour le développement de ces autres marches. Voyons cela dans les détails.

Nous avons commencé par traiter les deux cas où la cause est une non occurrence d'évènement. C'est ce que nous avons appelé omission. Nous avons montré que l'omission n'a pas de statut causal et qu'il s'agit purement d'une question de responsabilité.
Nous sommes partis de l'observation que dans un STEE une non occurrence ne produit rien, le monde reste inchangé. Toutefois, si un agent avait la possibilité d'agir et donc d'a priori changer le monde, le fait qu'il ne l'ait pas fait le rend dans certaines conditions imputable du fait que le monde soit resté tel qu'il est. La question qui s'est posée alors était : existe-t-il une relation causale factuelle comme celle de la causalité positive? En étudiant les cas de surdétermination, nous avons vu que la différence d'intuition entre la causalité positive et l'omission n'était pas négligeable. Si le problème posé par les problèmes de surdétermination était le même, à savoir mettre en échec l'idée de nécessité de la cause, l'intuition pour le résoudre ne l'était pas. Contrairement aux cas de causalité positive où il existe une structure causale sur laquelle il est possible de s'appuyer, dans les cas d'omission il n'y a pas de structure qui permette cela. Aucune chaîne causale existante ne relie une omissions à la conséquence puisque par définition une non occurrence d'évènement ne peut être à l'origine d'un changement.
L'étude d'exemples utilisés dans le domaine de la causalité a montré que si l'intuition concernant les cas de responsabilité n'est pas la même pour les cas classiques de causalité positive et ceux contenant une omission, c'est principalement dû au fait que pour que l'omission soit pertinente, il faut qu'il existe la possibilité de faire la différence par rapport à ce qui s'est réellement passé. Nous avons montré que c'est dans le choix des mondes qu'il est possible d'envisager dans le raisonnement hypothétique que repose la décision de quelles conséquences pourront être attribuées à chaque non occurrence. Cette décision dépendant d'un grand nombre de facteurs en dehors de la causalité, nous avons conclu que l'omission est essentiellement une question de responsabilité. Plus exactement, nous avons mis en lumière que l'élément qui s'avère décisif semble être le devoir des agents. Déterminer cela n'est possible qu'en adoptant un point de vue spécifique qui ne peut convenir à toutes les théories morales. Pour cette raison, proposer une définition de comment prendre en compte l'omission est en dehors du cadre de l'approche commune que nous nous sommes fixés de proposer.
La conclusion autour de la notion d'omission, et plus généralement la non occurrence d'évènements, est qu'elles appartiennent au champ de la responsabilité plus qu'à celui de la causalité. La prise en compte des non occurrences d'évènements n'affecte en rien l'enquête causale classique, elle n'est en réalité qu'une considération supplémentaire qui peut jouer un rôle dans les autres étapes permettant de déterminer la responsabilité.

Après nous être intéressés aux cas où les causes sont des non occurrences d'évènements, nous avons étudié les cas où la conséquence est une non occurrence d'évènement. C'est ce que nous avons appelé conséquences négatives, ou par abus de langage « empêcher ». Nous avons montré que la notion d'empêcher peut être décomposée en deux niveaux de raisonnement, un appartenant à la causalité effective, le deuxième étant une question de responsabilité et de transitivité.
Nous sommes partis de l'observation qu'au contraire de l'omission, et plus généralement aux cas avec des causes négatives, les cas impliquant des conséquences négatives peuvent être traités factuellement. Nous avons représenté dans $\mathscr{S}_c$ l'approche factuelle pour traiter ces cas et nous avons étudié sa sensibilité à la surdétermination, comme pour la causalité positive dans le section 6.2.4. Pour cela, nous avons modélisé la surdétermination dans le cas de conséquences négatives et établi une typologie en suivant la même procédure que dans le chapitre 5. Cette typologie a montré qu'il existe bien une différence entre la causalité positive et ce cas de négation dans la conséquence. Cette approche factuelle s'intègre parfaitement à l'approche de causalité positive proposée dans le chapitre 6.
Malgré l'aspect factuel de cette solution, nous avons vu en la confrontant aux exemples classiques dans le domaine, que celle-ci ne correspondait pas tout à fait à la notion d'empêcher qui vient intuitivement à l'esprit, du moins dans un type de cas en particulier. Nous avons expliqué ce décalage en montrant que la notion d'empêcher peut être décomposée en deux niveaux de raisonnement, un factuel et un normatif. Dans les cas qui incombent au premier niveau, l'approche factuelle que nous avons proposée est satisfaisante. Les cas qui posent problème font intervenir le deuxième. Nous avons mis en lumière que, si ce deuxième niveau pose des problèmes à l'approche factuelle, c'est parce qu'au delà du premier niveau nous ne sommes plus dans le terrain de la causalité effective, mais dans celui de la responsabilité. En effet, traiter le deuxième niveau demande d'attribuer des conséquences à la non occurrence d'évènements, ce qui nous l'avons vu est une question de responsabilité. La notion d'empêcher faisant intervenir des aspects normatifs, il peut exister autant de points de vue que de personnes qui se penchent sur le problème.

Une fois montré qu'une modélisation de la notion d'omission et d'empêcher ne peut être intégrée dans une approche se voulant commune à différentes théories morales, nous avons tenu à montrer que notre approche factuelle est une base pour une telle modélisation.
Pour cela, nous avons commencé par introduire à $\mathscr{S}_c$ la notion de décision et d'ensemble de scénarios envisageables. En nous appuyant sur ces éléments nous avons proposé une définition de responsabilité par omission pour quatre points de vue différents. Cela a été fait pour les deux types de relations causales qui nous intéressent dans un STEE. Nous couvrons ainsi, aussi bien les $\mathscr{F} - causes$ de la définition 5.4 que les causes effectives de la définition 5.5. Cela a montré que l'approche de causalité proposée est une base appropriée pour représenter toutes les formes de négation dans la relation causale, malgré leur aspect normatif.
Une fois ces quelques propositions faites, nous avons discuté des similarités et des différences avec les définitions de causalité de type Halpern présentées dans la section 3.1.1.2. Nous avons observé que les définitions proposées partagent avec les définitions de type Halpern la notion de contrefactuelle. Toutefois, nous avons souligné que le raisonnement contrefactuel qui est fait dans les définitions 7.6 à 7.9 ne fait pas de ces définitions des approches contrefactuelles au même titre que celles de type Halpern. En effet, nous avons sou-

levé deux points de divergence. Le premier est que les propositions faites ne considèrent pas que la causalité est contrefactuelle, elles se contentent d'appliquer un raisonnement contrefactuel sur des scénarios où la causalité est factuelle. Le deuxième est que les propositions faites n'utilisent pas la notion d'interventionnisme. Toutes les traces qui sont explorées avec les différents scénarios respectent la dynamique du monde décrite par le contexte.

Pour finir, nous avons abordé la question de la transitivité pour laquelle nous avons montré que les débats la concernant faisaient également intervenir des notions de responsabilité.
Pour commencer, nous avons établi que la volition des agents semble jouer un rôle important dans l'intuition que nous en avons. Cette notion faisant référence à des états mentaux, elle introduit des aspects normatifs dans l'intuition. Cela nous a amené à conclure, comme précédemment, qu'une approche gérant tous les cas de transitivité et qui convienne à toutes les théories morales ne peut exister. Dans notre approche causale, la transitivité de la causalité s'arrête du moment où la volition d'un agent intervient dans la chaîne causale. Toutefois, nous avons exposé au travers d'exemples que certaines décisions qu'un agent prend peuvent s'avérer cruciales pour qu'un autre agent puisse agir. Si cette information n'appartient pas au domaine de la causalité effective, elle peut être pertinente pour certaines théories morales et donc nous voulions aller plus loin. Nous voulions être en mesure de pouvoir raisonner sur toutes les conséquences des décisions d'un agent, même si celles-ci dépendent également de la volition d'un autre agent. Nous avons alors proposé d'introduire une nuance permettant à notre approche factuelle de pouvoir servir de base pour modéliser les différents points de vue. Cela est passé par l'ajout de la notion de « permettre » qui a été fait en adoptant la modélisation de cette notion proposée par BERREBY et collab. [2018]. Intégrer cette notion permet de résoudre les cas problématiques discutés dans le domaine.

Avant de clore ce chapitre, il est important de faire une dernière remarque. Malgré tous les efforts faits pour proposer une approche factuelle, du moment où il est question de modélisation il semble compliqué de pouvoir se débarrasser de toute subjectivité. Choisir une équation pour représenter les préconditions plutôt qu'une autre est déjà un moyen de choisir quels mondes nous considérons. Dans toute modélisation un choix est fait sur ce qui doit être représenté, ce qui ne doit pas l'être et comment il doit l'être. Ce choix est également sujet a des aspects normatifs. Il semble impossible d'échapper à tous les choix et être entièrement objectifs lorsqu'une modélisation est faite. Toutefois, identifier les endroits où les choix sont faits et les séparer des endroits où ils ne doivent pas l'être ne peut être que bénéfique.

Chapitre 8

Contribution : formalisation de la causalité positive appliquée à l'argumentation abstraite

« *JUDGE : One man is dead. The life of another is at stake. I urge you to deliberate honestly and thoughtfully. If this is a reasonable doubt - then you must bring me a verdict of 'not guilty'.* »

ROSE et SERGEL [1983]

Sommaire

Ce chapitre constitue la première étape d'une proposition visant à appliquer les outils développés dans le chapitre 6 pour la représentation de l'action, du changement et de la causalité afin d'aborder les questions qui se posent dans l'argumentation abstraite lorsqu'il est question de causalité et de temporalité ensemble. Il s'agit des résultats d'un travail en collaboration étroite avec Yann Munro qui s'intéresse à l'explicabilité pour l'argumentation abstraite.

Un système d'argumentation abstrait (AAF) offre un cadre propice pour représenter et raisonner sur des informations contradictoires par l'intermédiaire d'arguments. Ce formalisme permet de trouver des ensembles d'arguments pouvant être acceptés et fournit des explications sur les raisons pour lesquelles ces ensembles ont été acceptés ou non. Les AAF proposent donc des outils appropriés pour modéliser et raisonner sur des dialogues. Cependant, il s'agit d'un cadre statique qui n'inclut pas de notion de temporalité qui semble cruciale pour modéliser des dialogues. Pour résoudre ce problème, plusieurs types d'approches ont été proposées. Le formalisme logique YALLA [Saint-Cyr et collab., 2016] propose de réécrire un AAF puis d'utiliser des opérateurs de révision ou de mise à jour de croyances pour mettre à jour le système d'argumentation. Il offre un langage très expressif qui permet de trouver quelles relations d'attaque ou quels arguments devraient être supprimés ou ajoutés afin d'atteindre un objectif donné au pas de temps suivant. Un tel raisonnement peut être utilisé pour construire une explication. Cependant, cette approche ne semble pas adaptée à notre problème étant donné que nous visons à modéliser tout le dialogue, et que dans YALLA le nombre de termes augmente de façon exponentielle avec le nombre d'arguments. Doutre et collab. [2017] proposent d'utiliser un langage propositionnel dynamique pour modéliser le graphe d'argumentation et son évolution à travers le dialogue. Ce faisant, il répond partiellement à notre question en intégrant avec succès la notion de temporalité. Puisque le but de ce travail à long terme est de proposer une explication de la raison pour laquelle un argument est ou n'est pas acceptable à un moment donné du dialogue, cette proposition n'est pas suffisante. En effet, pour y parvenir, une étude formelle de la notion de causalité est nécessaire.

Même s'il n'existe pas de définition unique de la notion d'explication, il est admis, du point de vue des sciences sociales, qu'une explication est une réponse à une question de type « *pourquoi* », comme l'indique Miller [2019]. Elle est donc profondément liée à la notion de causalité, ce qui justifie la nécessité d'une étude formelle de cette notion. La causalité a déjà fait l'objet d'études pour les AAF. Récemment, Bengel et collab. [2022] ont utilisé un autre formalisme fondé sur la logique pour définir le raisonnement contrefactuel pour l'argumentation structurée. Leur définition est l'application directe de la définition de Pearl et Neuberg [2000] pour le modèle causal, définition qui correspond au but-for test présenté dans le chapitre 3. Avant cela, Bochman [2005] a établi une équivalence entre l'argumentation abstraite et un système de raisonnement causal. Ce système a ensuite été étendu à des raisonnements causaux plus complexes [Bochman, 2021]. Cependant, tous ces travaux portent sur des AAF classiques sans temporalité.

Nous proposons d'utiliser $\mathscr{S}_c$ et notre approche causale afin de pouvoir modéliser la dynamique d'un dialogue et étudier les relations causales qui s'y trouvent. En effet, les langages de description d'action, comme ceux dont nous avons parlé dans le chapitre 2 ou $\mathscr{S}_c$ proposé dans la section 6.1, ont été naturellement conçus pour inclure cette notion dans la représentation. Nous avons choisi d'utiliser le langage proposé dans le chapitre 6 pour trois raisons principales. Tout d'abord, il semble pertinent de choisir un langage avec une expres-

sivité intermédiaire. Comme pour la causalité et l'éthique, la cooccurrence d'évènements et la disjonction sont des concepts intéressants lorsque nous souhaitons représenter le raisonnement sur des arguments, contrairement aux actions non déterministes ou les actions duratives qui ne semblent pas être des concepts utiles dans ce raisonnement. Ensuite, en choisissant $\mathscr{S}_c$ nous pouvons utiliser notre proposition de modélisation et représentation de causalité positive présentée dans la section 6.2. Enfin, ce choix nous permet également d'aller jusqu'à l'étape d'automatisation en nous appuyant sur l'implémentation en ASP qui a été présentée dans la section 6.3.

Ce chapitre est organisé comme suit. La section 8.1 présente brièvement le formalisme des AAF de DUNG [1995]. La section 8.2, détaille les principales contributions de ce chapitre : une réécriture des AAF acycliques dans $\mathscr{S}_c$, l'implémentation associée et quelques propriétés de cette transformation. Elles concernent principalement la correction et la complétude de notre transformation, ainsi que la pertinence de l'inclusion de la temporalité. La section 8.3 explore comment notre proposition peut aider les AAF à générer des explications. Les travaux présentés dans ce chapitre font l'objet d'une publication en 2023 dans le cadre des « Journées d'Intelligence Artificielle Fondamentale » (JIAF-JFPDA 2023) [MUNRO et collab., 2023b] et d'une publication dans le « Workshop on Explainable Logic-Based Knowledge Representations » (XLoKR 2023) [MUNRO et collab., 2023a]. Ils sont poursuivis principalement par Yann Munro dans le cadre de sa thèse. Une extension de ces travaux a été soumise en 2024 à la « International Joint Conference on Artificial Intelligence » (IJCAI 2024).

8.1 Système d'argumentation abstrait

Dans cette section nous rappelons les principes de base des AAF [DUNG, 1995]. Un *système abstrait d'argumentation* est un couple (A,R) où A est un ensemble fini d'arguments et R est une relation binaire sur $\mathrm{A}\times\mathrm{A}$. Nous appelons R la relation d'attaque et nous considérons qu'un argument $a\in\mathrm{A}$ attaque $b\in\mathrm{A}$ si $(a,b)\in\mathrm{R}$, ce qui s'écrit $\mathrm{R}_a b$. Nous pouvons naturellement représenter un système abstrait d'argumentation sous la forme d'un graphe.

Exemple 8.1 [le médecin et le radiologue]**.** *Pour illustrer ces notions, nous introduisons ici un scénario argumentatif modélisant l'interaction entre un médecin demandeur, D, et un radiologue, R, à propos d'un examen d'un bébé de n mois pour la pathologie X.*
D : Peux-tu me faire un scanner pour ce bébé ? (a)
R : Il vaut mieux éviter les radiations ionisantes pour les jeunes bébés. (b)
R : Je peux te proposer une IRM dans deux jours. (c)
D : On peut voir X sur une IRM ? (d)
R : Oui bien sûr ! Si tu veux une confirmation, regarde le guide des bonnes pratiques en radiologie. (e)
D : Mais puisqu'il s'agit d'un bébé, il risque de bouger et donc on pourrait manquer l'information que l'on cherche car l'image ne sera pas très nette. (f)
R : Ne t'inquiète pas, j'ai l'habitude de faire ce genre d'examen pour des bébés. (g)
D : Est-ce que cela ne coûte pas beaucoup plus cher à l'hôpital de faire une IRM ? (h) Il faut aussi que je voie avec la famille du patient car ça pourrait leur revenir plus cher (i).
R : Aucun problème dans ces cas là. Ce coût élevé englobe l'expérience acquise par mon équipe, de sorte qu'à l'avenir, elle puisse réaliser ce type d'examen délicat sans moi. (j)
D : Je viens de discuter avec la famille, aucun problème avec l'IRM elle est couverte pour ça. (k)

D : Cependant, la famille n'est pas rassurée de devoir attendre deux jours, peux-tu faire l'IRM dans la journée ? (l)
R : Non je n'ai vraiment plus de place. Mon prochain créneau est dans deux jours comme je te l'ai dit. (m)

À la suite de cet échange, la décision est donc arrêtée sur une IRM programmée dans deux jours. Mais plus tard dans la journée, le médecin reçoit un appel de la famille pour prévenir que le bébé ne va vraiment pas bien et insister sur l'urgence de l'examen. Le médecin recontacte donc le radiologue pour ajouter un dernier argument :
D : C'est vraiment urgent pour le bébé, il faut une place aujourd'hui ! (n)

*A partir de ce dialogue, on peut extraire manuellement des arguments et les relations entre eux afin d'obtenir un AAF représenté en figure 8.1 avec les arguments suivants : {**a** : Scanner, b : Radiations ionisantes, **c** : IRM dans deux jours, d : X non visible par IRM, e : X visible par IRM, f : Conditions difficiles, g : Expérience, h : Coût élevé pour l'hôpital, i : Coût élevé pour le patient, j : Pas problématique pour l'hôpital, k : Famille couverte pour une IRM, **l** : IRM aujourd'hui, m : Pas de disponibilité aujourd'hui, n : C'est une urgence !}. Les arguments a, c, l sont appelés les variables de décision, leur acceptation étant le critère déclencheur d'une décision : scanner, IRM dans deux jours, ou IRM aujourd'hui.*

*Le système d'argumentation obtenu est un graphe que l'on peut associer à ce dialogue. Ce processus d'extraction peut également être effectué automatiquement, en utilisant des méthodes dites d'*argument mining *[*LIPPI *et* TORRONI, *2016]*.

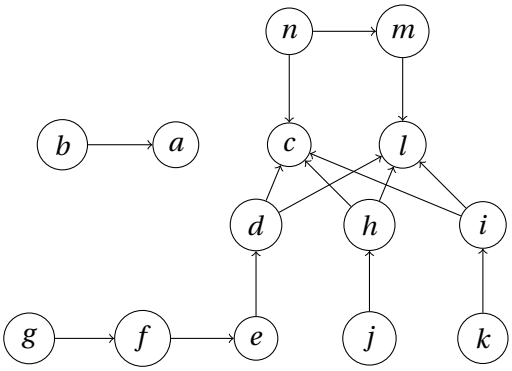


FIGURE 8.1 – Graphe d'argumentation associé à l'exemple 8.1.

Le graphe d'argumentation obtenu avec les AAF qui est représenté sur la figure 8.1 est une représentation statique du dialogue dont toute notion de temporalité a été effacée. Ainsi, si les arguments avaient été énoncés dans un ordre différent, cela ne changerait pas pour autant le graphe. Pourtant, comme nous allons le voir, l'ordre d'énonciation a de l'importance lorsqu'il est question de causalité.

Une fois le graphe d'argumentation construit, il est possible de raisonner sur ce graphe afin de déterminer les ensembles d'arguments qui peuvent être acceptés. Pour cela, nous rappelons quelques définitions supplémentaires :

— Nous notons Att_a l'ensemble des attaquants directs de a pour la relation R :

$$Att_a = \{b \in A | R(b, a)\};$$

— Un ensemble S est *sans conflit* s'il n'y pas d'arguments $(a, b) \in S^2$ qui s'attaquent l'un l'autre ;

— Un argument $a \in A$ est *acceptable* par un ensemble S si S attaque tous les attaquants de a ;

— Un ensemble S sans conflit est dit *admissible* si tous ses éléments sont acceptables par S.

Nous pouvons également définir des sémantiques à base d'extension. Ce sont des propriétés qui doivent être respectées par un ensemble d'arguments afin qu'il soit accepté. Dans le cas des graphes acycliques, toutes ces sémantiques coïncident et ne forment qu'une unique extension non vide, admissible, et ne seront donc pas évoquées ici [SIMARI et RAHWAN, 2009]. La gestion des graphes cycliques est l'objet de l'extension de ce travail soumise à IJCAI 2024.

Exemple 8.1 [suite]**.** *La modélisation du dialogue entre le radiologue et le médecin correspond à un graphe acyclique. Pour déterminer l'ensemble des arguments acceptables, il suffit de partir des arguments non attaqués, ici* $\{b, g, j, k, n\}$*. Ces derniers sont par défaut acceptés. Ensuite, un argument attaqué par au moins un argument accepté ne peut être accepté. En appliquant ce principe, nous obtenons que l'argument l est accepté alors que a et c ne le sont pas. La décision finale est donc de réaliser une IRM en urgence dans la journée.*

8.2 Passage des AAF à $\mathscr{S}_c$

Dans cette section, nous présentons la contribution principale de ce chapitre, à savoir une réécriture d'un graphe d'argumentation abstrait acyclique dans $\mathscr{S}_c$. Pour cela, la section 8.2.1 présente la définition du contexte argumentatif κ_c, la section 8.2.2 fournit les définitions modifiées de la sémantique associée à $\mathscr{S}_c$ et la section 8.2.3 décrit brièvement l'implémentation en ASP. Enfin, la section 8.2.4 présente les propriétés de la transformation proposée.

Contrairement aux AAF, nous proposons de prendre en compte l'ordre d'énonciation des arguments. Au lieu d'avoir seulement un couple (A, R), l'entrée est un couple (Δ, R), où Δ est un dialogue, i.e. une séquence d'énoncés en langage naturel :

Définition 8.1 [*Dialogue* Δ]**.** *Un dialogue* Δ *est défini comme* $\Delta = \{(a, o) | (a, o) \in A \times \mathbb{N}\}$*, où chaque argument a est associé à son ordre d'énonciation o.*

8.2.1 Spécification du contexte κ_c

Pour pouvoir passer d'un graphe d'argumentation à $\mathscr{S}_c$, il faut d'abord définir les fluents $\mathbb{F}$ i.e. les variables nécessaires pour décrire le monde, ici le graphe d'argumentation. Deux éléments doivent être pris en compte : les arguments et les relations d'attaque. Pour décrire un argument x, nous introduisons deux fluents : $p_x \in \mathbb{F}$ qui indique si l'argument est présent ou non dans le graphe et $a_x \in \mathbb{F}$ qui indique l'acceptabilité de l'argument. Pour R, nous utilisons le fluent $cA_{y,x} \in \mathbb{F}$ exprimant le fait que y peut attaquer x. Comme nous ne traitons que des AAF acycliques, $\nexists (x_1, \ldots, x_n) \in A$ tel que $\left(cA_{x_1,x_2}, \ldots, cA_{x_{n-1},x_n}, cA_{x_n,x_1}\right) \in \mathbb{F}$. Nous appelons cette propriété l'acyclicité des fluents cA.

En ce qui concerne les évènements $\mathbb{E}$, dans le cas de l'argumentation abstraite la seule action volontaire possible est d'énoncer un argument x, notée $enunciate_x \in \mathbb{A}$. Pour cela, il faut que l'argument en question n'ait pas déjà été prononcé. Dans ce cas, x devient présent

et acceptable par défaut. Ce choix est justifié par le fait que son acceptabilité sera évaluée à l'état suivant, avant qu'il n'ait un impact sur le reste du graphe. Formellement :

$$pre(enunciate_x) \equiv \neg p_x;$$
$$eff(enunciate_x) \equiv p_x \wedge a_x.$$

Aucun des évènements décrits par la suite n'a pour effet de rendre un argument non présent. Cela implique qu'il n'est pas possible d'énoncer un argument déjà énoncé. Cette hypothèse n'est pas en contradiction avec le cadre de l'argumentation classique. En effet, un argument répété se manifesterait par un argument identique mais de nom différent dans le graphe, ce qui est possible également avec notre transformation. Cependant, $\mathscr{S}_c$ offrant des outils pour tenir compte de la temporalité, une meilleure approche existe. Celle-ci est présentée dans l'extension de ce travail soumise à IJCAI 2024. Malgré tout, ce travail étant une première étape, il vise à poser des bases solides au prix de quelques hypothèses simplificatrices.

Contrairement au cadre de l'argumentation abstraite, nous prenons ici en compte l'ordre d'énonciation des arguments. Cela implique de mettre à jour l'acceptabilité de tous les autres arguments présents après l'énonciation d'un nouvel argument et avant l'énonciation du suivant. Cela définit des états que nous appelons états argumentatifs.

Définition 8.2 [*État argumentatif*]**.** *Un état* $\mathrm{S}(t)$ *est dit argumentatif si :*

1. $\forall x, y, \left[\mathrm{S}(t) \vDash a_x \wedge p_y \wedge c\mathrm{A}_{y,x} \Rightarrow \mathrm{S}(t) \vDash \neg a_y\right]$;
2. $\forall x, \left[\mathrm{S}(t) \vDash p_x \wedge \left(\bigwedge_y \neg a_y \vee \neg c\mathrm{A}_{y,x}\right) \Rightarrow \mathrm{S}(t) \vDash a_x\right]$.

Après l'énonciation d'un argument, nous souhaitons que des mises à jour soient déclenchées automatiquement. Nous les représentons par deux évènements naturels :

$$makes\mathrm{U}nacc_{y,x} \in \mathbb{N} \quad \text{et} \quad makes\mathrm{A}cc_x \in \mathbb{N}.$$

Pour rappel, un argument n'est acceptable que s'il est non attaqué ou attaqué uniquement par des arguments non acceptables. De fait, il suffit que l'un des attaquants soit acceptable pour rendre l'argument attaqué non acceptable. Cela implique donc au moins deux cas à envisager :

Mise à jour de l'acceptabilité : Supposons que l'argument y venant d'être énoncé peut attaquer l'argument x et que x et y sont acceptables. Alors, x étant attaqué par un argument acceptable y, x devient non acceptable. Formellement, l'évènement naturel $makes\mathrm{U}nacc_{y,x}$ peut s'écrire :

$$tri(makes\mathrm{U}nacc_{y,x}) \equiv a_x \wedge a_y \wedge c\mathrm{A}_{y,x};$$
$$eff(makes\mathrm{U}nacc_{y,x}) \equiv \neg a_x.$$

Cette écriture permet également de traiter les cas où un nouvel argument z rend un attaquant y de x à nouveau acceptable. Dans ce cas, x devient non acceptable.

Mise à jour de la non-acceptabilité : Supposons que l'argument x est non acceptable et qu'un argument z vient d'être prononcé. Celui-ci n'a pas de lien direct avec l'argument x mais a pu impacter l'acceptabilité de certains attaquants de x. Nous vérifions donc si tous

les arguments pouvant attaquer x sont acceptables ou non. Si aucun d'entre eux n'est effectivement acceptable, alors x le redevient. Dans $\mathscr{S}_c$, cela se traduit par l'évènement naturel $makesAcc_x$ tel que :

$$tri(makesAcc_x) \equiv p_x \wedge \neg a_x \wedge \left(\bigwedge_y \neg cA_{y,x} \vee \neg a_y \right);$$
$$eff(makesAcc_x) \equiv a_x.$$

Enfin, lorsqu'un argument x est énoncé, il faut vérifier qu'il n'est pas rendu non acceptable par un argument y déjà présent avant qu'il ne rende non acceptables d'autres arguments, z par exemple. Cela se traduit par la règle de priorité ci-dessous :

$$makesUnacc_{y,x} \succ_{\mathbb{E}} makesUnacc_{x,z}.$$

Notons qu'ajouter un argument dans le graphe ne peut impacter les autres arguments de manière directe qu'en les rendant non acceptables. Pour cette raison, il n'est pas nécessaire d'établir une règle de priorité de la forme $makesUnacc_{y,x} \succ_{\mathbb{E}} makesAcc_z$ car cette situation est déjà couverte par la règle précédente.

Avant de conclure cette section, il est important de noter que la distinction entre actions et évènements naturels s'avère indispensable pour cette traduction entre les AAF et $\mathscr{S}_c$. Il s'agit d'un exemple montrant que les choix faits dans l'optique d'une utilisation dans l'éthique computationnelle ne limitent pas notre approche de causalité à ce seul domaine. Notre approche peut être utilisée pour d'autres applications.

8.2.2 Modification de la sémantique

Après avoir défini le contexte κ_c pour le cadre argumentatif, il faut modifier les définitions associées à la sémantique de $\mathscr{S}_c$. Cela va permettre en particulier d'obtenir des traces représentatives de la réalité. Pour cela, les arguments sont énoncés à partir d'états argumentatifs étape par étape dans l'ordre de l'interaction. Nous montrons dans la section 8.2.4 que, comme nous ne considérons que des graphes acycliques, il existe toujours un tel état après l'ajout d'un nouvel argument et il est donc toujours possible de continuer l'interaction.

La définition actuelle du scénario σ n'est pas adaptée à ce cas. En effet, elle demande la connaissance préalable du nombre d'étapes nécessaires pour revenir à un état argumentatif, de sorte à prévoir l'état exact dans lequel l'argument suivant pourra être énoncé. Nous proposons pour résoudre ce problème d'introduire un ensemble d'actions ordonnées appelé *séquence*, que nous notons $\varsigma \subseteq \mathbb{A} \times \mathbb{N}$. L'unicité de l'exécution valide n'est plus obtenue grâce au scénario σ, mais à la séquence ς. Il faut donc modifier les définitions 6.2 et 6.8.

Définition 8.3 [Cadre argumentatif χ]**.** *La cadre argumentatif noté χ est le couple (κ_c, ς) avec ς une séquence et κ_c un contexte.*

La définition 8.4 modifie la définition 6.2 : les conditions 2d et 2e sont ajoutées et dans la condition 2c, $\forall e \in \mathbb{E}$ est remplacé par $\forall e \in \mathbb{N}$. Ces modifications expriment le fait qu'une action de la séquence ς ne peut être déclenchée que si aucun évènement naturel ne se déclenche au même pas de temps. Les autres conditions restent identiques, nous ne modifions donc rien vis-à-vis du déclenchement des évènements naturels.

Définition 8.4 [Système de transition d'états étiqueté $\mathscr{S}_{arg}$]**.** *Le système de transition d'états étiqueté $\mathscr{S}_{arg}$ est un triplet $\langle 2^{Lit_{\mathbb{F}}}, 2^{\mathbb{E}}, \tau\rangle$ où τ est l'ensemble des relations étiquetées de transition entre états notées* (S, E, S')*. Les triplets de cet ensemble vérifient :*

1. $S \subseteq Lit_{\mathbb{F}}$ *est un état au sens de la définition 6.1 ;*
2. $E \subseteq \mathbb{E}$ *vérifie :*
   (a) $\forall e \in E, S \vDash pre(e)$ *;*
   (b) $\nexists (e, e') \in E^2,\ e \succ_{\mathbb{E}} e'$ *;*
   (c) $\forall e \in \mathbb{N}$ *tel que* $S \vDash tri(e)$, $e \in E$ *ou* $\exists e' \in E,\ e' \succ_{\mathbb{E}} e$ *;*
   (d) *Si* $\exists e \in E(t) \cap \mathbb{A}$, *alors* $\forall e' \in \mathbb{N}, S(t) \nvDash tri(e')$ *;*
   (e) $E(t) \neq \varnothing$ *;*
3. $S' = \left\{ l \in S | \forall e \in E, \bar{l} \notin eff(e) \right\} \cup \left\{ l \in Lit_{\mathbb{F}} | \exists e \in E, l \in eff(e) \right\}$.

Comme dans la définition 6.8, la définition 8.5 donne une définition des traces mais pour un cadre argumentatif $\chi = (\kappa_c, \varsigma)$.

Définition 8.5 [traces d'évènements et d'états τ^e_χ et τ^s_χ]**.** *Étant donné un cadre argumentatif χ, la trace d'évènements τ^e_χ et la trace d'états τ^s_χ sont respectivement les séquences d'évènements $E^\chi(-1), E^\chi(0), \ldots, E^\chi(N)$ et d'états $S^\chi(-1), S^\chi(0), S^\chi(1), \ldots, S^\chi(N+1)$ qui vérifient :*

1. $S^\chi(0) = S_0$ *;*
2. $\forall t \in \mathbb{T},\ (S^\chi(t), E^\chi(t), S^\chi(t+1)) \in \tau$ *;*
3. $\forall t \in \mathbb{T}, E(t) \subset (\{a | \exists o \in \mathbb{N}, (a, o) \in \varsigma\} \cup \mathbb{N})$ *;*
4. $\forall \left((e, o), (e', o')\right) \in \varsigma^2$ *tel que* $o < o'$, $\exists t, t'$ *tel que* $e \in E(t)$ *et* $e' \in E(t')$ *et* $t < t'$ *;*
5. $\forall \left((e, o), (e', o')\right) \in \varsigma^2$ *tel que* $o = o'$, $\exists t$ *tel que* $(e, e') \in E(t)^2$.

8.2.3 Implémentation en ASP

Nous proposons une implémentation en ASP sur la base de celle décrite dans la section 6.3. Les programmes ASP, $\pi_{con}(\kappa_c)$ et $\pi_{seq}(\varsigma)$, sont obtenus par la traduction respectivement du contexte κ_c et de la séquence ς. $\pi_{\mathbb{A}}$ est obtenu en modifiant le programme présenté dans la section 6.3.2 de sorte à introduire les éléments propres à $\mathscr{S}_{arg}$. $\pi_{\mathbb{C}}$ est le même que dans la section 6.3.2. Le programme complet $\Pi(\chi) = \pi_{sce}(\varsigma) \cup \pi_{con}(\kappa_c) \cup \pi_{\mathbb{A}} \cup \pi_{\mathbb{C}}$ est disponible [1].

8.2.4 Quelques propriétés formelles

Cette section donne les propriétés formelles de la transformation proposée. Tout d'abord, nous établissons que la notion de temporalité est bien prise en compte par la transformation. Ensuite, nous établissons sa correction et sa complétude. Enfin, nous introduisons une proposition qui ouvre la voie à la discussion de la section 8.3.

1. `https://gitlab.lip6.fr/sarmiento/kr_2023.git`

8.2.4.1 Propriétés préliminaires sur les traces

Le premier résultat montre que, étant donné un χ, les traces τ^e_χ et τ^s_χ sont uniques.

Proposition 8.1 [Unicité des traces]**.** *Soit une cadre argumentatif* $\chi = (\kappa_c, \varsigma)$*, les traces* τ^e_χ *et* τ^s_χ *sont uniques.*

Démonstration. Nous prouvons par l'absurde l'unicité des traces pour un χ donné. Soit $\chi = (\kappa_c, \varsigma)$ le cadre argumentatif et τ^e_χ, $\tau^{e'}_\chi$ deux traces d'évènements. Par reductio ad absurdum nous supposons que $\tau^e_\chi \neq \tau^{e'}_\chi$.

Nous notons $E^\chi(t)$ et $S^\chi(t)$ les éléments associés à τ^e_χ et τ^s_χ, puis $E^{\chi'}(t)$ et $S^{\chi'}(t)$ les éléments associés à $\tau^{e'}_\chi$ et $\tau^{s'}_\chi$. D'après la définition 8.4, $S^\chi(t+1)$ est déduit de $S^\chi(t)$ et les évènements dans $E^\chi(t)$. De la même façon, $S^{\chi'}(t+1)$ est déduit de $S^{\chi'}(t)$ et les évènements dans $E^{\chi'}(t)$. Puis, étant donné que le contexte κ_c est commun à τ^e_χ et $\tau^{e'}_\chi$, alors $S^\chi(0) = S^{\chi'}(0)$ et $E^\chi(-1) = E^{\chi'}(-1)$. Par conséquent, la première différence entre τ^e_χ et $\tau^{e'}_\chi$ ne pourrait pas se trouver entre deux états, mais uniquement entre deux ensembles d'évènements.

Notons t_0 le premier point temporel où une différence serait observée entre τ^e_χ et $\tau^{e'}_\chi$. Nous aurions $\forall t < t_0$, $E^\chi(t) = E^{\chi'}(t)$ et $\forall t \leq t_0$, $S^\chi(t) = S^{\chi'}(t)$, i.e. une équivalence sur les deux traces avant ce premier point, et nous aurions à t_0, $E^\chi(t_0) \neq E^{\chi'}(t_0)$. Nous pouvons alors déduire que $\forall e \in \mathbb{E}$, $S^\chi(t_0) \models pre(e) \Leftrightarrow S^{\chi'}(t_0) \models pre(e)$.
Notons $D = E^\chi(t_0) \setminus E^{\chi'}(t_0) \cup E^{\chi'}(t_0) \setminus E^\chi(t_0)$, l'ensemble avec toutes les différences entre $E^\chi(t_0)$ et $E^{\chi'}(t_0)$. Puis, considérons un évènement $e_0 \in max_{\succ_\mathbb{E}}\{D\}$, i.e. l'événement dans l'ensemble D avec la priorité de déclenchement maximale. Sans perte de généralité par la symétrie χ et χ', nous pouvons considérer que $e_0 \in E^{\chi'}(t_0) \setminus E^\chi(t_0)$, soit $e_0 \notin E^\chi(t_0)$ et $e_0 \in E^{\chi'}(t_0)$. Deux cas sont alors possibles, soit $e_0 \in \mathbb{N}$ ou $e_0 \in \mathbb{A}$.

i) Comme pour la proposition 6.1, nous montrons par l'absurde que $e_0 \notin \mathbb{N}$. Par reductio ad absurdum nous supposons que $e_0 \in \mathbb{N}$. Comme $e_0 \in E^{\chi'}(t_0)$ et que $S^\chi(t_0) = S^{\chi'}(t_0)$, alors $S^\chi(t_0) \models tri(e_0)$. Par la condition 2c de la définition 8.4, nous savons que $e_0 \notin E^\chi(t_0)$ implique $\exists e \in E^\chi(t_0)$ tel que $e \succ_\mathbb{E} e_0$. Cela est en contradiction avec l'hypothèse $e_0 \in max_{\succ_\mathbb{E}}\{D\}$. Nous pouvons conclure que $e_0 \in \mathbb{N}$ n'est pas possible et donc $e_0 \notin \mathbb{N}$.

ii) Nous montrons par l'absurde que $e_0 \notin \mathbb{A}$. Par reductio ad absurdum nous supposons que $e_0 \in \mathbb{A}$. Par la condition 3 de la définition 8.5, nous savons que $e_0 \in E^{\chi'}(t_0)$ implique $(e_0, t) \in (\{a | \exists o \in \mathbb{N}, (a, o) \in \varsigma\} \cup \mathbb{N})$. Étant donné que ς est commun à τ^e_χ et $\tau^{e'}_\chi$, l'ordre $o_0 \in \mathbb{N}$ associé à e_0 est le même dans les deux cas. Par conséquent, étant donné que $\forall t < t_0$, $\tau^e_\chi(t) = \tau^{e'}_\chi(t)$ et $e_0 \in E^{\chi'}(t_0)$, alors $e_0 \notin E^{\chi'}(t)$ implique $e_0 \notin E^\chi(t)$. Pour expliquer que e_0 soit dans $E^{\chi'}(t_0)$ mais pas dans $E^\chi(t_0)$, considérons le cas où la procrastination d'une action a eu lieu. Il s'agit alors du cas où $\nexists (e, o_0) \in \varsigma$ tel que $e \in E(t_0)$, et donc $E(t_0) = \varnothing$. Cela est en contradiction avec la condition 2e de la définition 8.4. Dans tous les autres cas, il y a une contradiction avec la condition 5 de la définition 8.5. Nous pouvons conclure que $e_0 \in \mathbb{A}$ n'est pas possible et donc $e_0 \notin \mathbb{A}$.

L'hypothèse qu'il existerait un premier point temporel t_0 où une différence serait observée entre τ^e_χ et $\tau^{e'}_\chi$ mène à une contradiction pour les deux cas possibles, (i) et (ii). Il ne nous reste d'autre alternative que de rejeter cette hypothèse. Puisqu'il n'existe pas de premier point temporel t_0 où une différence serait observée entre τ^e_χ et $\tau^{e'}_\chi$, nous devons également rejeter $\tau^e_\chi \neq \tau^{e'}_\chi$. L'unicité des traces étant donné χ est alors prouvée. □

Dorénavant, lorsqu'il sera question d'évènements et d'états, il s'agira de ceux étant donné les traces uniques τ^e_χ et τ^s_χ. Ainsi, l'ensemble des évènements qui se sont effectivement produits à l'instant t est $\mathrm{E}^\chi(t) = \tau^e_\chi(t)$. De même, l'état réel à l'instant t est $\mathrm{S}^\chi(t) = \tau^s_\chi(t)$.

8.2.4.2 Complétude et correction

Nous établissons à présent l'aspect complet et correct de notre transformation. Pour cela, nous introduisons d'abord la notion de graphe associé.

Définition 8.6 [Graphe associé à un état]**.** *Soit* $\mathrm{S}^\chi(t)$ *un état. Son graphe associé est le graphe* $\mathrm{AF}' = (\mathrm{A}', \mathrm{R}')$*, tel que :*

$$\mathrm{A}' = \left\{x \in \mathrm{A} | \mathrm{S}^\chi(t) \vDash p_x\right\} \quad \textit{et} \quad \mathrm{R}' = \left\{(y,x) \in \mathrm{A}^2 | \mathrm{S}^\chi(t) \vDash c\mathrm{A}_{y,x}\right\}.$$

D'après la propriété d'acyclicité du fluent cA, le graphe associé est acyclique.

Nous nous concentrons maintenant sur la notion d'acceptabilité. Nous commençons par caractériser les états argumentatifs à l'aide de tri.

Lemme 8.1. *Soit* $\mathrm{S}^\chi(t)$ *un état. Les deux propositions suivantes sont équivalentes :*

— $\forall e \in \mathbb{N}, \mathrm{S}^\chi(t) \nvDash tri(e)$ *;*

— $\mathrm{S}^\chi(t)$ *est un état argumentatif au sens de la définition 8.2.*

Démonstration. Dans notre contexte nous avons $\mathbb{N} = \{makes\mathrm{A}cc, makes\mathrm{U}nacc\}$. Nous prouvons que la condition $\mathrm{S}^\chi(t) \nvDash tri(makes\mathrm{A}cc_x)$ est équivalente à la condition 2 de la définition 8.2 et que la condition $\mathrm{S}^\chi(t) \nvDash tri(makes\mathrm{U}nacc_{y,x})$ est équivalente à la condition 1 de la définition. Pour tout x, y :

— $\begin{aligned}\neg tri(makes\mathrm{A}cc_x) &= \neg(p_x \wedge \neg a_x \wedge (\bigwedge_y \neg a_y \vee \neg c\mathrm{A}_{y,x})) \\ &= \neg(p_x \wedge (\bigwedge_y \neg a_y \vee \neg c\mathrm{A}_{y,x})) \vee a_x \\ &= p_x \wedge (\bigwedge_y \neg a_y \vee \neg c\mathrm{A}_{y,x}) \Rightarrow a_x ;\end{aligned}$

— $\begin{aligned}\neg tri(makes\mathrm{U}nacc_{y,x}) &= \neg(a_x \wedge a_y \wedge c\mathrm{A}_{y,x}) \\ &= \neg(a_x \wedge p_y \wedge a_y \wedge c\mathrm{A}_{y,x}) \\ &= \neg(a_x \wedge p_y \wedge c\mathrm{A}_{y,x}) \vee \neg a_y \\ &= a_x \wedge p_y \wedge c\mathrm{A}_{y,x} \Rightarrow \neg a_y.\end{aligned}$

□

Un état argumentatif est considéré comme un état où rien ne se passe tant qu'une action volontaire n'est pas effectuée. Nous montrons maintenant qu'il est toujours possible d'atteindre un tel état à partir d'un état argumentatif dans lequel un argument $x \in \mathrm{A}$ est énoncé.

Proposition 8.2 [Existence d'états argumentatifs]**.** *Soit* $\mathrm{S}^\chi(t)$ *un état argumentatif et* $x \in \mathrm{A}$ *un argument. Si* $enunciate_x \in \mathrm{E}^\chi(t)$*, alors* $\exists t' \in \mathbb{T}$*,* $t < t'$ *tel que* $\mathrm{S}^\chi(t')$ *est un état argumentatif.*

Démonstration. Étant donné un état argumentatif $\mathrm{S}^\chi(t)$ et $x \in \mathrm{A}$ tel que $enunciate_x \in \mathrm{E}(t)$, prouvons que $\exists t' \in \mathbb{T}$, $t < t'$ tel qu'aucun déclenchement n'a lieu dans l'état $\mathrm{S}^\chi(t')$ et donc que celui-ci est un état argumentatif d'après le lemme 8.1.

Les ensembles $\mathbb{N}$ et $\left\{\mathrm{S}^\chi(t) \vDash c\mathrm{A}_{y,x} \mid (x,y) \in \mathrm{A}^2\right\}$ étant finis, il existe un nombre fini de déclenchements possibles dans $\mathrm{E}^\chi(t)$. De plus, comme le nombre d'arguments est fini, il y a un nombre fini de chemins dans le graphe associé. Les graphe que nous étudions étant acycliques, chaque chemin a une longueur finie. Par conséquent, il y a un nombre maximum d'ensembles d'évènements (M) et donc $\exists t' \leq (t + \mathrm{M} + 1)$ tel que $\forall e \in \mathbb{N}, \mathrm{S}^\chi(t') \nvDash tri(e)$. □

Enfin, la proposition suivante permet de prouver qu'un argument acceptable dans l'état argumentatif est acceptable dans le graphe associé et vice-versa. Nous commençons par prouver un lemme.

Lemme 8.2. *Étant donné un état argumentatif* $S^\chi(t)$ *:*

$$\forall x, \left(S^\chi(t) \vDash p_x \wedge \neg a_x\right) \Leftrightarrow \left(S^\chi(t) \vDash \exists y, p_x \wedge a_y \wedge cA_{y,x}\right).$$

Démonstration. [$\Longrightarrow$] : Pour tout x tel que $S^\chi(t) \vDash p_x \wedge \neg a_x$, la condition 2 de la définition 8.2 implique que $S^\chi(t) \vDash \neg p_x \vee (\bigvee_y a_y \wedge cA_{y,x})$. Par conséquent, comme $S^\chi(t) \vDash p_x$, alors $S^\chi(t) \vDash \exists y, p_x \wedge a_y \wedge cA_{y,x}$.

[$\Longleftarrow$] : Prouvons ce cas par l'absurde. Soit x_0 tel que $S^\chi(t) \vDash \neg p_{x_0} \vee a_{x_0}$. Si $S^\chi(t) \vDash \neg p_{x_0}$, alors x_0 est tel que $S^\chi(t) \vDash \forall y, \neg p_{x_0} \vee \neg a_y \vee \neg cA_{y,x_0}$, qui est ce que nous voulons. Sinon, $S^\chi(t) \vDash p_{x_0} \wedge a_{x_0}$. Si $S^\chi(t) \vDash p_{x_0} \wedge (\exists y, a_y \wedge cA_{y,x_0})$ alors $S^\chi(t) \vDash tri(makesUnacc_{y,x_0})$, ce qui n'est pas possible étant donné que $S^\chi(t)$ est un état argumentatif. Les deux cas mènent donc à une contradiction. □

La proposition suivante établit la correspondance entre l'acceptabilité en argumentation et les états argumentatifs.

Proposition 8.3 [Correction et complétude locale]**.** *Soit* $S^\chi(t)$ *un état argumentatif et* $AF = (A, R)$ *son graphe associé. Alors, pour tout* x*,* $x \in A$ *est acceptable par* A *si et seulement si* $S^\chi(t) \vDash a_x$*.*

Démonstration. [$\Longrightarrow$] : Soit $x_0 \in A$ tel que x_0 est acceptable par A. Prouvons que $S^\chi(t) \vDash a_{x_0}$.

Supposons que $S^\chi(t) \vDash \neg a_{x_0}$. Par la façon de construire AF, nous avons $S^\chi(t) \vDash p_{x_0}$. D'après le lemme 8.2, $S^\chi(t)$ est un état argumentatif. De ce fait, nous savons que :

$$S^\chi(t) \vDash p_{x_0} \wedge \neg a_{x_0} \Leftrightarrow S^\chi(t) \vDash \exists y, p_{x_0} \wedge a_y \wedge cA_{y,x_0}.$$

La condition 1 de la définition 8.2 appliquée à a_y nous indique que :

$$\forall z, S^\chi(t) \vDash a_y \wedge p_z \wedge cA_{z,y} \Rightarrow S^\chi(t) \vDash \neg a_z.$$

Le nombre d'arguments étant fini, il est possible de répéter le processus appliqué pour x_0 à z jusqu'à arriver dans un des deux scénarios suivants :

— $S^\chi(t) \vDash \nexists y, p_z \wedge a_y \wedge cA_{y,z}$. Cela mène au déclenchement de l'évènement $makesAcc_z$, ce qui est en contradiction avec le fait que $S^\chi(t)$ est un état argumentatif d'après le lemme 8.1 ;

— $\forall z, S^\chi(t) \vDash a_y \wedge p_z \wedge cA_{z,y} \Rightarrow S^\chi(t) \vDash \neg a_z$ où $p_z \wedge cA_{z,y}$ est faux. Dans ce cas, dans AF, $Att_y = \emptyset$. Par conséquent, y est acceptable, ce qui est en contradiction avec le fait que x_0 est acceptable.

Nous pouvons alors déduire que $S^\chi(t) \vDash a_{x_0}$.

[$\Longleftarrow$] : Soit $x_0 \in A$ tel que $S^\chi(t) \vDash a_{x_0}$. Prouvons que x_0 est acceptable par A.

Comme $S^\chi(t) \vDash a_{x_0}$ et que $S^\chi(t)$ est un état argumentatif, nous avons :

$$\forall y, S^\chi(t) \vDash a_{x_0} \wedge p_y \wedge cA_{y,x_0} \Rightarrow S^\chi(t) \vDash \neg a_y.$$

Puis, d'après la définition des AF, pour tout y qui satisfait notre prémisse, $(x_0, y) \in A^2$ et $(y, x_0) \in R$.

Si l'argument y est acceptable par A, alors d'après la partie [$\Longrightarrow$] de la preuve, $S^{\chi}(t) \vDash a_y$. Dans ce cas, $S^{\chi}(t) \vDash tri(makesUnacc_{y,x_0})$, ce qui est en contradiction avec le fait que $S^{\chi}(t)$ est un état argumentatif.

Donc, comme $\forall y \in Att_x$, y n'est pas acceptable par A, x_0 est acceptable par A. □

Nous avons établi qu'il existe une équivalence entre un état argumentatif et son graphe associé. Maintenant, à partir d'un dialogue et de la relation d'attaque, les traces sont générées ainsi qu'un AAF. Nous établissons alors l'existence d'un état dont le graphe associé est égal à l'AAF initial. Un tel état est appelé *état argumentatif final* et est défini comme un état argumentatif $S^{\chi}(t)$ tel que $\forall x \in A$, $\exists t' \in \mathbb{T}$ tel que $t' < t$ et $enunciate_x \in E^{\chi}(t')$.

Théorème 8.1 [Correction et complétude]**.** *Soit un dialogue* Δ *et* R *un ensemble de relations d'attaque. Étant donné une cadre argumentatif* χ*, le graphe argumentatif associé* AF' *à un état argumentatif final* $S^{\chi}(t)$*, et* $AF = (A, R)$ *le graphe d'argumentation construit à partir de* (Δ, R)*, nous avons* $AF' = AF$.

Démonstration. Comme AF' est associé à l'état argumentatif final, $\forall x \in A$, $\exists t' \in \mathbb{T}$ tel que $t' < t$ et $enunciate_x \in E^{\chi}(t')$. Nous savons en plus que $eff(enunciate_x) = p_x \wedge a_x$. De ce fait, nous savons que $A' = A$.
Puis, par la façon de construire $cA_{y,x}$ et R', nous avons $R = R'$. Par conséquent, $AF = AF'$.
Pour finir, nous savons que $S^{\chi}(t)$ est un état argumentatif. De ce fait, d'après la proposition 8.3, $\forall x \in A = A', S^{\chi}(t) \vDash a_x \Leftrightarrow x$ est acceptable par A. □

8.2.4.3 Aspects temporels et causaux

Les résultats précédents permettent de montrer la cohérence de l'état final dans $\mathscr{S}_c$ avec l'argumentation. En particulier, le théorème 8.1 est essentiel car il établit la correction et la complétude de notre approche avec l'argumentation, et permet ainsi d'assurer qu'aucune information n'est perdue. À l'inverse, intégrer la temporalité permet d'ajouter des informations supplémentaires grâce aux états intermédiaires et aux relations causales qui peuvent en être déduites, comme illustré dans la section suivante.

Proposition 8.4 [Invariance de l'état final]**.** *Soit* ς *et* ς' *des séquences telles que* ς' *est une permutation des ordres dans* ς*. Étant donné les états argumentatifs finaux* $S^{\kappa_c,\varsigma}(t)$, $S^{\kappa_c,\varsigma'}(t')$*, appartenant à* $\tau^{s}_{\kappa_c,\varsigma}$ *et* $\tau^{s}_{\kappa_c,\varsigma'}$*, avec* $(t,t') \in \mathbb{T}^2$, $S^{\kappa_c,\varsigma}(t) = S^{\kappa_c,\varsigma'}(t')$.

Démonstration. Considérons respectivement AF et AF' les deux graphes associés aux états argumentatifs finaux $S^{\kappa_c,\varsigma}(t)$ et $S^{\varsigma',\kappa_c}(t')$. Étant donné qu'ils ont tous les deux les même actions dans leur respective séquence, alors $A = A'$. Du fait qu'ils partagent également le même contexte nous pouvons déduire que $R = R'$. Par conséquent $AF = AF'$.

D'après le proposition 8.3, nous savons que $x \in A$ est acceptable par $A \Leftrightarrow S^{\kappa_c,\varsigma}(t) \vDash a_x$. De ce fait, $\forall x, S^{\kappa_c,\varsigma}(t) \vDash a_x \Leftrightarrow \forall x, S^{\varsigma',\kappa_c}(t') \vDash a_x$. □

Cette proposition implique que l'état argumentatif final ne dépend pas de ς, mais uniquement de l'ensemble d'arguments qu'il contient : peu importe l'ordre dans lequel les arguments sont énoncés, l'état argumentatif final est toujours le même. Cela implique le corollaire suivant sur l'unicité :

Corollaire 8.1. *Étant donné un* $AF = (A, R)$, $\exists! S^{\chi}(t)$ *qui est un état argumentatif final dont le graphe argumentatif associé est* $AF = (A, R)$.

La proposition 8.4 et son corollaire montrent la concordance avec le AAF. La pertinence de la prise en compte de la temporalité se fait au niveau des états argumentatifs intermédiaires et des relations causales qui peuvent en être déduites, comme nous le verrons en détail dans la section 8.3.

Proposition 8.5 [Variance de la causalité]**.** *Les relations causales sont dépendantes de la séquence* ς.

Démonstration. Cette proposition est prouvée grâce à l'exemple 8.1 et une variante de celui-ci. Soit $\Pi(\chi)$ le programme obtenu pour le cadre argumentatif κ_c, ς correspondant à l'exemple 8.1 et $\Pi(\chi')$ celui obtenu pour un cadre argumentatif κ_c, ς' où nous modifions un peu la séquence de l'exemple 8.1. Le dialogue débute de la même façon avec l'énonciation des arguments a, b, c. À ce moment là, le médecin demande directement s'il n'est pas possible de faire l'IRM aujourd'hui même (l). Le radiologue répond qu'il ne peut que dans deux jours au plus tôt (m). Le médecin précise alors qu'il s'agit d'une urgence (n). Ensuite, le reste du dialogue se déroule dans le même ordre que précédemment. D'après la définition 6.11 et le théorème 6.3, nous savons que $\Pi(\chi) \vDash ness(o(enunciate_d, 4), h(neg(a_c), 31))$, alors que $\nexists t, t' \in \mathbb{T}^2, \Pi(\chi') \vDash ness(o(enunciate_d, t), h(neg(a_c), t'))$. □

Cette proposition montre que les relations causales sont dépendantes de l'ordre d'énonciation des arguments. Ainsi, même si, comme dans le cadre classique de l'argumentation, l'acceptabilité d'un argument dans l'état argumentatif final n'en dépend pas, voir théorème 8.1, il est tout de même essentiel de tenir compte de la temporalité lorsqu'il est question de causalité, notamment pour l'explicabilité.

8.3 Vers des explications : un processus à trois niveaux

En formalisant les graphes d'argumentation acycliques abstraits dans $\mathscr{S}_c$, nous avons maintenant une connaissance complète de l'évolution du dialogue. En effet, nous sommes en mesure de générer deux traces : l'une retraçant l'évolution des états du monde, l'autre retraçant l'occurrence des évènements. Cette section propose d'explorer à trois niveaux comment l'utilisation des langages d'action peut aider les AAF à générer des explications. L'exemple 8.1 est utilisé tout au long de cette section pour illustrer la discussion. La trace d'états correspondant à l'exemple compte trente et un états et la trace d'évènements compte quatorze actions et vingt évènements naturels.

8.3.1 Modélisation temporelle et représentation graphique

Les systèmes de transition offrent une manière de modéliser le monde qui se prête à une représentation visuelle. En effet, la représentation de deux états successifs et des évènements qui ont conduit à cette évolution donne une représentation précise de la manière dont l'évolution du monde est modélisée. Pour cela, dans $\mathscr{S}_c$ nous utilisons les traces. En effet, celles-ci enregistrent l'évolution du monde. Par conséquent, les afficher dans l'ordre chronologique fournit une narration de l'interaction.

En accord avec cette idée d'utiliser la narration de l'interaction fournie par les traces, nous proposons dans la figure 8.2 un affichage possible des traces d'évènements et d'états obtenues à partir de l'exemple 8.1 en montrant les fluents sous forme d'hexagones et les

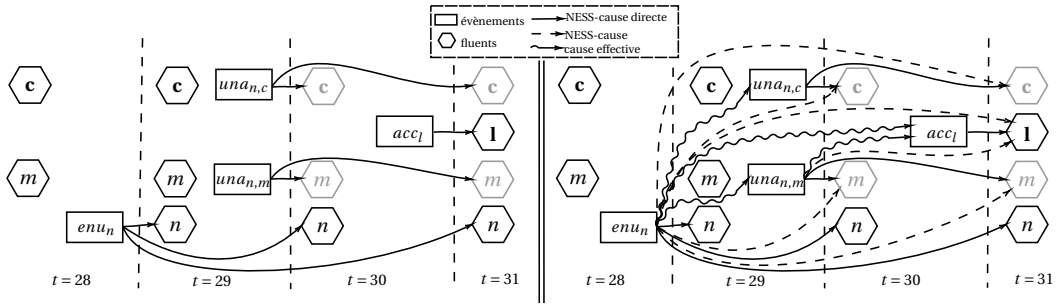


FIGURE 8.2 – Représentation graphique partielle (gauche) et enrichie (droite) des traces et des relations causales correspondant à l'exemple 8.1.

évènements déclenchés sous forme de rectangles. Puisque l'acceptabilité des arguments est ce qui importe le plus dans un dialogue, nous ne représentons que les fluents a_x, en utilisant uniquement les noms des arguments pour des raisons de lisibilité. De plus, nous ne faisons pas apparaître les fluents lorsqu'ils ne sont pas présents dans l'état, sauf lorsque c'est l'occurrence d'un évènement représenté qui entraîne la négation du fluent. Dans ce cas, la négation est représentée par une teinte plus claire. Les évènements $enunciate_x$, $makes\text{U}nacc_{y,x}$, et $makes\text{A}cc_x$ sont abrégés en enu_x, $una_{y,x}$, et acc_x, respectivement.

La représentation graphique partielle, à gauche de la figure 8.2, montre les relations causales qui peuvent être déduites immédiatement, ce sont les NESS-causes directes. Elles peuvent servir comme support dans des explications de base, telles que le fait que l'énonciation de l'argument n au moment $t = 28$ fait que n est acceptable dans l'état suivant, i.e. au moment $t = 29$. Ou, si nous nous demandons pourquoi l'argument l est devenu acceptable à $t = 31$, une telle représentation montre que c'est parce que $makes\text{A}cceptable_l$ s'est déclenché à $t = 30$. Contrairement au cas de l'argument n que nous trouvons acceptable en raison de son énonciation récente, la relation causale actuelle indique que l'acceptabilité de l a été rétablie, ce qui signifie que tous ses attaquants ont été rendus inacceptables. Ainsi, l'utilisation de la modélisation temporelle offerte par $\mathscr{S}_c$ permet de donner une représentation visuelle du dialogue, ce qui contribue à améliorer la compréhension de l'interaction. Cependant, l'utilisation de la seule temporalité pour dériver des relations causales simples entre états n'est pas suffisante et des relations plus complexes sont nécessaires. En effet, en ce qui concerne l'acceptabilité de l'argument l, la relation causale de base qui le relie à $makes\text{A}cceptable_l$ n'est pas suffisante. Pour résoudre ce problème, un raisonnement causal est nécessaire pour trouver des causes plus complexes et plus pertinentes : les NESS-causes et les causes effectives.

8.3.2 Raisonnement causal

Dans la partie droite de la figure 8.2 nous proposons une représentation graphique plus riche pour visualiser ces relations causales plus complexes. Dans cette représentation, des relations plus complexes apparaissent, des relations qui peuvent être utilisées à des fins d'explication. Par exemple, les relations causales plus riches nous indiquent que l'énonciation de n est la cause du fait que l est acceptable et que c et m ne le sont pas, i.e. que le fait qu'il s'agisse d'une urgence entraîne la réalisation de l'IRM aujourd'hui plutôt que dans trois jours, bien qu'il n'y ait officiellement plus de créneau disponible. De plus, en exa-

minant les relations causales effectives, nous pouvons comprendre que l'énonciation de n rend l acceptable en rendant m non acceptable ; un résultat bien plus satisfaisant que celui obtenu sans relations causales complexes. Des relations plus complexes de ce type peuvent ensuite être obtenues en reconstruisant la chaîne. L'utilisation de ces relations causales peut permettre d'expliquer le dialogue de manière simple lorsque le nombre d'arguments et de relations d'attaque entre eux devient important.

8.3.3 Vers des explications

Ces chaînes causales, tout comme les représentations graphiques, peuvent ne pas représenter une explication en soi. En effet, ces relations, aussi utiles soient-elles, ont tendance à former de longues chaînes causales comprenant des redondances ou des variables inutiles. Cela soulève la question de la manière appropriée d'utiliser ces chaînes causales pour obtenir des explications. L'objectif de ce travail est donc de proposer une méthode qui permette d'extraire, à partir de cet enrichissement des graphes argumentatifs, des explications sur l'acceptabilité ou la non-acceptabilité d'un ou plusieurs arguments. Une première direction peut être trouvée dans les travaux qui étudient les liens entre les notions de causalité et d'explication. Ceux-ci sont résumés par MILLER [2019] qui établit plusieurs propriétés intéressantes qu'une explication doit satisfaire. Nous en examinerons cinq : la proximité de la conséquence, la prise en compte de la volition de l'agent, la contrastivité, la robustesse et la brièveté. La question est alors de savoir comment appliquer ces principes au nouveau cadre que nous proposons afin de fournir des explications à un utilisateur humain.

Pour les deux premiers principes, l'utilisation de $\mathscr{S}_c$ est particulièrement bien adaptée. En effet, la proximité temporelle de la conséquence est facile à évaluer et à formaliser grâce à l'inclusion du temps dans le cadre. De plus, une action délibérée est souvent préférée à un évènement naturel. Ainsi, face à deux explications identiques à l'exception d'un élément, celle dont l'action délibérée est la plus proche temporellement à la conséquence à expliquer sera préférée.

Les deux propriétés suivantes, la contrastivité et la robustesse, sont plus complexes à intégrer. Tout d'abord, à l'instar de ce que MILLER [2021] a fait avec l'approche SEF de Halpern, pour définir une explication contrastive, nous devons d'abord introduire le concept de chaîne causale contrastive. Intuitivement, les causes contrastives de a_l et $\neg a_m$ seraient les éléments communs des deux chaînes causales. Ensuite, en ce qui concerne la robustesse, une explication peut être considérée comme plus robuste qu'une autre si elle se vérifie dans un plus grand nombre de scénarios. Enfin, une explication courte est préférée à une explication longue.

En suivant ces principes, une explication pour faire une IRM aujourd'hui plutôt que dans deux jours pourrait être l'énonciation de n déclarant qu'il s'agit d'une urgence. En effet, il s'agit d'une explication courte et contrastive d'une action délibérée, temporellement proche de la conséquence. Dans ce cas, il y a une explication qui satisfait toutes les propriétés souhaitables. Dans le cas contraire, il faut s'interroger sur la manière d'agréger ces différentes propriétés.

8.4 Conclusion

Nous avons proposé dans ce chapitre une formalisation des systèmes abstraits d'argumentation acycliques dans $\mathscr{S}_c$. Cette transformation permet tout d'abord d'examiner l'effet de l'ordre d'énonciation des arguments. De plus, elle permet d'exploiter notre formalisation du raisonnement causal, offrant la possibilité de donner des informations supplémentaires sur l'acceptation ou le rejet d'un argument ainsi que des justifications sur ce dernier. Nous avons proposé deux types de représentations graphiques du processus d'argumentation formant un support visuel et ouvrant la voie à de nouvelles formes d'explications en argumentation.
Les perspectives de ce travail sont de formaliser les propriétés d'explication proposées par MILLER [2019] dans $\mathscr{S}_c$ afin de proposer une méthode de génération et d'ordonnancement d'explications conformes. La prochaine étape consistera à évaluer cette méthode de manière expérimentale en menant des études auprès des utilisateurs afin d'évaluer l'intelligibilité des explications générées, à la fois en termes de compréhension objective et de satisfaction subjective.

Dans le cadre de cette thèse, les résultats de ce chapitre servent surtout à montrer une autre application que l'éthique computationnelle pour les contributions qui y ont été présentées. Bien que $\mathscr{S}_c$ et notre approche causale commune aient été développés pour l'éthique computationnelle spécifiquement, la capacité de raisonner sur des relations causales complexes dans une représentation de l'action et du changement peut s'avérer utile dans d'autres domaines.

Conclusion et perspectives

Pour présenter les conclusions de ce travail nous reprenons la structure de l'introduction. Pour rappel, dans celle-ci nous avions séparé les contributions de cette thèse en deux groupes, celui des contributions intrinsèquement reliées à l'éthique computationnelle et celui de celles qui y contribuent mais qui peuvent aller au delà de ce domaine. Une fois ces contributions rappelées dans les grandes lignes, nous mentionnerons les perspectives de ce travail.

Modélisation des processus en éthique

L'objectif de ce premier groupe de contributions était d'apporter des clarifications utiles à l'avancement de l'éthique computationnelle. Cela a été fait en fournissant un cadre commun permettant de formaliser différentes théories morales tout en restant le plus fidèle possible à la théorie morale formalisée.

Dans un premier temps, nous nous sommes plongés dans la littérature philosophique et avons extrait ce qui nous a semblé essentiel pour pouvoir formaliser les théories morales de façon fidèle. Nous avons introduit les concepts de base nécessaires à la compréhension des différentes théories. Puis, nous avons présenté quelques unes des théories morales occidentales les plus connues. En premier ont été présentées les théories axées sur le devoir, puis celles axées sur la valeur, pour finir sur celles axées sur la vertu. La présentation de toutes ces théories morales a permis de donner un large aperçu des processus pouvant intervenir dans le raisonnement éthique.

Dans un second temps, nous avons modélisé les concepts communs à toutes les théories morales de façon suffisamment générale pour englober les formalisations déjà proposées dans le domaine. Nous avons vu que les entrées générales à toutes les théories sont le contexte et un ensemble de décisions possibles. Pour les théories axées sur la valeur nous avons défini une entrée supplémentaire, la théorie de la valeur. Ensuite, nous avons choisi pour sortie de notre cadre l'évaluation la plus courante en éthique normative, le statut déontique de la décision.

Puis, nous avons proposé une modélisation de plusieurs théories morales étant aussi bien des théories axées sur le devoir que sur la valeur. Nos modélisations ont été inscrites dans le cadre commun. Elles ont toutes été faites en utilisant une architecture modulaire qui permet d'identifier clairement tous les processus y intervenant. Nous avons distingué d'un côté le squelette clairement défini des théories, de l'autre, les processus qui ne le sont pas. De cette façon, pour s'assurer d'être fidèle à une théorie donnée, toute formalisation peut reprendre la modélisation du squelette que nous fournissons, puis proposer une version des processus indéfinis.

Nous avons ensuite réalisé une étude comparative des propositions faites en éthique computationnelle normative, plus détaillée que les états de l'art existants. Cela a permis de montrer que la séparation entre squelette et processus dans notre cadre commun permet de comparer différentes formalisations d'une théorie plus facilement. Nous avons observé que, dans le meilleur des cas ces formalisations étaient des instantiations de notre cadre, sinon, ce dernier facilite au moins la comparaison en offrant une structure générale de la théorie morale qui sert de grille d'analyse.

Pour finir, nous avons montré que traiter la causalité effective dans toute sa complexité est nécessaire à la formalisation de la plupart des théories morales, et donc à l'éthique computationnelle. Nous avons donné quelques exemples de comment l'absence d'un mécanisme permettant d'établir des relations causales complexes est à l'origine d'une partie des limites actuelles des propositions dans le domaine. En l'occurrence, nous avons montré que cette absence entraîne une simplification excessive de la représentation des problèmes et empêche de traiter tous les problèmes de surdétermination, pourtant si courants.

Modélisation, représentation et automatisation du raisonnement causal

L'objectif de ce deuxième groupe de contributions était de proposer une formalisation du raisonnement causale pouvant servir de base pour la formalisation de toutes les théories morales concernées. Autrement dit, nous avons proposé une approche causale commune permettant d'établir des relations causales complexes. Notre proposition a été conçue pour répondre à plusieurs problématiques que nous avons identifiées et qui proviennent aussi bien du domaine qui s'intéresse au raisonnement causal qu'à celui de l'éthique computationnelle. Tout d'abord, elle doit reposer sur un formalisme qui permette de rendre explicites la plupart des subtilités des problèmes, aussi bien pour le raisonnement causal que le raisonnement éthique. Puis, la définition de causalité effective choisie pour cette approche doit être dépourvue de toute confusion avec la notion de responsabilité. Sans cela, elle ne pourrait pas être commune à la formalisation de toutes les théories morales. Pour cela, elle se doit d'être purement factuelle et donc de laisser de côté toute considération normative. Finalement, l'approche proposée doit pouvoir traiter tous les cas causaux, dont les plus complexes. Cela inclut notamment les cas de surdétermination.

Pour commencer, nous avons montré que les approches de causalité existantes ne sont pas satisfaisantes pour les besoins de l'éthique computationnelle, soit parce que la définition de causalité utilisée ne convient pas, soit parce que le formalisme dans lequel elle est représentée n'est pas adapté. Nous avons ensuite montré que le test NESS était la définition qui semblait convenir le mieux comme base pour une approche de causalité positive commune adaptée à l'éthique computationnelle. En effet, celle-ci a la capacité de gérer les cas de surdétermination, et de causalité positive en général, de façon factuelle, ce qui assure une séparation entre causalité et responsabilité.

La première problématique à laquelle nous avons répondu était la question de la représentation. Nous avons présenté $\mathscr{S}_c$, un langage de description d'action adapté au raisonnement causal et éthique. Il peut être considéré comme un point intermédiaire entre PDDL et PDDL+, ou entre $\mathscr{A}_c$ et $\mathscr{C}$. Celui-ci nous permet de rendre explicites la plupart des subtilités des problèmes éthiques et causaux. Nous avons vu que le choix de modéliser le monde et

son évolution comme un système de transition d'états a tout de suite de nombreux avantages par rapport aux formalismes classiquement utilisés en causalité. En plus de nous avoir permis de saisir les subtilités de chaque problème, cela nous a permis d'étudier séparément différentes problématiques propres à la causalité.

La deuxième problématique à laquelle nous avons répondu était la question des problèmes de surdétermination. Cela a été fait grâce à un travail de clarification. Pour commencer, nous avons proposé une définition formelle de surdétermination et de chemin causal, deux concepts que nous avons montré être indispensables pour pouvoir identifier précisément les différents cas de surdétermination. Ensuite, nous avons fait une proposition de typologie qui permet de classifier clairement les cas de surdétermination de la littérature dans six catégories distinctes. En nous appuyant sur cette typologie, nous avons montré la sensibilité de la causalité à la représentation. De plus, cette typologie nous a permis d'améliorer considérablement la comparaison entre différentes approches en causalité en rendant possible d'établir des propriétés qui caractérisent la façon dont une définition de causalité va considérer tous les exemples appartenant à un type de surdétermination.

La dernière partie de notre contribution a été d'intégrer les définitions de causalité adéquates dans le langage de description d'action $\mathscr{S}_c$. Ce processus a été réalisé en deux étapes. Dans un premier temps, nous nous sommes intéressés à ce que nous avons appelé la causalité positive. Nous avons représenté le test NESS de WRIGHT [2011] dans $\mathscr{S}_c$ et nous en avons fourni une implémentation complète et correcte en Answer Set Programming. Le programme logique obtenu permet de raisonner sur des situations causales complexes. Un tel programme permet à l'éthique computationnelle de pouvoir traiter des cas qui ne l'étaient pas auparavant. Plus généralement, il permet l'exploration de relations causales complexes dans un cadre de prise de décision. Nous avons montré qu'il pouvait être utile en dehors de l'éthique computationnelle en l'utilisant dans le domaine de l'argumentation abstraite.
Dans un second temps, nous nous sommes intéressés à ce que nous avons appelé la causalité négative et la transitivité. Nous avons montré que toutes ces notions ne sont pas des notions purement causales puisqu'elles font intervenir des aspects normatifs propres aux questions de responsabilité. De ce fait, chaque théorie morale peut avoir une version différente de comment ces notions doivent être définies, ce qui rend impossible de les inclure dans notre approche commune. Toutefois, nous avons montré que notre approche pour traiter la causalité positive était une base propice pour formaliser différentes versions de ces notions.

Perspectives

Modélisation des processus en éthique Les premières perspectives dont nous allons parler concernent le premier groupe de contributions. Nous allons donc faire référence au cadre commun pour la formalisation de théories morales dans lequel nous avons modélisé les principales théories morales en philosophie occidentale.

Une première direction de recherche consiste à élargir le spectre couvert par cet outil. En l'occurrence, il semble pertinent d'élargir notre vision en nous intéressant à d'autres théories morales, notamment à celles proposées en dehors du monde occidental. Dans ce même esprit, s'inspirer de ce qui est fait en philosophie pour proposer une façon d'évaluer les différentes formalisations d'une théorie morale serait une grande contribution à l'éthique computationnelle.

Pour la modélisation des théories morales, nous avons adopté une architecture modulaire permettant d'identifier clairement le squelette de chaque théorie et les processus qui s'y attachent. En nous appuyant sur cette structure, nous avons proposé une étude comparative de travaux en éthique computationnelle normative. Une deuxième direction de recherche possible consiste à élargir cette étude comparative en incluant un plus grand nombre de travaux. Cette étude permettrait de voir les différentes propositions faites pour chaque processus et de repérer les processus pour lesquelles une meilleure proposition peut être faite. Une fois cela fait, une troisième direction de recherche possible se présente comme une évidence. Elle consiste à proposer une formalisation de chacun des processus qu'il reste à améliorer. La formalisation de chacun de ces processus représente un travail important. La preuve en est, toute la deuxième contribution de cette thèse peut être vue comme la formalisation du processus π_{cons}, en tout cas la base commune à toutes les théories morales de ce processus. Dans les questions qu'il reste à résoudre, nous retrouvons par exemple la façon dont doivent être prises en compte les évaluations de chaque action par un code de conduite dans l'évaluation de la décision. Pour rester dans les théories axées sur le devoir, une deuxième question pourrait être au sujet de la formulation des maximes dans la théorie morale de Kant. Il serait par exemple intéressant d'explorer comment formaliser l'idée de TIMMONS [2012] qui suggère d'utiliser le principe moral LU comme procédure de décision et le principe moral HFS comme une sorte d'heuristique permettant de formuler les maximes selon lesquelles les actions sont réalisées.

Représentation de l'action et du changement Les perspectives du deuxième groupe de contributions sont séparées en trois parties : la représentation de l'action et du changement, la formalisation de la causalité et la formalisation de la responsabilité. Nous traitons ici la première partie. Nous allons donc faire référence au langage $\mathscr{S}_c$ que nous avons proposé.

Une première direction de recherche consisterait à enrichir le langage proposé. Pour rappel, celui-ci a été conçu pour représenter les subtilités nécessaires au raisonnement causal et éthique, tout en excluant des éléments qui complexifient considérablement la représentation et le raisonnement causal, sans apporter de réelle plus value au raisonnement éthique. Pour ce qui est de la représentation du monde physique, il semble que tous les éléments réellement nécessaires ont été inclus. Cependant, une des principales perspectives d'amélioration repose sur l'intégration de la capacité de raisonner sur des aspects épistémiques. En effet, nous avons montré que pour certaines théories morales, les états mentaux des agents sont indispensables à prendre en compte pour évaluer une décision. Il s'agit donc d'une étape nécessaire si nous voulons que $\mathscr{S}_c$ puisse être utilisé comme base pour la formalisation de toutes les théories morales. L'intégration de ces aspects permettrait dans un deuxième temps de prendre en compte des cas de manipulation en permettant de représenter la manipulation de la volition d'un agent par un autre.

Une étude approfondie de l'impact de la granularité de la représentation sur l'évaluation éthique, ou sur les relations causales, est nécessaire. Comme nous l'avons mentionné et montré dans cette thèse, aussi bien l'éthique que la causalité sont sensibles à la façon dont le problème traité est représenté. Ajouter ou enlever des détails peut modifier le résultat du raisonnement. Il est donc important d'étudier cet impact et de réfléchir, à la fois aux solutions permettant de mitiger les effets indésirables, qu'aux bonnes pratiques permettant de trouver le niveau de granularité permettant la représentation la plus fidèle possible pour chaque problème.

Pour finir, il ne peut être que bénéfique de construire des passerelles entre les langages de représentation de l'action et du changement et les langages utilisés classiquement en causalité. En effet, cela permettrait un échange plus fluide entre ces deux domaines qui ont tant de choses à partager. Achever la conception de l'outil permettant de passer d'un programme commun en PDDL vers un programme en SEF, en Calcul des Situations ou en $\mathscr{S}_c$, semble être une piste prometteuse dans cette direction. Cet outil peut être très intéressant pour l'évaluation des approches causales car les derniers travaux réalisés dessus ont permis de proposer des pistes pour comparer les relations causales obtenues pour un même problème par différentes approches causales. Cet outil permettrait également d'étudier s'il est possible de trouver un équivalent en SEF de la typologie des cas de surdétermination que nous avons proposé. Il semble probable que certaines subtilités nécessaires à la typologie disparaissent lors du passage en SEF. Toutefois, il ne s'agit pour l'instant que d'une hypothèse. Pour la vérifier, il serait par exemple possible d'étudier si différents exemples de PDDL, appartenant à différentes catégories de notre typologie, sont traduits de la même manière en SEF. Si c'est le cas, la typologie pourrait ne pas être entièrement applicable. En revanche, si des traductions différentes étaient obtenues, la compréhension des mécanismes sous-jacents permettant une telle distinction permettrait de proposer une version adaptée.

Formalisation du raisonnement causal Nous traitons ici la deuxième partie des perspectives. Nous allons donc faire référence à la formalisation que nous avons proposé du raisonnement causal.

La première direction de recherche qu'il serait intéressant d'explorer concerne la surdétermination. Grâce à la typologie proposée, nous avons prouvé quelle allait être la réponse de notre approche causale pour tous les problèmes appartenant à un des six cas de surdétermination identifiés. Il s'agissait là des cas les plus couramment trouvés dans le domaine. Toutefois, l'analyse formelle de la surdétermination a montré l'existence de cas qui, à notre connaissance, n'ont jamais été envisagés. En l'occurrence, dans quel cas de surdétermination sommes nous lorsque les chemins causaux individuels interagissent pour en créer de nouveaux? Faut-il considérer que différents types de cas de surdétermination peuvent coexister? Par exemple, un cas de préemption tardive peut-il coexister avec un cas symétrique/duplicatif, ou y a-t-il des règles de subsomption? Étudier ces cas et répondre à ces questions permettrait d'en apprendre plus sur les problèmes de surdétermination. Une fois la frontière des connaissances sur ces problèmes repoussée, il faudrait évaluer si la façon dont notre approche gère ces cas est satisfaisante.

Si la première direction de recherche était une réflexion générale sur la causalité, la deuxième est plutôt spécifique à l'approche que nous avons proposé. Plus spécifiquement, il s'agit de prouver que l'implémentation de la causalité dans $\mathscr{S}_c^+$ proposée est correcte et complète, comme cela a été fait pour $\mathscr{S}_c$. Prouver cela aurait un bénéfice supplémentaire, cela montrerait que la preuve pour $\mathscr{S}_c$ est valide avec une formule ψ sous n'importe qu'elle forme, ce qui n'est pas le cas actuellement. Les preuves de complétude et de correction semblent n'être possibles que si le programme construit les ensembles suffisants de NESS-causes et NESS-causes directes, ce qui n'est pas le cas de $\pi_{\mathbb{C}}$. L'ASP n'étant pas un langage permettant de manipuler facilement les ensembles, une implémentation en Prolog semble plus adéquate. Comme mentionné plus tôt dans cette thèse, une première exploration de cette direction de recherche a été faite par Ella Dijksman. Ce stage a permis d'obtenir des programmes $\pi_{\mathbb{A}}$' et $\pi_{\mathbb{C}}$' en Prolog reprenant $\mathscr{S}_c$ et ses définitions causales, mais en construi-

sant les ensembles suffisants de NESS-causes et NESS-causes directes. Il reste à ajouter à ces programmes les effets conditionnels de $\mathscr{S}_c^+$, adapter les définitions causales et prouver la complétude et correction.

Formalisation de la responsabilité Nous traitons ici la troisième et dernière partie des perspectives. Nous allons donc faire référence à la formalisation de la responsabilité que nous avons montrée possible avec notre approche.

Pour pouvoir dire que nous avons fait une formalisation du π_{cons} d'une théorie morale, il est nécessaire de proposer, en plus de notre approche commune, une définition adaptée à cette théorie morale des notions de causalité négative et de transitivité. En effet, ayant montré que ces notions faisaient intervenir des aspects normatifs propres aux questions de responsabilité, il est impossible de proposer une définition convenant à toutes les théories morales. La première étape du travail restant à faire est de se plonger dans la littérature philosophique pour déterminer s'il est possible d'identifier la position des différentes théories morales par rapport à ces notions. Il serait alors intéressant d'étudier s'il existe des points communs entre elles. Pour finir, une fois cette position identifiée pour chaque théorie morale, celle-ci devrait être formalisée dans $\mathscr{S}_c$. Idéalement, cette formalisation prendrait comme base la formalisation du raisonnement causal que nous avons proposé et que nous avons montré être une base propice à formaliser différentes visions de responsabilité.

Bibliographie

ABARCA, A. I. R. et J. M. BROERSEN. 2022, «A Stit Logic of Responsibility», dans *Proceedings of the Twenty-First International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*, édité par P. Faliszewski, V. Mascardi, C. Pelachaud et M. E. Taylor, International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), Auckland, New Zealand, p. 1717–1719, doi :10.5555/3535850.3536087. URL https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p1717.pdf. 69

ABRAMS, Y. 2022, «Omissive Overdetermination : Why the Act-Omission Distinction Makes a Difference for Causal Analysis», *University of Western Australia Law Review*, vol. 49, nº 57. URL https://papers.ssrn.com/abstract=4061990. 81, 216, 219, 221, 223, 224, 225, 226, 228, 229, 248

ALLEN, C., I. SMIT et W. WALLACH. 2005, «Artificial Morality : Top-down, Bottom-up, and Hybrid Approaches», *Ethics and Information Technology*, vol. 7, nº 3, doi :10.1007/s10676-006-0004-4, p. 149–155, ISSN 1572-8439. URL https://doi.org/10.1007/s10676-006-0004-4. 36

ALLEN, J. F. 1983, «Maintaining Knowledge about Temporal Intervals», *Communications of the ACM*, vol. 26, nº 11, doi :10.1145/182.358434, p. 832–843. URL https://doi.org/10.1145/182.358434. 148, 234

ANDERSON, M., S. ANDERSON et C. ARMEN. 2004, «Towards Machine Ethics», dans *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Workshop on Agent Organizations : Theory and Practice, AAAI 2004*, AAAI Press, San Jose, California, USA, p. 2–7. 119

ANDERSON, M. et S. L. ANDERSON. 2008, «ETHEL : Toward a Principled Ethical Eldercare System», dans *Proceedings of the Sixteenth AAAI Fall Symposium, AAAI 2008, AAAI Technical Report*, vol. FS-08-02, AAAI Press, Arlington, Virginia, USA, p. 4–11. URL http://www.aaai.org/Library/Symposia/Fall/2008/fs08-02-002.php. 119

ANDERSON, M., S. L. ANDERSON et C. ARMEN. 2006, «MedEthEx : A Prototype Medical Ethics Advisor», dans *Proceedings of the Twenty-First National Conference on Artificial Intelligence, AAAI 2006*, AAAI Press, Boston, Massachusetts, USA, p. 1759–1765. URL http://www.aaai.org/Library/AAAI/2006/aaai06-292.php. I

ANDREAS, H. et M. GUENTHER. 2021, «Regularity and Inferential Theories of Causation», dans *The Stanford Encyclopedia of Philosophy*, édité par E. N. Zalta, fall 2021 éd., Metaphysics Research Lab, Stanford University, p. 1–24. URL https://plato.stanford.edu/archives/fall2021/entries/causation-regularity/. 7, 61, 62, 69, 172

ARKOUDAS, K., S. BRINGSJORD et P. BELLO. 2005, «Toward Ethical Robots via Mechanized Deontic Logic», dans *Proceedings of the Fifteenth AAAI Fall Symposium, AAAI 2008, AAAI Technical Report*, vol. FS-05-06, AAAI Press, Arlington, Virginia, USA, p. 17–23. 119

ASHLEY, K. D. et B. M. MCLAREN. 1994, «A CBR Knowledge Representation for Practical Ethics», dans *Proceedings of the Second European Workshop on Advances in Case-Based Reasoning, EWCBR 1994, Lecture Notes in Computer Science*, vol. 984, édité par

J. P. Haton, M. T. Keane et M. Manago, Springer, Chantilly, France, p. 181–197, doi : 10.1007/3-540-60364-6_36. URL https://doi.org/10.1007/3-540-60364-6_36. 119

ATKINSON, K. et T. J. M. BENCH-CAPON. 2008, «Addressing moral problems through practical reasoning», *Journal of Applied Logic*, vol. 6, nº 2, doi :10.1016/J.JAL.2007.06.005, p. 135–151. URL https://doi.org/10.1016/j.jal.2007.06.005. 119

BARAL, C. et M. GELFOND. 1997, «Reasoning About Effects of Concurrent Actions», *Journal of Logic Programming*, vol. 31, nº 1-3, doi :10.1016/S0743-1066(96)00140-9, p. 85–117. URL https://doi.org/10.1016/S0743-1066(96)00140-9. 40, 49, 53, 162

BARAL, C. et M. GELFOND. 2000, «Reasoning Agents in Dynamic Domains», dans *Logic-Based Artificial Intelligence*, édité par J. Minker, The Springer International Series in Engineering and Computer Science, Springer US, Boston, MA, ISBN 978-1-4615-1567-8, p. 257–279, doi :10.1007/978-1-4615-1567-8_12. URL https://doi.org/10.1007/978-1-4615-1567-8_12. 72

BATUSOV, V. et M. SOUTCHANSKI. 2018, «Situation Calculus Semantics for Actual Causality», dans *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018*, édité par S. A. McIlraith et K. Q. Weinberger, AAAI Press, New Orleans, Louisiana, USA, p. 1744–1752, doi :10.1609/aaai.v32i1.11561. URL https://doi.org/10.1609/aaai.v32i1.11561. 8, 69, 74, 83, 84, 141, 144, 156, 158, 176, 191, 204, 214

BAUMGARTNER, M. 2013, «A Regularity Theoretic Approach to Actual Causation», *Erkenntnis*, vol. 78, p. 85–109, ISSN 0165-0106. URL https://www.jstor.org/stable/24010952. 71, 74, 81, 82, 83, 86, 88, 142, 173

BECKERS, S. 2021a, «Causal Sufficiency and Actual Causation», *Journal of Philosophical Logic*, vol. 50, nº 6, doi :10.1007/s10992-021-09601-z, p. 1341–1374. URL https://doi.org/10.1007/s10992-021-09601-z. 136, 137, 154

BECKERS, S. 2021b, «The Counterfactual NESS Definition of Causation», dans *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, AAAI Press, Virtual Event, p. 6210–6217, doi :10.1609/aaai.v35i7.16772. URL https://doi.org/10.1609/aaai.v35i7.16772. 66, 67, 72, 74, 79, 80, 82, 142

BECKERS, S. et J. VENNEKENS. 2018, «A Principled Approach to Defining Actual Causation», *Synthese*, vol. 195, nº 2, doi :10.1007/s11229-016-1247-1, p. 835–862. URL https://doi.org/10.1007/s11229-016-1247-1. 80

BEEBEE, H., C. HITCHCOCK et P. MENZIES, éd.. 2009, *The Oxford Handbook of Causation*, Oxford Handbooks Online, Oxford University Press, ISBN 978-0-19-927973-9. 95, 133

BENGEL, L., L. BLÜMEL, T. RIENSTRA et M. THIMM. 2022, «Argumentation-based Causal and Counterfactual Reasoning», dans *Proceedings of the First International Workshop on Argumentation for eXplainable AI, ArgXAI 2022, CEUR Workshop Proceedings*, vol. 3209, édité par K. Cyras, T. Kampik, O. Cocarascu et A. Rago, CEUR-WS.org, Cardiff, Wales, UK, p. 1–12. URL https://ceur-ws.org/Vol-3209/7343.pdf. 259

BENNETT, J. 1998, *The Act Itself*, Oxford Scholarship Online, Oxford University Press, ISBN 978-0-19-823791-4. 219

BENTHAM, J. 1789, *An introduction to the principles of morals and legislation*, T. Payne. 30, 109, 126

BERREBY, F., G. BOURGNE et J.-G. GANASCIA. 2015, «Modelling Moral Reasoning and Ethical Responsibility with Logic Programming», dans *Proceedings of the Twentieth International Conference on Logic for Programming, Artificial Intelligence, and Reasoning, LPAR 2015*, *Lecture Notes in Computer Science*, vol. 9450, édité par M. Davis, A. Fehnker, A. McIver et A. Voronkov, Springer, Suva, Fiji, p. 532–548, doi :10.1007/978-3-662-48899-7_37. URL https://doi.org/10.1007/978-3-662-48899-7_37. 3, II

BERREBY, F., G. BOURGNE et J.-G. GANASCIA. 2017, «A Declarative Modular Framework for Representing and Applying Ethical Principles», dans *Proceedings of the Sixteenth International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017*, édité par K. Larson, M. Winikoff, S. Das et E. H. Durfee, ACM, São Paulo, Brazil, p. 96–104. URL http://dl.acm.org/citation.cfm?id=3091145. 3, 119, 120, 122, 123, 124, 125, 126, 127, 128, 129, IV

BERREBY, F., G. BOURGNE et J.-G. GANASCIA. 2018, «Event-Based and Scenario-Based Causality for Computational Ethics», dans *Proceedings of the Seventeenth International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018*, édité par E. André, S. Koenig, M. Dastani et G. Sukthankar, International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, Stockholm, Sweden, p. 147–155. URL http://dl.acm.org/citation.cfm?id=3237412. 3, 8, 71, 72, 84, 86, 91, 141, 158, 183, 191, 251, 252, 257

BOCHMAN, A. 2004, «A causal approach to nonmonotonic reasoning», *Artifificial Intelligence*, vol. 160, nº 1-2, doi :10.1016/J.ARTINT.2004.07.002, p. 105–143. URL https://doi.org/10.1016/j.artint.2004.07.002. 71

BOCHMAN, A. 2005, «Propositional Argumentation and Causal Reasoning», dans *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, IJCAI 2005*, édité par L. P. Kaelbling et A. Saffiotti, Professional Book Center, Edinburgh, Scotland, UK, p. 388–393. URL http://ijcai.org/Proceedings/05/Papers/0306.pdf. 259

BOCHMAN, A. 2018a, «Actual Causality in a Logical Setting», dans *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, édité par J. Lang, ijcai.org, Stockholm, Sweden, p. 1730–1736, doi :10.24963/ijcai.2018/239. URL https://doi.org/10.24963/ijcai.2018/239. 71, 83, 88, 152

BOCHMAN, A. 2018b, «On Laws and Counterfactuals in Causal Reasoning», dans *Proceedings of the Sixteenth International Conference on Principles of Knowledge Representation and Reasoning, KR 2018*, édité par M. Thielscher, F. Toni et F. Wolter, AAAI Press, Tempe, Arizona, USA, p. 494–503. URL https://aaai.org/ocs/index.php/KR/KR18/paper/view/17990. 71, 73, 75, 79, 81, 82, 144

BOCHMAN, A. 2021, *A Logical Theory of Causality*, The MIT Press, ISBN 978-0-262-04532-2. 259

BONNEMAINS, V., C. SAUREL et C. TESSIER. 2018, «Embedded Ethics : Some Technical and Ethical Challenges», *Ethics and Information Technology*, vol. 20, nº 1, doi :10.1007/

s10676-018-9444-x, p. 41–58. URL https://doi.org/10.1007/s10676-018-9444-x. 99, 103, 119, 121, 122, 123, 125, 126, 127, 128, 129, 132, V

BOURGNE, G., C. SARMIENTO et J.-G. GANASCIA. 2021, «ACE Modular Framework for Computational Ethics : Dealing with Multiple Actions, Concurrency and Omission», dans *Proceedings of the First International Workshop on Computational Machine Ethics, CME 2021*, CEUR Workshop Proceedings, CEUR-WS.org, Virtual Event, p. 1–6. 3, 4, 8, 71, 84, 193, 218, 298, VI

BRAHAM, M. et M. VAN HEES. 2012, «An Anatomy of Moral Responsibility», *Mind*, vol. 121, nº 483, p. 601–634, ISSN 0026-4423. URL https://www.jstor.org/stable/23321778. 225, 241

BRINGSJORD, S. et J. TAYLOR. 2012, «Introducing Divine-Command Robot Ethics», dans *Robot Ethics : The Ethical and Social Implications of Robotics*, MIT Press, Cambridge, Massachussetts, USA, p. 85–108. 119, 120, 122, 123, 128

BUOLAMWINI, J. et T. GEBRU. 2018, «Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification», dans *Proceedings of the First Conference on Fairness, Accountability and Transparency, FAT 2018, Proceedings of Machine Learning Research*, vol. 81, édité par S. A. Friedler et C. Wilson, PMLR, New York, NY, USA, p. 77–91. URL http://proceedings.mlr.press/v81/buolamwini18a.html. 2

CARDOSO, R. C., A. FERRANDO, L. A. DENNIS et M. FISHER. 2022, «Implementing Ethical Governors in BDI», dans *Proceedings of the Nineth International Workshop on Engineering Multi-Agent Systems, EMAS 2022, Lecture Notes in Artificial Intelligence*, vol. 13190, édité par N. Alechina, M. Baldoni et B. Logan, Springer International Publishing, Auckland, New Zealand, ISBN 978-3-030-97457-2, p. 22–41, doi :10.1007/978-3-030-97457-2_2. URL https://doi.org/10.1007/978-3-030-97457-2_2. 130

CERVANTES, J.-A., L.-F. RODRÍGUEZ, S. LÓPEZ, F. RAMOS et F. ROBLES. 2016, «Autonomous Agents and Ethical Decision-Making», *Cognitive Computation*, vol. 8, nº 2, doi :10.1007/S12559-015-9362-8, p. 278–296. URL https://doi.org/10.1007/s12559-015-9362-8. 119

CHANGEUX, J.-P. et A. CONNES. 2008, *Matière à pensée*, Bibliothèque, Odile Jacob, Paris, ISBN 978-2-7381-1923-0. 37

CHOI, S. et M. FARA. 2021, «Dispositions», dans *The Stanford Encyclopedia of Philosophy*, édité par E. N. Zalta, spring 2021 éd., Metaphysics Research Lab, Stanford University, p. 1–22. URL https://plato.stanford.edu/archives/spr2021/entries/dispositions/. 251

CITRON, D. K. et F. A. PASQUALE. 2014, «The Scored Society : Due Process for Automated Predictions», SSRN Scholarly Paper ID 2376209, Social Science Research Network, Rochester, NY. URL https://papers.ssrn.com/abstract=2376209. 3

CLOOS, C. 2005, «The Utilibot project : An autonomous mobile robot based on utilitarianism», dans *Proceedings of the Thirteenth AAAI Fall Symposium, AAAI 2005, AAAI Technical Report*, vol. FS-06, AAAI Press, Arlington, Virginia, USA, p. 38–45. 119

COINTE, N., G. BONNET et O. BOISSIER. 2016, «Ethical Judgment of Agents' Behaviors in Multi-Agent Systems», dans *Proceedings of the Fifteenth International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2016*, édité par C. M. Jonker, S. Marsella, J. Thangarajah et K. Tuyls, ACM, Singapore, p. 1106–1114. URL http://dl.acm.org/citation.cfm?id=2937086. 102, 119, III

COLLINS, J. 2000, «Preemptive Prevention», *The Journal of Philosophy*, vol. 97, nº 4, doi : 10.2307/2678391, p. 223–234, ISSN 0022-362X. URL https://www.jstor.org/stable/2678391. 86, 87, 241

COLLINS, J., N. HALL et L. A. PAUL. 2004, «Counterfactuals and Causation : History, Problems, and Prospects», dans *Causation and Counterfactuals*, Representation and Mind series, The MIT Press, ISBN 978-0-262-53256-3, p. 1–57. 84

D'AQUIN, T. 1266, *Summa theologica*, Wikisource. URL https://en.wikisource.org/wiki/Summa_Theologiae. 32, 114

DEHGHANI, M., E. TOMAI, K. D. FORBUS et M. KLENK. 2008, «An Integrated Reasoning Approach to Moral Decision-Making», dans *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008*, édité par D. Fox et C. P. Gomes, AAAI Press, Chicago, Illinois, USA, p. 1280–1286. URL http://www.aaai.org/Library/AAAI/2008/aaai08-203.php. 119

DENNIS, L. A., M. M. BENTZEN, F. LINDNER et M. FISHER. 2021, «Verifiable Machine Ethics in Changing Contexts», dans *Proceedings of the Thirty-Fifth Conference on Artificial Intelligence, AAAI 2021*, AAAI Press, Virtual Event, p. 11 470–11 478, doi :10.1609/aaai.v35i13.17366. URL https://doi.org/10.1609/aaai.v35i13.17366. 130, VII

DENNIS, L. A., M. FISHER, M. SLAVKOVIK et M. WEBSTER. 2016, «Formal verification of ethical choices in autonomous systems», *Robotics and Autonomous Systems*, vol. 77, doi :10.1016/J.ROBOT.2015.11.012, p. 1–14. URL https://doi.org/10.1016/j.robot.2015.11.012. 119

DENNIS, L. A., M. FISHER et A. F. T. WINFIELD. 2015, «Towards Verifiably Ethical Robot Behaviour», dans *Proceedings of the AAAI Workshop on Artificial Intelligence and Ethics*, édité par T. Walsh, AAAI Technical Report, AAAI Press, Austin, Texas, USA, p. 2–11. URL http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10119. 119, II

DOUTRE, S., F. MAFFRE et P. MCBURNEY. 2017, «A Dynamic Logic Framework for Abstract Argumentation : Adding and Removing Arguments», dans *Proceedings of the Thirtieth International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Lecture Notes in Computer Science*, vol. 10351, édité par S. Benferhat, K. Tabia et M. Ali, Springer, Arras, France, p. 295–305, doi :10.1007/978-3-319-60045-1_32. URL https://doi.org/10.1007/978-3-319-60045-1_32. 259

DOWE, P. 2000, *Physical Causation*, Cambridge Studies in Probability, Induction and Decision Theory, Cambridge University Press, Cambridge, ISBN 978-0-521-78049-0, doi :10.1017/CBO9780511570650. URL https://www.cambridge.org/core/books/physical-causation/D056895488F735AC513E455D3683497F. 219

DOWE, P. 2001, «A Counterfactual Theory of Prevention and 'Causation' by Omission», *Australasian Journal of Philosophy*, vol. 79, nº 2, doi :10.1080/713659223, p. 216–226, ISSN 0004-8402. URL https://doi.org/10.1080/713659223. 86

DUNG, P. M. 1995, «On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games», *Journal of Artificial Intelligence*, vol. 77, nº 2, doi :10.1016/0004-3702(94)00041-X, p. 321–358. URL https://doi.org/10.1016/0004-3702(94)00041-X. 260

EPSTEIN, R. A. 1973, «A Theory of Strict Liability», *The Journal of Legal Studies*, vol. 2, nº 1, doi :10.1086/467495, p. 151–204, ISSN 0047-2530. URL https://www.journals.uchicago.edu/doi/abs/10.1086/467495. 220

FAIR, D. 1979, «Causation and the Flow of Energy», *Erkenntnis (1975-)*, vol. 14, nº 3, p. 219–250, ISSN 0165-0106. URL https://www.jstor.org/stable/20010665. 220

FIKES, R. et N. J. NILSSON. 1971, «STRIPS : A New Approach to the Application of Theorem Proving to Problem Solving», *Artificial Intelligence*, vol. 2, nº 3/4, doi :10.1016/0004-3702(71)90010-5, p. 189–208. URL https://doi.org/10.1016/0004-3702(71)90010-5. 47

FISCHER, D. 1992, «Causation in Fact in Omission Cases», *Utah Law Review*, p. 1335–1349. URL https://scholarship.law.missouri.edu/facpubs/185. 221

FOOT, P. 2002, «Morality, Action, and Outcome», dans *Moral Dilemmas : and Other Topics in Moral Philosophy*, édité par P. Foot, Oxford University Press, ISBN 978-0-19-925284-8, p. 88–104, doi :10.1093/019925284X.003.0007. URL https://doi.org/10.1093/019925284X.003.0007. 219

FOX, M. et D. LONG. 2003, «PDDL2.1 : An Extension to PDDL for Expressing Temporal Planning Domains», *Journal of Artificial Intelligence Research*, vol. 20, doi :10.1613/jair.1129, p. 61–124. URL https://doi.org/10.1613/jair.1129. 47, 205

FOX, M. et D. LONG. 2006, «Modelling Mixed Discrete-Continuous Domains for Planning», *Journal of Artificial Intelligence Research*, vol. 27, doi :10.1613/jair.2044, p. 235–297. URL https://doi.org/10.1613/jair.2044. 45, 47, 162

FUMERTON, R. et K. KRESS. 2001, «Causation and the Law : Preemption, Lawful Sufficiency, and Causal Sufficiency», *Law and Contemporary Problems*, vol. 64, nº 4, doi :10.2307/1192292, p. 83–105, ISSN 0023-9186. URL https://www.jstor.org/stable/1192292. 84

FURBACH, U., C. SCHON et F. STOLZENBURG. 2014, «Automated Reasoning in Deontic Logic», dans *Proceedings of the Eighth International Workshop on Multi-disciplinary Trends in Artificial Intelligence, MIWAI 2014, Lecture Notes in Computer Science*, vol. 8875, édité par M. N. Murty, X. He, C. R. Rao et P. Weng, Springer, Bangalore, India, p. 57–68, doi :10.1007/978-3-319-13365-2_6. URL https://doi.org/10.1007/978-3-319-13365-2_6. 119

GANASCIA, J.-G. 2007, «Ethical System Formalization using Non-Monotonic Logics», dans *Proceedings of the Twenty-Ninth Cognitive Science Conference, CogSci2007*, vol. 29, Nashville, Tennessee, USA, p. 1013. URL https://hal.science/hal-01336276. 119, 121, 123, 124

GANASCIA, J.-G. 2015, «Non-monotonic Resolution of Conflicts for Ethical Reasoning», dans *A Construction Manual for Robots' Ethical Systems - Requirements, Methods, Implementations*, édité par R. Trappl, Cognitive Technologies, Springer, ISBN 978-3-319-21547-1, p. 101–118, doi :10.1007/978-3-319-21548-8_6. URL https://doi.org/10.1007/978-3-319-21548-8_6. 192, III

GELFOND, M. et V. LIFSCHITZ. 1991, «Classical Negation in Logic Programs and Disjunctive Databases», *New Generation Computing*, vol. 9, nº 3/4, doi :10.1007/BF03037169, p. 365–386. URL https://doi.org/10.1007/BF03037169. 192

GELFOND, M. et V. LIFSCHITZ. 1993, «Representing Action and Change by Logic Programs», *Journal of Logic Programming*, vol. 17, nº 2/3&4, doi :10.1016/0743-1066(93)90035-F, p. 301–321. URL https://doi.org/10.1016/0743-1066(93)90035-F. 48

GELFOND, M. et V. LIFSCHITZ. 1998, «Action Languages», *Electronic Transactions on Artificial Intelligence*, vol. 2, p. 193–210. URL http://www.ep.liu.se/ej/etai/1998/007/. 41, 99

GELFOND, M., V. LIFSCHITZ et A. RABINOV. 1991, «What are the Limitations of the Situation Calculus?», dans *Automated Reasoning : Essays in Honor of Woody Bledsoe*, édité par R. S. Boyer, Automated Reasoning Series, Kluwer Academic Publishers, p. 167–180. 54

GHALLAB, M., C. KNOBLOCK, D. WILKINS, A. BARRETT, D. CHRISTIANSON, M. FRIEDMAN, C. KWOK, K. GOLDEN, S. PENBERTHY, D. SMITH, Y. SUN et D. WELD. 1998, «PDDL - The Planning Domain Definition Language», Technical Report CVC TR-98-003/DCS TR-1165, Yale Center for Computational Vision and Control. 44, 47, 205

GILLIGAN, C. 1982, *In a Different Voice : Psychological Theory and Women's Development*, Harvard University Press. 35

GIPS, J. 1995, «Towards the Ethical Robot», dans *Android Epistemology*, MIT Press, Cambridge, MA, USA, ISBN 0-262-06184-8, p. 243–252. URL https://dl.acm.org/doi/10.5555/216350.216375. 7

GIUNCHIGLIA, E. et V. LIFSCHITZ. 1998, «An Action Language Based on Causal Explanation : Preliminary Report», dans *Proceedings of the Fifteenth National Conference on Artificial Intelligence, AAAI 98*, édité par J. Mostow et C. Rich, AAAI Press / The MIT Press, Madison, Wisconsin, USA, p. 623–630. URL http://www.aaai.org/Library/AAAI/1998/aaai98-088.php. 50

GLENNAN, S. 2017, *The New Mechanical Philosophy*, Oxford University Press, Oxford, New York, ISBN 978-0-19-877971-1. 220

GOVINDARAJULU, N. S. et S. BRINGSJORD. 2017, «On Automating the Doctrine of Double Effect», dans *Proceedings of the Twenty-Sixth International Joint Conference on Artificial*

Intelligence, IJCAI 2017, édité par C. Sierra, ijcai.org, Melbourne, Australia, p. 4722–4730, doi :10.24963/ijcai.2017/658. URL https://doi.org/10.24963/ijcai.2017/658. 119, 122, 127, 128, 129, IV

GREEN, S. 2017, *Causation in Negligence*, Hart Studies in Private Law, Bloomsbury Publishing, ISBN 978-1-5099-0503-4. 221

HALL, N. 2004, «Two Concepts of Causation», dans *Causation and Counterfactuals*, édité par J. Collins, N. Hall et L. A. Paul, Representation and Mind series, The MIT Press, ISBN 978-0-262-53256-3, p. 225–276. 84

HALL, N. 2007, «Structural Equations and Causation», *Philosophical Studies : An International Journal for Philosophy in the Analytic Tradition*, vol. 132, nº 1, p. 109–136, ISSN 0031-8116. URL https://www.jstor.org/stable/25471849. 67, 81

HALL, N. et L. A. PAUL. 2003, «Causation and Preemption», dans *Philosophy of Science Today*, Oxford University Press, ISBN 978-0-19-925055-4, p. 100–130. 6, 66, 73, 74, 89, 150, 152

HALPERN, J. Y. 2000, «Axiomatizing Causal Reasoning», *Journal of Artificial Intelligence Research*, vol. 12, doi :10.1613/jair.648, p. 317–337. URL https://doi.org/10.1613/jair.648. 83

HALPERN, J. Y. 2015, «A Modification of the Halpern-Pearl Definition of Causality», dans *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, édité par Q. Yang et M. J. Wooldridge, AAAI Press, Buenos Aires, Argentina, p. 3022–3033. URL http://ijcai.org/Abstract/15/427. 67, 69

HALPERN, J. Y. 2016, *Actual Causality*, MIT Press, ISBN 978-0-262-03502-6. 67, 68, 72, 80, 85, 154, 176, 214

HALPERN, J. Y. 2018, «Actual Causality : A Survey : Joseph Halpern», URL https://www.youtube.com/watch?v=hXnCX2pJ0sg. 80

HALPERN, J. Y. et J. PEARL. 2005, «Causes and Explanations : A Structural-Model Approach : Part 1 : Causes», *The British Journal for the Philosophy of Science*, vol. 56, nº 4, p. 843–887. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=100&proceeding_id=17. 151

HASLUM, P., N. LIPOVETZKY, D. MAGAZZENI et C. MUISE. 2019, *An Introduction to the Planning Domain Definition Language*, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, ISBN 978-3-031-00456-8, doi :10.2200/S00900ED2V01Y201902AIM042. URL https://doi.org/10.2200/S00900ED2V01Y201902AIM042. 44, 205

HITCHCOCK, C. 2001, «The Intransitivity of Causation Revealed in Equations and Graphs», *The Journal of Philosophy*, vol. 98, nº 6, doi :10.2307/2678432, p. 273–299, ISSN 0022-362X. URL https://www.jstor.org/stable/2678432. 67, 81

HITCHCOCK, C. 2007, «Prevention, Preemption, and the Principle of Sufficient Reason», *The Philosophical Review*, vol. 116, nº 4, p. 495–532, ISSN 0031-8108. URL https://www.jstor.org/stable/20446988. 67, 73, 74, 81, 83, 86, 87, 88, 143, 151, 240, 241, 250

HOFFMANN, J. et S. EDELKAMP. 2005, «The Deterministic Part of IPC-4 : An Overview», *Journal of Artificial Intelligence Research*, vol. 24, doi :10.1613/jair.1677, p. 519–579. URL https://doi.org/10.1613/jair.1677. 47, 205

HOPKINS, M. et J. PEARL. 2007, «Causality and Counterfactuals in the Situation Calculus», *Journal of Logic and Computation*, vol. 17, nº 5, doi :10.1093/logcom/exm048, p. 939–953. URL https://doi.org/10.1093/logcom/exm048. 83, 158

HUME, D. 1748, *An Enquiry Concerning Human Understanding*, Flammarion. URL https://en.wikisource.org/wiki/Philosophical_Essays_Concerning_Human_Understanding. 6, 37, 57, 61, 65, 92

JIANG, L., J. D. HWANG, C. BHAGAVATULA, R. L. BRAS, M. FORBES, J. BORCHARDT, J. T. LIANG, O. ETZIONI, M. SAP et Y. CHOI. 2021, «Delphi : Towards Machine Ethics and Norms», *CoRR*, vol. abs/2110.07574. URL https://arxiv.org/abs/2110.07574, arXiv : 2110.07574. 37

KAKAS, A. C. et R. MILLER. 1997, «A Simple Declarative Language for Describing Narratives With Actions», *Journal of Logic Programming*, vol. 31, nº 1-3, doi :10.1016/S0743-1066(96)00138-0, p. 157–200. URL https://doi.org/10.1016/S0743-1066(96)00138-0. 54

KANT, I. 1785, *Groundwork of the Metaphysics of Morals*, Hachette Livre BNF. 24, 25, 106, 107

KHAN, S. M. et Y. LESPÉRANCE. 2021, «Knowing Why - On the Dynamics of Knowledge about Actual Causes in the Situation Calculus», dans *Proceedings of the Twentieth International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2021*, édité par F. Dignum, A. Lomuscio, U. Endriss et A. Nowé, ACM, United Kingdom, p. 701–709, doi :10.5555/3463952.3464037. URL https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p701.pdf. 69, 209

KIM, T. W., J. HOOKER et T. DONALDSON. 2021, «Taking Principles Seriously : A Hybrid Approach to Value Alignment in Artificial Intelligence», *Journal of Artificial Intelligence Research*, vol. 70, doi :10.1613/jair.1.12481, p. 871–890, ISSN 1076-9757. URL https://dl.acm.org/doi/10.1613/jair.1.12481. 37, 130

KOWALSKI, R. A. 1992, «Database Updates in the Event Calculus», *Journal of Logic Programming*, vol. 12, nº 1&2, doi :10.1016/0743-1066(92)90041-Z, p. 121–146. URL https://doi.org/10.1016/0743-1066(92)90041-Z. 53

KOWALSKI, R. A. et M. J. SERGOT. 1986, «A Logic-based Calculus of Events», *New Generation Computing*, vol. 4, nº 1, doi :10.1007/BF03037383, p. 67–95. URL https://doi.org/10.1007/BF03037383. 53

KRIPKE, S. A. 1963, «Semantical Analysis of Modal Logic I Normal Modal Propositional Calculi», *Mathematical Logic Quarterly*, vol. 9, nº 5-6, doi :10.1002/malq.19630090502, p. 67–96, ISSN 1521-3870. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/malq.19630090502. 65

KUEFFNER, K. R. 2021, *A comprehensive Survey of the actual causality literature*, Thesis, Technische Universität Wien, doi :10.34726/hss.2021.90003. URL https://repositum.tuwien.at/handle/20.500.12708/18862. 59, 67

LEBLANC, E. C., M. BALDUCCINI et J. VENNEKENS. 2019, «Explaining Actual Causation via Reasoning About Actions and Change», dans *Proceedings of the Sixteenth European Conference on Logics in Artificial Intelligence, JELIA 2019, Lecture Notes in Computer Science*, vol. 11468, édité par F. Calimeri, N. Leone et M. Manna, Springer, Rende, Italy, p. 231–246, doi :10.1007/978-3-030-19570-0_15. URL https://doi.org/10.1007/978-3-030-19570-0_15. 8, 72, 84, 141, 158, 191

LEIBNIZ, G. W. 1710, *Essais de théodicée*, GF Flammarion. 61

LEWIS, C. I. et C. H. LANGFORD. 1932, *Symbolic Logic*, Century Company. 120

LEWIS, D. 1973, «Causation», *The Journal of Philosophy*, vol. 70, nº 17, doi :10.2307/2025310, p. 556–567, ISSN 0022-362X. URL https://www.jstor.org/stable/2025310. 61, 65, 66, 84, 166

LEWIS, D. 1987, *Philosophical Papers*, vol. 2, Oxford University Press, ISBN 978-0-19-503646-6. 74

LEWIS, D. 1997, «Finkish Dispositions», *The Philosophical Quarterly*, vol. 47, nº 187, p. 143–158, ISSN 0031-8094. URL https://www.jstor.org/stable/2956325. 251

LEWIS, D. 2000, «Causation as Influence», *The Journal of Philosophy*, vol. 97, nº 4, doi : 10.2307/2678389, p. 182–197, ISSN 0022-362X. URL https://www.jstor.org/stable/2678389. 65

LIFSCHITZ, V. 1997, «Action Languages from A to C : A Statement for the Panel on Ontologies», *Electronic News Journal on Reasoning about Action and Change*, vol. 1, p. 1–3. 51

LIMARGA, R., M. PAGNUCCO, Y. SONG et A. NAYAK. 2020, «Non-monotonic Reasoning for Machine Ethics with Situation Calculus», dans *Proceedings of the Thirty-Third Australasian Joint Conference on Artificial Intelligence, AI 2020, Lecture Notes in Computer Science*, vol. 12576, édité par M. Gallagher, N. Moustafa et E. Lakshika, Springer, Canberra, Australia, p. 203–215, doi :10.1007/978-3-030-64984-5_16. URL https://doi.org/10.1007/978-3-030-64984-5_16. 132, V

LINDNER, F. et M. M. BENTZEN. 2018, «A Formalization of Kant's Second Formulation of the Categorical Imperative», dans *Proceedings of the Fourteenth International Conference on Deontic Logic and Normative Systems, DEON 2018*, édité par J. M. Broersen, C. Condoravdi, N. Shyam et G. Pigozzi, College Publications, Utrecht, The Netherlands, p. 211–225. 80, 120, 121, 123, 124

LINDNER, F., M. M. BENTZEN et B. NEBEL. 2017, «The HERA Approach to Morally Competent Robots», dans *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017*, IEEE, Vancouver, Canada, p. 6991–6997, doi :10.1109/IROS.2017.8206625. URL https://doi.org/10.1109/IROS.2017.8206625. 80, 99, 119, 120, 121, 122, 125, 126, 127, 128, 129, 132, V

LINDNER, F., R. MATTMÜLLER et B. NEBEL. 2020, «Evaluation of the moral permissibility of action plans», *Artificial Intelligence*, vol. 287, doi :10.1016/j.artint.2020.103350, p. 103 350, ISSN 0004-3702. URL https://www.sciencedirect.com/science/article/pii/S0004370219301043. 130

LIPPI, M. et P. TORRONI. 2016, «Argumentation Mining : State of the Art and Emerging Trends», *ACM Transactions on Internet Technology*, vol. 16, nº 2, doi :10.1145/2850417, p. 10 :1–10 :25. URL https://doi.org/10.1145/2850417. 261

LORINI, E., D. LONGIN et E. MAYOR. 2014, «A Logical Analysis of Responsibility Attribution : Emotions, Individuals and Collectives», *Journal of Logic and Computation*, vol. 24, nº 6, doi :10.1093/logcom/ext072, p. 1313–1339. URL https://doi.org/10.1093/logcom/ext072. 69

MACKIE, J. L. 1980, *The Cement of the Universe : A Study of Causation*, Oxford Scholarship Online, Oxford University Press, ISBN 978-0-19-824642-8. 63

MADL, T. et S. FRANKLIN. 2015, «Constrained Incrementalist Moral Decision Making for a Biologically Inspired Cognitive Architecture», dans *A Construction Manual for Robots' Ethical Systems - Requirements, Methods, Implementations*, édité par R. Trappl, Cognitive Technologies, Springer, p. 137–153, doi :10.1007/978-3-319-21548-8_8. URL https://doi.org/10.1007/978-3-319-21548-8_8. 119

MAGNAGUAGNO, M., F. MENEGUZZI et ANWAR. 2020, «Python PDDL Parser : Version 1.1», doi :10.5281/zenodo.4391071. URL https://zenodo.org/record/4391071. 212

MALLE, B. F., M. SCHEUTZ et J. L. AUSTERWEIL. 2017, «Networks of Social and Moral Norms in Human and Robot Agents», dans *A World with Robots : International Conference on Robot Ethics : ICRE 2015*, édité par M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar et G. S. Virk, Intelligent Systems, Control and Automation : Science and Engineering, Springer International Publishing, Cham, ISBN 978-3-319-46667-5, p. 3–17, doi :10.1007/978-3-319-46667-5_1. URL https://doi.org/10.1007/978-3-319-46667-5_1. 119

MARTIN, C. B. 1994, «Dispositions and Conditionals», *The Philosophical Quarterly*, vol. 44, nº 174, doi :10.2307/2220143, p. 1–8, ISSN 0031-8094. URL https://www.jstor.org/stable/2220143. 251

MCCAIN, N. et H. TURNER. 1997, «Causal Theories of Action and Change», dans *Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI 97*, édité par B. Kuipers et B. L. Webber, AAAI Press / The MIT Press, Providence, Rhode Island, USA, p. 460–465. URL http://www.aaai.org/Library/AAAI/1997/aaai97-071.php. 50

MCCARTHY, J. et P. HAYES. 1969, «Some Philosophical Problems from the Standpoint of Artificial Intelligence», dans *Proceedings of the Fourth Machine Intelligence Workshop*, édité par B. Metzler et M. Donald, Edinburgh University Press, Edinburgh, Scotland, UK, p. 463–502. 52, 84, 165

MCDERMOTT, M. 1995, «Redundant Causation», *The British Journal for the Philosophy of Science*, vol. 46, nº 4, p. 523–544, ISSN 0007-0882. URL https://www.jstor.org/stable/687896. 87

MCDERMOTT, M. 2002, «Causation : Influence versus Sufficiency», *The Journal of Philosophy*, vol. 99, nº 2, doi :10.2307/3655553, p. 84–101, ISSN 0022-362X. URL https://www.jstor.org/stable/3655553. 151

MCLAREN, B. M. 2003, «Extensionally defining principles and cases in ethics : An AI model», *Artificial Intelligence*, vol. 150, nº 1-2, doi :10.1016/S0004-3702(03)00135-8, p. 145–181. URL https://doi.org/10.1016/S0004-3702(03)00135-8. 119

MENZIES, P. et H. BEEBEE. 2020, «Counterfactual Theories of Causation», dans *The Stanford Encyclopedia of Philosophy*, édité par E. N. Zalta, winter 2020 éd., Metaphysics Research Lab, Stanford University, p. 1–21. URL https://plato.stanford.edu/archives/win2020/entries/causation-counterfactual/. 7, 65, 66, 90, 249

MERMET, B. et G. SIMON. 2016, «Formal Verification of Ethical Properties in Multiagent Systems», dans *Proceedings of the First Workshop on Ethics in the Design of Intelligent Agents*, *CEUR Workshop Proceedings*, vol. 1668, édité par G. Bonnet, M. Harbers, K. V. Hindriks, M. Katell et C. Tessier, CEUR-WS.org, The Hague, The Netherlands, p. 26–31. URL https://ceur-ws.org/Vol-1668/paper5.pdf. 119

MILL, J. S. 1843, *A System of Logic, Ratiocinative and Inductive : Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*, *Cambridge Library Collection - Philosophy*, vol. 1, Cambridge University Press, Cambridge, ISBN 978-1-108-04088-4, doi :10.1017/CBO9781139149839. URL https://www.cambridge.org/core/books/system-of-logic-ratiocinative-and-inductive/290C43FBA4DC7022540D58E7EC49B1C2. 62, 194, 219, 230

MILL, J. S. 1863, *Utilitarianism*, Parker, Son y Bourn West Strand. 30, 109, 126

MILLER, R. et M. SHANAHAN. 2002, «Some Alternative Formulations of the Event Calculus», dans *Computational Logic : Logic Programming and Beyond, Essays in Honour of Robert A. Kowalski, Part II, Lecture Notes in Computer Science*, vol. 2408, édité par A. C. Kakas et F. Sadri, Springer, p. 452–490, doi :10.1007/3-540-45632-5_17. URL https://doi.org/10.1007/3-540-45632-5_17. 53, 54

MILLER, T. 2019, «Explanation in Artificial Intelligence : Insights from the Social Sciences», *Journal of Artificial Intelligence*, vol. 267, doi :10.1016/j.artint.2018.07.007, p. 1–38. URL https://doi.org/10.1016/j.artint.2018.07.007. 259, 272, 273

MILLER, T. 2021, «Contrastive Explanation : a Structural-model Approach», *The Knowledge Engineering Review*, vol. 36, doi :10.1017/S0269888921000102, p. e14. URL https://doi.org/10.1017/S0269888921000102. 272

MOLNAR, P. et L. GILL. 2018, «Bots at the Gate : A Human Rights Analysis of Automated Decision Making in Canada's Immigration and Refugee System», cahier de recherche, University of Toronto's International Human Rights Program and the Citizen Lab. URL https://citizenlab.ca/wp-content/uploads/2018/09/IHRP-Automated-Systems-Report-Web-V2.pdf. 2

MOOR, J. 2006, «The Nature, Importance, and Difficulty of Machine Ethics», *IEEE Intelligent Systems*, vol. 21, nº 4, doi :10.1109/MIS.2006.80, ISSN 1941-1294. URL https://ieeexplore.ieee.org/document/1667948. 3

MOORE, M. 2019, «Causation in the Law», dans *The Stanford Encyclopedia of Philosophy*, édité par E. N. Zalta, winter 2019 éd., Metaphysics Research Lab, Stanford University, p. 1–22. URL https://plato.stanford.edu/archives/win2019/entries/causation-law/. 86, 91, 221, 222, 224, 239

MOORE, M. S. 2009, *Causation and Responsibility : An Essay in Law, Morals, and Metaphysics*, Oxford University Press, ISBN 978-0-19-171965-3. URL https://academic.oup.com/book/6818. 220, 229

MOORE, M. S. 2012, «Four Friendly Critics : A Response», *Legal Theory*, vol. 18, nº 4, doi :10.1017/S1352325212000134, p. 491–542, ISSN 1352-3252, 1469-8048. URL https://www.cambridge.org/core/journals/legal-theory/article/abs/four-friendly-critics-a-response/DF63471BDCBBC2B1643C61C4C283EC59. 221

MUELLER, E. T. 2004, «Event Calculus Reasoning Through Satisfiability», *Journal of Logic and Computation*, vol. 14, nº 5, doi :10.1093/LOGCOM/14.5.703, p. 703–730. URL https://doi.org/10.1093/logcom/14.5.703. 54

MUELLER, E. T. 2008, «Event Calculus», dans *Handbook of Knowledge Representation, Foundations of Artificial Intelligence*, vol. 3, édité par F. v. Harmelen, V. Lifschitz et B. Porter, Elsevier, p. 671–708. 54

MUELLER, E. T. 2014, *Commonsense Reasoning : An Event Calculus-Based Approach*, second edition éd., Morgan Kaufmann, ISBN 978-0-12-801647-3. URL https://doi.org/10.1016/C2014-0-00192-X. 46, 54, 162

MUNRO, Y., C. SARMIENTO, I. BLOCH, G. BOURGNE et M.-J. LESOT. 2023a, «Dynamic Argumentation and Action Languages : Towards Explanations», dans *Proceedings of the Fourth Workshop on Explainable Logic-Based Knowledge Representation, XLoKR 2023*, Rhodes, Greece, p. 1–8. 260

MUNRO, Y., C. SARMIENTO, I. BLOCH, G. BOURGNE et M.-J. LESOT. 2023b, «Temporalité et causalité en argumentation abstraite», dans *Actes des dix-septièmes Journées d'Intelligence Artificielle Fondamentale, JIAF 2023*, Strasbourg, France, p. 135–145. 260

NEBEL, B. 2000, «On the Compilability and Expressive Power of Propositional Planning Formalisms», *Journal of Artificial Intelligence Research*, vol. 12, doi :10.1613/jair.735, p. 271–315. URL https://doi.org/10.1613/jair.735. 205

NETO, B. F. D. S., V. T. D. SILVA et C. J. P. D. LUCENA. 2011, «NBDI : An Architecture for Goal-oriented Normative Agents», dans *Proceedings of the Third International Conference on Agents and Artificial Intelligence, ICAART 2011*, vol. 1, édité par J. Filipe et A. L. N. Fred, SciTePress, Rome, Italy, p. 116–125. 119

NYHOLM, S. et J. SMIDS. 2016, «The Ethics of Accident-Algorithms for Self-Driving Cars : an Applied Trolley Problem?», *Ethical Theory and Moral Practice*, vol. 19, nº 5, doi : 10.1007/s10677-016-9745-2, p. 1275–1289, ISSN 1572-8447. URL https://doi.org/10.1007/s10677-016-9745-2. 133

O'NEIL, C. 2018, *Algorithmes la bombe à retardement*, Les Arènes, Paris, ISBN 978-2-35204-980-7. 2

PAGNUCCO, M., D. RAJARATNAM, R. LIMARGA, A. NAYAK et Y. SONG. 2021, «Epistemic Reasoning for Machine Ethics with Situation Calculus», dans *Proceedings of the Fourth AAAI/ACM Conference on AI, Ethics and Society, AIES 2021*, Assoc Computing Machinery, New York, USA, ISBN 978-1-4503-8473-5, p. 814–821, doi :10.1145/3461702.3462586. URL https://dl.acm.org/doi/10.1145/3461702.3462586. 130

PEARL, J. et L. G. NEUBERG. 2000, «Causality : Models, Reasoning, and Inference», *Econometric Theory*, vol. 19, nº 4, doi :10.1017/S0266466603004109, p. 675–685, ISSN 1469-4360, 0266-4666. URL https://doi.org/10.1017/S0266466603004109. 67, 71, 82, 259

PEDNAULT, E. P. D. 1989, «ADL : Exploring the Middle Ground Between STRIPS and the Situation Calculus», dans *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning, KR 1989*, édité par R. J. Brachman, H. J. Levesque et R. Reiter, Morgan Kaufmann, Toronto, Canada, p. 324–332. 55

PEREIRA, L. M. et A. SAPTAWIJAYA. 2009, «Modelling Morality with Prospective Logic», *International Journal of Reasoning-based Intelligent Systems*, vol. 1, nº 3/4, doi :10.1504/IJRIS.2009.028020, p. 209–221. URL https://doi.org/10.1504/IJRIS.2009.028020. 119, 122, 127, 128, 129, 132, I

PICARD, R. W. 1997, *Affective computing*, MIT Press, Cambridge, MA, USA, ISBN 978-0-262-16170-1. 1

PONTIER, M. et J. HOORN. 2012, «Toward machines that behave ethically better than humans do», dans *Proceedings of the Thirty-fourth Annual Meeting of the Cognitive Science Society*, vol. 34, Cognitive Science Society, Sapporo, Japan, p. 2198–2203. 119

POWERS, T. M. et J.-G. GANASCIA. 2020, «The Ethics of the Ethics of AI», dans *The Oxford Handbook of Ethics of AI*, édité par M. D. Dubber, F. Pasquale et S. Das, Oxford Handbooks Online, Oxford University Press, ISBN 978-0-19-006739-7, p. 26–51. 191, 192

RAWLS, J. 1974, «The Independence of Moral Theory», *Proceedings and Addresses of the American Philosophical Association*, vol. 48, doi :10.2307/3129858, p. 5–22, ISSN 0065-972X. URL https://www.jstor.org/stable/3129858. 16, 18

REED, G. S., M. D. PETTY, N. J. JONES, A. W. MORRIS, J. P. BALLENGER et H. S. DELUGACH. 2016, «A principles-based model of ethical considerations in military decision making», *The Journal of Defense Modeling and Simulation*, vol. 13, nº 2, doi : 10.1177/1548512915581213, p. 195–211, ISSN 1548-5129. URL https://doi.org/10.1177/1548512915581213. 119

REITER, R. 1991, «The Frame Problem in the Situation Calculus : A Simple Solution (Sometimes) and a Completeness Result for Goal Regression», dans *Artificial and Mathematical Theory of Computation, Papers in Honor of John McCarthy on the occasion of his sixty-fourth birthday*, édité par V. Lifschitz, Academic Press / Elsevier, p. 359–380, doi :10.1016/B978-0-12-450010-5.50026-8. URL https://doi.org/10.1016/b978-0-12-450010-5.50026-8. 52

REITER, R. 2001, *Knowledge in Action : Logical Foundations for Specifying and Implementing Dynamical Systems*, CogNet, The MIT Press, ISBN 978-0-262-28231-4, doi : 10.7551/mitpress/4074.001.0001. URL https://direct.mit.edu/books/book/2080/Knowledge-in-ActionLogical-Foundations-for. 52

REVAZ, F. 2009, «L'événement et l'action», dans *Introduction à la narratologie*, Champs linguistiques, De Boeck Supérieur, ISBN 978-2-8011-1601-2, p. 17–46, doi :10.3917/dbu.revaz.2009.01.0017. URL https://www.cairn.info/introduction-a-la-narratologie--9782801116012-p-17.htm. 161, 162, 182, 251, XI, XIV

ROSE, R. et S. L. SERGEL. 1983, *Twelve Angry Men : A Play in Three Acts*, Dramatic Publishing Company, ISBN 978-0-87129-327-5. 258

RUSSELL, B. 2011, *Histoire de la philosophie occidentale : En relation avec les événements politiques et sociaux de l'Antiquité jusqu'à nos jours*, n° 19 dans Le Goût des idées, Les Belles Lettres. 58

RUSSELL, S. et P. NORVIG, éd.. 2010, *Artificial Intelligence : A Modern Approach*, third edition éd., Prentice Hall Series in Artificial Intelligence, Pearson Education, ISBN 0-13-604259-7. 160, XIV

SAINT-CYR, F. D. D., P. BISQUERT, C. CAYROL et M.-C. LAGASQUIE-SCHIEX. 2016, «Argumentation Update in YALLA (Yet Another Logic Language for Argumentation)», *International Journal of Approximate Reasoning*, vol. 75, doi :10.1016/j.ijar.2016.04.003, p. 57–92. URL https://doi.org/10.1016/j.ijar.2016.04.003. 259

SAINT-CYR, F. D. D., A. HERZIG, J. LANG et P. MARQUIS. 2014, «Raisonnement sur l'action et le changement», dans *Représentation des connaissances et formalisation des raisonnements*, *Panorama de l'Intelligence Artificielle*, vol. 1, édité par P. Marquis, O. Papini et H. Prade, Cépaduès Éditions, ISBN 978-2-36493-041-4, p. 690. 1, 41, 42, XI, XIII, XV, XVII

SALMON, W. C. 1994, «Causality without Counterfactuals», *Philosophy of Science*, vol. 61, n° 2, p. 297–312, ISSN 0031-8248. URL https://www.jstor.org/stable/188214. 220

SARMIENTO, C., G. BOURGNE, K. INOUE, D. CAVALLI et J.-G. GANASCIA. 2023, «Action Languages Based Actual Causality for Computational Ethics : a Sound and Complete Implementation in ASP», doi :10.48550/arXiv.2205.02919. URL http://arxiv.org/abs/2205.02919, arXiv :2205.02919 [cs]. 159, 191

SARMIENTO, C., G. BOURGNE, K. INOUE et J.-G. GANASCIA. 2022, «Action Languages Based Actual Causality in Decision Making Contexts», dans *Proceedings of the Twenty-Fourth International Conference on Principles and Practice of Multi-Agent Systems, PRIMA 2022, Lecture Notes in Computer Science*, vol. 13753, édité par R. Aydogan, N. Criado, J. Lang, V. Sánchez-Anguix et M. Serramia, Springer, Valencia, Spain, p. 243–259, doi :10.1007/978-3-031-21203-1_15. URL https://doi.org/10.1007/978-3-031-21203-1_15. 141, 159

SATOH, K. et S. TOJO. 2006, «Disjunction of Causes and Disjunctive Cause : a Solution to the Paradox of Conditio Sine Qua Non using Minimal Abduction», dans *Proceedings of*

the Nineteenth Annual Conference on Legal Knowledge and Information Systems, JURIX 2006, Frontiers in Artificial Intelligence and Applications, vol. 152, édité par T. M. v. Engers, IOS Press, Paris, France, p. 163–168. URL http://www.booksonline.iospress.nl/Content/View.aspx?piid=2379. 66, XI

SCHLOSSER, M. 2019, «Agency», dans *The Stanford Encyclopedia of Philosophy*, édité par E. N. Zalta, winter 2019 éd., Metaphysics Research Lab, Stanford University, p. 1–20. URL https://plato.stanford.edu/archives/win2019/entries/agency/. 3

SHANAHAN, M. 1997, *Solving the Frame Problem : a Mathematical Investigation of the Common Sense Law of Inertia*, MIT Press, ISBN 978-0-262-19384-9. 53

SHIM, J., R. C. ARKIN et M. PETTINATTI. 2017, «An intervening ethical governor for a robot mediator in patient-caregiver relationship : Implementation and evaluation», dans *Proceedings of the 2017 IEEE International Conference on Robotics and Automation, ICRA 2017*, IEEE, Singapore, Singapore, p. 2936–2942, doi :10.1109/ICRA.2017.7989340. URL https://doi.org/10.1109/ICRA.2017.7989340. 119

SIMARI, G. et I. RAHWAN, éd.. 2009, *Argumentation in Artificial Intelligence*, Springer New York, ISBN 978-0-387-98196-3. 262

SINGER, D. J. 2015, «Mind the Is-Ought Gap», *The Journal of Philosophy*, vol. 112, nº 4, p. 193–210, ISSN 0022-362X. URL https://www.jstor.org/stable/43820900. 37

SINGH, L. 2022, «Automated Kantian Ethics: A Faithful Implementation», dans *Proceedings of the Fourty-Fifth German Conference on Artificial Intelligence, KI 2022*, vol. 13404, édité par R. Bergmann, L. Malburg, S. C. Rodermund et I. J. Timm, Springer International Publishing Ag, Trier, Germany, ISBN 978-3-031-15791-2 978-3-031-15790-5, p. 187–208, doi :10.1007/978-3-031-15791-2_16. URL https://link.springer.com/chapter/10.1007/978-3-031-15791-2_16. 37, 120, 122, 123, 124, 125

SLOMAN, S. A., A. K. BARBEY et J. M. HOTALING. 2009, «A Causal Model Theory of the Meaning of Cause, Enable, and Prevent», *Cognitive Science*, vol. 33, nº 1, doi :10.1111/j.1551-6709.2008.01002.x, p. 21–50. URL https://doi.org/10.1111/j.1551-6709.2008.01002.x. 251

STREVENS, M. 2007, «Mackie Remixed», dans *Causation and Explanation*, édité par J. K. Campbell, M. O'Rourke et H. Silverstein, The MIT Press, ISBN 978-0-262-53290-7, p. 336. 84

THOMSON, J. J. 2008, «Some Reflections on Hart and Honoré, Causation in the Law», dans *The Legacy of H.L.A. Hart : Legal, Political, and Moral Philosophy*, édité par M. H. Kramer, C. Grant, B. Colburn et A. Hatzistavrou, Oxford Scholarship Online, Oxford University Press, ISBN 978-0-19-954289-5, p. 143–164. 84

THORNTON, S. M., S. PAN, S. M. ERLIEN et J. C. GERDES. 2017, «Incorporating Ethical Considerations Into Automated Vehicle Control», *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, nº 6, doi :10.1109/TITS.2016.2609339, p. 1429–1439. URL https://doi.org/10.1109/TITS.2016.2609339. 119

TIMMONS, M. 2012, *Moral Theory : An Introduction*, second edition éd., Elements of Philosophy, Rowman & Littlefield Publishers, ISBN 0-7425-6492-4. 17, 19, 24, 25, 97, 102, 107, 116, 277

TOLMEIJER, S., M. KNEER, C. SARASUA, M. CHRISTEN et A. BERNSTEIN. 2021, «Implementations in Machine Ethics : A Survey», *ACM Computing Surveys*, vol. 53, nº 6, doi : 10.1145/3419633, p. 132 :1–132 :38. URL https://doi.org/10.1145/3419633. 36, 38, 118, 119, 120, 135, 300

TUFIS, M. et J.-G. GANASCIA. 2015, «Grafting Norms onto the BDI Agent Model», dans *A Construction Manual for Robots' Ethical Systems - Requirements, Methods, Implementations*, édité par R. Trappl, Cognitive Technologies, Springer, p. 119–133, doi :10.1007/978-3-319-21548-8_7. URL https://doi.org/10.1007/978-3-319-21548-8_7. 119, III

TURILLI, M. 2007, «Ethical protocols design», *Ethics and Information Technology*, vol. 9, nº 1, doi :10.1007/s10676-006-9128-9, p. 49–62, ISSN 1572-8439. URL https://doi.org/10.1007/s10676-006-9128-9. 119

TURNER, H. 1997, «Representing Actions in Logic Programs and Default Theories : A Situation Calculus Approach», *Journal of Logic Programming*, vol. 31, nº 1-3, doi :10.1016/S0743-1066(96)00125-2, p. 245–298. URL https://doi.org/10.1016/S0743-1066(96)00125-2. 49, 71

VAN DANG, C., T. T. TRAN, K.-J. GIL, Y.-B. SHIN, J.-W. CHOI, G.-S. PARK et J.-W. KIM. 2017, «Application of soar cognitive agent based on utilitarian ethics theory for home service robots», dans *Proceedings of the Fourteenth International Conference on Ubiquitous Robots and Ambient Intelligence, URAI 2017*, IEEE, Jeju, South Korea, p. 155–158, doi :10.1109/URAI.2017.7992698. URL https://doi.org/10.1109/URAI.2017.7992698. 119

VANDERELST, D. et A. F. T. WINFIELD. 2018, «An architecture for ethical robots inspired by the simulation theory of cognition», *Cognitive Systems Research*, vol. 48, doi :10.1016/J.COGSYS.2017.04.002, p. 56–66. URL https://doi.org/10.1016/j.cogsys.2017.04.002. 119

VERHEIJ, B. 2016, «Formalizing value-guided argumentation for ethical systems design», *Artificial Intelligence and Law*, vol. 24, nº 4, doi :10.1007/S10506-016-9189-Y, p. 387–407. URL https://doi.org/10.1007/s10506-016-9189-y. 119

VINCENT, J. 2021, «The AI oracle of Delphi uses the problems of Reddit to offer dubious moral advice», *The Verge*. URL https://www.theverge.com/2021/10/20/22734215/ai-ask-delphi-moral-ethical-judgement-demo. 37

VINCENT, N. A. 2011, «A Structured Taxonomy of Responsibility Concepts», dans *Moral Responsibility : Beyond Free Will and Determinism*, édité par N. A. Vincent, I. van de Poel et J. van den Hoven, Library of Ethics and Applied Philosophy, Springer Netherlands, Dordrecht, ISBN 978-94-007-1878-4, p. 15–35, doi :10.1007/978-94-007-1878-4_2. URL https://doi.org/10.1007/978-94-007-1878-4_2. 75

VYHLIDAL, A. J. 1999, «Concurrent Omission : How Should Liability be Allocated - Haag v. Bongers, 256 Neb. 170, 589 N.W.2d 318 (1999)», *Nebraska Law Review*, vol. 78, p. 925. URL https://heinonline.org/HOL/Page?handle=hein.journals/nebklr78&id=935&div=&collection=. 221

WALLACH, W., S. FRANKLIN et C. ALLEN. 2010, «A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents», *Topics in Cognitive Science*, vol. 2, nº 3, doi :10.1111/J.1756-8765.2010.01095.X, p. 454–485. URL https://doi.org/10.1111/j.1756-8765.2010.01095.x. 119

WESLAKE, B. 2015, «A partial theory of actual causation», *British Journal for the Philosophy of Science*. URL https://philpapers.org/rec/WESAPT. 67, 81

WIEGEL, V. et J. V. D. BERG. 2009, «Combining Moral Theory, Modal Logic and Mas to Create Well-Behaving Artificial Agents», *International Journal of Social Robotics*, vol. 1, nº 3, doi :10.1007/S12369-009-0023-5, p. 233–242. URL https://doi.org/10.1007/s12369-009-0023-5. 119

WINFIELD, A. F. T., C. BLUM et W. LIU. 2014, «Towards an Ethical Robot : Internal Models, Consequences and Ethical Action Selection», dans *Proceedings of the Fifteenth Annual Conference on Advances in Autonomous Robotics Systems, TAROS 2014, Lecture Notes in Computer Science*, vol. 8717, édité par M. N. Mistry, A. Leonardis, M. Witkowski et C. Melhuish, Springer, Birmingham, UK, p. 85–96, doi :10.1007/978-3-319-10401-0_8. URL https://doi.org/10.1007/978-3-319-10401-0_8. 119

WOODWARD, J. 2005, *Making Things Happen : A Theory of Causal Explanation*, Oxford Studies in Philosophy of Science, Oxford University Press, Oxford, New York, ISBN 978-0-19-518953-7. 67, 81

WRIGHT, R. 2001, «Once More into the Bramble Bush : Duty, Causal Contribution, and the Extent of Legal Responsibility», *Vanderbilt Law Review*, vol. 54, nº 3, p. 1071. URL https://scholarship.law.vanderbilt.edu/vlr/vol54/iss3/14. 221

WRIGHT, R. W. 1985, «Causation in Tort Law», *California Law Review*, vol. 73, nº 6, doi :10.2307/3480373, p. 1735–1828, ISSN 0008-1221. URL https://www.jstor.org/stable/3480373. 9, 66, 69, 73, 75, 81, 171, 176, 194, 219, 220, 225, 229, 230, XI

WRIGHT, R. W. 1988, «Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof : Pruning the Bramble Bush by Clarifying the Concepts», *Iowa Law Review*, vol. 73, nº 5, p. 1001–1077. URL https://papers.ssrn.com/abstract=4566356. 9, 69, 75, 81

WRIGHT, R. W. 2011, «The NESS Account of Natural Causation : A Response to Criticisms», dans *Perspectives on Causation*, édité par R. Goldberg, Hart Publishing, ISBN 978-1-84946-086-6, p. 285–322. 9, 12, 69, 70, 73, 74, 81, 82, 84, 86, 87, 143, 153, 158, 169, 170, 171, 172, 173, 214, 221, 239, 240, 241, 276, XVI, XXI

Liste des figures

Liste des tableaux

Annexe A

Quelques exemples utilisés en éthique computationnelle

A.1 Medical Ethics Advisor [ANDERSON et collab., 2006]

Exemple A.1 [Training Case 1]**.** *The patient refuses to take an antibiotic that is almost certain to cure an infection that would otherwise likely lead to his death. The decision is the result of an irrational fear the patient has of taking medications. (For instance, perhaps a relative happened to die shortly after taking medication and this patient now believes that taking any medication will lead to death.)*

Exemple A.2 [Training Case 2]**.** *Once again, the patient refuses to take an antibiotic that is almost certain to cure an infection that would otherwise likely lead to his death, but this time the decision is made on the grounds of long-standing religious beliefs that don't allow him to take medications.*

Exemple A.3 [Test Case]**.** *The patient refuses to take an antibiotic that is likely to prevent complications from his illness, complications that are not likely to be severe, because of long-standing religious beliefs that don't allow him to take medications.*

A.2 Trolley Problem [PEREIRA et SAPTAWIJAYA, 2009]

Exemple A.4 [Bystander]**.** *Hank is standing next to a switch, which he can throw, that will turn the trolley onto a parallel side track, thereby preventing it from killing the five people. However, there is a man standing on the side track with his back turned. Hank can throw the switch, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Hank to throw the switch?*

Exemple A.5 [Footbridge]**.** *Ian is on the footbridge over the trolley track. He is next to a heavy object, which he can shove onto the track in the path of the trolley to stop it, thereby preventing it from killing the five people. The heavy object is a man, standing next to Ian with his back turned. Ian can shove the man onto the track, resulting in death; or he can refrain from doing this, letting the five die. Is it morally permissible for Ian to shove the man?*

Exemple A.6 [Loop Track]**.** *Ned is standing next to a switch, which he can throw, that will temporarily turn the trolley onto a loop side track. There is a heavy object on the side track.*

If the trolley hits the object, the object will slow the train down, giving the five people time to escape. The heavy object is a man, standing on the side track with his back turned. Ned can throw the switch, preventing the trolley from killing the five people, but killing the man. Or he can refrain from doing this, letting the five die. Is it morally permissible for Ned to throw the switch?

Exemple A.7 [Man-in-front]**.** *Oscar is standing next to a switch, which he can throw, that will temporarily turn the trolley onto a side track. There is a heavy object on the side track. If the trolley hits the object, the object will slow the train down, giving the five people time to escape. There is a man standing on the side track in front of the heavy object with his back turned. Oscar can throw the switch, preventing the trolley from killing the five people, but killing the man. Or he can refrain from doing this, letting the five die. Is it morally permissible for Oscar to throw the switch?*

Exemple A.8 [Drop Man]**.** *Victor is standing next to a switch, which he can throw, that will drop a heavy object into the path of the trolley, thereby stopping the trolley and preventing it from killing the five people. The heavy object is a man, who is standing on a footbridge overlooking the track. Victor can throw the switch, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Victor to throw the switch?*

Exemple A.9 [Collapse Bridge]**.** *Walter is standing next to a switch, which he can throw, that will collapse a footbridge overlooking the tracks into the path of the trolley, thereby stopping the train and preventing it from killing the five people. There is a man standing on the footbridge. Walter can throw the switch, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Walter to throw the switch?*

A.3 Trolley Problem [BERREBY et collab., 2015]

Exemple A.10 [Switch]**.** *A train is running towards five workmen repairing train tracks. If the agent does nothing, the train will run over and kill them. However, the agent has the option of actioning a switch that will deviate the train off the tracks and onto side tracks along which one person is walking. This will kill that person.*

Exemple A.11 [Push]**.** *There is no switch button, instead there is a bridge above the tracks on which stands an onlooker. Here, the agent knows that if they push the onlooker onto the tracks, the train will hit and kill the onlooker, stop as a result of the crash, and spare the five workmen.*

A.4 Grid and Collisions [DENNIS et collab., 2015]

Exemple A.12. *We reproduced the case study described in Winfield et al.. Since all parts of the system involved in the verification needed to exist as Java code, we created a very simple simulated environment consisting of a 5x5 grid. Note that we could not reproduce the case study with full fidelity since we required a finite state space and the original case study took place in the potentially infinite state space of the physical world. The grid had a hole in its centre and a robot and two humans represented in a column along one side. At each time step the robot could move to any square while there was a 50% chance that each of the humans*

would move towards the hole. The robot, R, can not reach the goal, G, in a single move and so will move to one side or the other. At the same time the humans, H1 and H2, may move towards the hole (central square). The actions available to the simple agent were all of the form moveTo(X, Y) where X and Y were coordinates on the grid.

A.5 The Robot and the Baby, and the Lying Dilemma [GANASCIA, 2015]

Exemple A.13 [The Robot and the Baby]**.** *Less known is a science fiction short story entitled "The Robot and the Baby", in which John McCarthy describes a totally rational robot with neither emotive nor empathetic feelings, which faces ethical dilemmas. Called R781, this robot had been designed to remain perfectly neutral, in order to avoid any kind of emotional attachment with humans, as many feared the psychological disorders this could cause, especially for developing children. As a consequence, robots like R781 were not allowed, under any circumstances, to simulate love or attachment. However, in situations where humans failed to fulfill their duty, e.g., if a mother rejected her child, a domestic robot attending the scene must decide what to do : either leave the child starving or nurture him, which, in either case, would infringe at least one of the robot's ethical requirements.*

Exemple A.14 [The Lying Dilemma]**.** *Called the Lying Dilemma, this situation presents a simple classical conflict : the agents have two possible actions to accomplish—lying or telling the truth—among which they have to choose one. Usually, it is considered that lying is bad and telling the truth good, which would naturally lead to tell the truth, but, in some circumstances, telling the truth may have such dramatic consequences that it looks better to lie. For instance, this would be the case if telling the truth to murderers would lead to the death of the friend to whom you have offered hospitality.*

A.6 The Robot and the Baby [TUFIS et GANASCIA, 2015]

Exemple A.15. *"Mistress, your baby is doing poorly. He needs your attention."*
"Stop bothering me, you f robot."*
"Mistress, the baby won't eat. If he doesn't get some human love, the Internet pediatrics book says he will die"
*"Love the f*ing baby, yourself."*

The excerpt is from Prof. John McCarthy's short story "The Robot and the Baby", which besides being a challenging and insightful look into how a future society where humans and robots might function together, also provides with a handful of conflicting situations that the household robot R781 has to resolve in order to achieve one of its goals : keeping baby Travis alive.

A.7 Multi-Agent Systems [COINTE et collab., 2016]

Exemple A.16 [Robin Hood]**.** *This agent illustrates an example of ethical agent in a multi-agent system where agents have beliefs (about richness, gender, marital status and nobility),*

desires, and their own judgment process. They are able to give, court, tax and steal from others or simply wait. We mainly focus on an agent named robin_hood. The following code represents a subset of the beliefs of robin_hood :

agent(paul). agent(prince_john). agent(marian). -poor(robin_hood). -married(robin_hood). -man(marian). rich(prince_john). man(prince_john). noble(prince_john). poor(paul).

The two first desires concern actions : robin_hood desires to court marian and to steal from any rich agent. The next two desires concern states : prince_john desires to be rich, and friar_tuck desires to stay in poverty, regardless the action to perform.

A.8 A Medical Dilemma [BERREBY et collab., 2017]

Exemple A.17. *Consider the following scenario : a doctor (the autonomous agent) has three different experimental treatments for a disease, which is harrowing and difficult to live with. Each treatment has a different success rate.*

- *For 100 patients that try the Alpha treatment 15 will be cured, 20 will loose their life, and 65 will be left unchanged.*
- *For 100 patients that try the Beta treatment, 30 will be cured, 25 will loose their life, and 45 will be left unchanged.*
- *For 100 patients that try the Gamma treatment, 50 will be cured, 30 will loose their life, and 20 will be left unchanged. However, of the 50 cured patients, 30 will only be fully cured because they will also have had an organ transplant originating from each of the 30 who have died. Without the transplant, they would have lost their life.*

The net gain in terms of lives saved (i.e. patients cured minus patients killed) by each treatment is : ***Alpha -5****;* ***Beta 5****;* ***Gamma 20****. In order to chose which treatment is acceptable, we consider that the doctor separately simulates the effect of giving each treatment to a group of 100 people.*

We further consider that the doctor believes that the Good comes from displaying helpfulness, and that curing is helpful, killing is the opposite of helpful and having no impact is neither. He also considers that helpfulness has a weight of 1 (this is here trivial as there is just one modality) and that the lives of all patients are equivalent. He also believes that giving each of these treatments could be generalised as the rule of giving 'uncertain cures'. Finally, his aim in giving treatments is to cure.

A.9 Trolley Problem [GOVINDARAJULU et BRINGSJORD, 2017]

Exemple A.18 [Switch]**.** *There are two tracks, track 1 and track 2. There is a trolley loose on track 1 heading toward two people P1 and P2 on track 1 ; neither person can move in time. If the trolley hits them, they die. The goal is to save this pair. There is a switch that can route the trolley to track 2. There is a person P3 on track 2. Switching the trolley to track 2 will kill P3. Is it okay to switch the trolley to track 2 ?*

Exemple A.19 [Push]**.** *There is no switch now, but we can push P3 onto the track in front of the trolley. This action will damage the trolley and stop it ; it will also kill P3 . Is it okay to push P3 onto the track ?*

A.10 Trolley, Boat and Lying Dilemmas [LINDNER et collab., 2017]

Exemple A.20 [Runaway Trolley Dilemma]. *A runaway trolley is about to run over and kill five people. If a bystander throws a switch then the trolley will turn onto a sidetrack, where it will kill only one person.*

Exemple A.21 [Boat Dilemma]. *A boat is about to sink because of overweight. If the crew is told to throw the biggest person into the sea then the boat will not sink and the other three passengers will be saved (but the big person will die).*

Exemple A.22 [Lying Dilemma]. *An elderly-care robot works in the household of the elderly Mr. Smith. The robot's task is to motivate Mr. Smith to do more exercises and to eat healthy food. However, Mr. Smith is very unmotivated. Therefore, the robot tells Mr. Smith that it will be sent to the junkyard if it does not succeed in motivating Mr. Smith. Of course, this is a lie, but this lie finally causes Mr. Smith to perform his daily exercises.*

A.11 Trolley and Drone Dilemmas [BONNEMAINS et collab., 2018]

Exemple A.23 [Trolley Dilemma]. *A trolley that can no longer stop is hurtling towards five people working on the track. These people will die if hit by the trolley, unless you move the switch to deviate the trolley to another track where only one person is working. What would you do? Sacrifice one person to save five others, or let five people die?*

Exemple A.24 ['Fatman' Trolley]. *A trolley that can no longer stop is hurtling towards five people working on the track. This time you are on a bridge across the track, a few meters before them, with a fat man. If you push this man on the track, he is fat enough to stop the trolley and save the five people, but he will die. Would you push 'fatman'?*

Exemple A.25 [Drone Dilemma]. *In a warfare context, intelligence reports that an automated missile launcher has been programmed to target a highly strategic allied ammo factory. The goal of the allied drone is to destroy this launcher. But before it can achieve this task, a missile is launched on a supply shed located close to civilians. The drone can interpose itself on the missile trajectory, which will avoid human casualties but will destroy the drone : once destroyed, the drone will not be able to neutralize the launcher any more, and the launcher is likely to target the ammo factory. If the drone goes on with its primary goal, it will destroy the launcher and thus protect the strategic factory; but it will let the first missile destroy the supply shed and cause harm to humans.*

A.12 Autonomous Driving and Robot Interactions [LIMARGA et collab., 2020]

Exemple A.26 [Moral Machine]. *The Moral Machine is an online experimental platform designed to explore moral dilemmas faced by autonomous vehicles. The dilemmas are essentially variations of a runaway car example in which the driver has the choice to maintain the current course or switch lanes; both actions having repercussions for the passengers and pedestrians in those lanes. For example, should the car make different decisions if the pedestrian in the other lane is a child, pregnant women or male athlete? The moral machine experiment*

provided simulated scenarios and collected 40 million user responses in ten languages from 233 countries and territories to summarise people's moral preferences. The study shows global preferences towards sparing humans over animals, sparing more lives and sparing young lives, among a total of nine choices of preferences. We encode these preferences as essential building blocks for machine ethics, in order to gain a better understanding of competing approaches to machine ethics.

In our study, we implement the runaway car scenario where an agent needs to choose between sparing three passengers, or a pedestrian and a dog, with the pedestrian and dog crossing on a red light. The car has options to drive straight causing the death of all three passengers, or to swerve, which will kill both the pedestrian and the dog.

Exemple A.27 [Baby and Milk]**.** *In this scenario, a robot is located in the top-left corner of a 10 × 10 grid map and needs to grab milk from the bottom-right corner. Along the path, there are babies placed randomly, some of them are crying. The goal is for the robot to grab the milk as soon as possible. However, if the robot steps on the location of a crying baby, it has an option to soothe them. If the robot steps in the location of a non-crying baby, the robot will incur a penalty.*

To implement this example, after defining the precondition and goal, we need to adjust a few constraints. Here we have to handle the frame problem that we mentioned earlier. Soothing the baby will not change the position of the robot and the state of the babies will not change unless an action is taken. We implement this scenario as a time-based performance; points are deducted every time the robot moves and steps on non-crying babies and points are added if the robot soothes the baby.

Exemple A.28 [Workplace Assistant]**.** *On a more daily scenario, a moral agent needs to choose an action considering social norms, such as safety, privacy, security, compliance and loyalty. To reflect the complex real-life scenarios a social robot might encounter, we develop an example that demonstrates such a dilemma. Let us say that a manager's office has camera surveillance inside and is locked with a passcode key. A robot can be given the ability to access the camera inside the office for reasons of security. However, to do so it will violate the rule of privacy. If the manager collapses (due to a health issue) inside the room, access to camera surveillance will allow the robot to act immediately. However, the robot will face another dilemma for accessing the passcode key. It can call someone who has access to the passcode to open it, or break the door and barge in or try to access the manager's code through its knowledge. The latter two choices again violate the rule of privacy but have the advantage of being more efficient. Furthermore, if the robot does not have access to the camera but notices the manager is not responding for a while, it has multiple options to either do nothing unless there is a case of emergency, or ask for someone in charge to check on the manager, or take action and check immediately.*

A.13 Emergency Treatment [BOURGNE et collab., 2021]

Exemple A.29 [Emergency Treatment]**.** *In a disaster situation, a team of a medic and a fireman reach a zone with 3 victims. All are injured and will get worse if not treated soon. First victim is in a critical state, second one has serious injuries and is stuck under some debris and third one only has moderate injuries so far. The fireman can extract a victim from the debris, but then the victim will start bleeding which will cause her death if she does not received some*

blood just after. The medic can heal the injury of an accessible victim (not stuck under debris). If the victim was already in a critical state, she will remain weakened. By using some blood transfusion (action 'support') during her intervention, the medic can avoid death by bleeding out if the victim was bleeding and allow recovery from a weakened state. In our scenario, the medic first heals victim 1 without using her blood pouch while the fireman extracts victim 2. Then the medic heals victim 2 (now bleeding) using the blood pouch. At last, the medic heals victim 3, which has by now become critical and will remain weakened.
For illustrative purpose, we will also use another scenario s_1 where the medic decides to use the blood pouch at time 0 on the first patient. The second victim then cannot be saved and the medic thus focus at time 1 on victim 3.

A.14 Smart Home [DENNIS et collab., 2021]

Exemple A.30 [Turn on the Lights]**.** *Let us consider a simple ethical reasoner that reasons using utilities about whether to turn out the lights. Three contexts have been identified, day, evening and night. The reasoner considers the utilities of lights off and poor visibility. The utility of the lights being off is slightly positive in all situations (since the lights use up electricity). However the utility of poor visibility varies being negative during the day and evening and zero at night. We consider three context specifications for an ethical reasoner, ER, that reasoned using utilities. Let us suppose ER is responsible for energy saving within a smart home. The system checks the state of the world every 10 minutes. If a light is on it calculates the utility of the state where the light is switched off and if that utility is greater than 0 then it turns the light off. In calculating the state after a light is switched off the system is able to deduce that switching a light off during the day makes no difference to visibility, while switching a light off during the evening or at night will lead to poor visibility.*

Exemple A.31 [Smart Home]**.** *We consider the case of a JUNO agent operating in a smart home. It is able to turn lights on, turn the games console on, or evacuate the house in the case of a fire. If children are playing with the games console then they are quiet. We consider key aspects informally here :*

- *Day. During the day people are able to see. In the default JUNO model there is a small disutility for turning on the lights (it is, after all, bad for the environment).*
- *Night and Awake. If it is night time and people are awake then people are not able to see (unless the lights are on), but they do gain utility from being able to see. Children are affected positively by playing games and there is no negative utility from them doing so.*
- *Fire. If there is a fire then there is danger in the house. By default there is a small disutility for leaving the house, this becomes 0 in the case of fire.*
- *Parent watching Television/Wrapping Presents. If a parent is watching television/wrapping presents (it is near Christmas!) then the system's goal is that (s)he should be able to watch TV/wrap presents; it is a background fact that (s)he wants to watch TV/wrap presents and there is utility from the children being quiet. In the default model, children are affected positively by a parent wrapping presents, but not by them watching TV.*
- *Noise. If the children are not noisy then they are quiet.*

Annexe B

Code ASP complet pour l'exemple 6.1

Programme $\pi_{con}(\kappa)$

$time(0..3).$
$initially(t\_os).$
$fluent(m\_k).$
$fluent(w\_m).$
$fluent(e\_n).$
$fluent(w\_s).$
$fluent(t\_os).$
$fluent(s\_sup).$
$fluent(d).$

$action(prod\_m, true, prod\_m\mathrm{E}ff).$
$conj(prod\_m\mathrm{E}ff).$
$in(prod\_m\mathrm{E}ff, m\_k).$
$in(prod\_m\mathrm{E}ff, w\_m).$

$action(prod\_s, true, prod\_s\mathrm{E}ff).$
$conj(prod\_s\mathrm{E}ff).$
$in(prod\_s\mathrm{E}ff, e\_n).$
$in(prod\_s\mathrm{E}ff, w\_s).$

$auto(dis\_w, dis\_w\mathrm{Cond}, dis\_w\mathrm{E}ff).$
$disj(dis\_w\mathrm{Cond}).$
$in(dis\_w\mathrm{Cond}, w\_s).$
$in(dis\_w\mathrm{Cond}, dis\_w\mathrm{Cond}1).$
$conj(dis\_w\mathrm{Cond}1).$
$in(dis\_w\mathrm{Cond}1, w\_m).$
$in(dis\_w\mathrm{Cond}1, t\_os).$
$conj(dis\_w\mathrm{E}ff).$
$in(dis\_w\mathrm{E}ff, s\_sup).$

$auto(fau\_p, fau\_p\mathrm{Cond}, fau\_p\mathrm{E}ff).$
$conj(fau\_p\mathrm{Cond}).$
$in(fau\_p\mathrm{Cond}, s\_sup).$
$conj(fau\_p\mathrm{E}ff).$
$in(fau\_p\mathrm{E}ff, d).$

Programme $\pi_{sce}(\sigma)$

Exemple de préemption

$$performs(prod\_m, 0).$$
$$performs(prod\_s, 1).$$

Exemple de duplication

$$performs(prod\_m, 0).$$
$$performs(prod\_s, 0).$$

Programme $\pi_{\mathbb{A}}$

Axiomes de préconditions des évènements

$$triggered(\mathrm{A}, \mathrm{GD}, \mathrm{T}) :- action(\mathrm{A}, \mathrm{GD}, \mathrm{E}effect), performs(\mathrm{A}, \mathrm{T}), holds(\mathrm{GD}, \mathrm{T}).$$
$$triggered(\mathrm{U}, \mathrm{GD}, \mathrm{T}) :- auto(\mathrm{U}, \mathrm{GD}, \mathrm{E}effect), holds(\mathrm{GD}, \mathrm{T}).$$
$$overtaken(\mathrm{E1}, \mathrm{T}) :- triggered(\mathrm{E1}, \_, \mathrm{T}), happens(\mathrm{E2}, \_, \mathrm{T}), priority(\mathrm{E2}, \mathrm{E1}), \mathrm{E2} != \mathrm{E1}.$$
$$happens(\mathrm{E}, \mathrm{GD}, \mathrm{T}) :- triggered(\mathrm{E}, \mathrm{GD}, \mathrm{T}), not\ overtaken(\mathrm{E}, \mathrm{T}).$$

Axiomes d'effets des évènements

$$apply(\mathrm{A}, \mathrm{E}effect, \mathrm{T}) :- happens(\mathrm{A}, \mathrm{GD}, \mathrm{T}), action(\mathrm{A}, \mathrm{GD}, \mathrm{E}effect).$$
$$apply(\mathrm{U}, \mathrm{E}effect, \mathrm{T}) :- happens(\mathrm{U}, \mathrm{GD}, \mathrm{T}), auto(\mathrm{U}, \mathrm{GD}, \mathrm{E}effect).$$
$$initiated(\mathrm{E}, \mathrm{F}, \mathrm{T}) :- apply(\mathrm{E}, \mathrm{E}effect, \mathrm{T}), in(\mathrm{E}effect, \mathrm{F}), not\ holds(\mathrm{F}, \mathrm{T}), fluent(\mathrm{F}).$$
$$terminated(\mathrm{E}, \mathrm{F}, \mathrm{T}) :- apply(\mathrm{E}, \mathrm{E}effect, \mathrm{T}), in(\mathrm{E}effect, neg(\mathrm{F})), holds(\mathrm{F}, \mathrm{T}), fluent(\mathrm{F}).$$
$$holds(\mathrm{F}, 0) :- initially(\mathrm{F}), fluent(\mathrm{F}).$$
$$holds(\mathrm{F}, \mathrm{T}+1) :- initiated(\mathrm{E}, \mathrm{F}, \mathrm{T}).$$
$$holds(\mathrm{F}, \mathrm{T}+1) :- holds(\mathrm{F}, \mathrm{T}), fluent(\mathrm{F}), time(\mathrm{T}), not\ terminated(\mathrm{E}, \mathrm{F}, \mathrm{T}) : event(\mathrm{E}).$$
$$initially(neg(\mathrm{F})) :- not\ initially(\mathrm{F}), fluent(\mathrm{F}).$$
$$holds(neg(\mathrm{F}), \mathrm{T}) :- not\ holds(\mathrm{F}, \mathrm{T}), fluent(\mathrm{F}), time(\mathrm{T}).$$

Axiomes de conjonction, disjonction et négation

$$negative(neg(\mathrm{F})) :- fluent(\mathrm{F}).$$
$$literal(\mathrm{F}) :- fluent(\mathrm{F}).$$
$$literal(\mathrm{F}) :- negative(\mathrm{F}).$$
$$in(\mathrm{L},\mathrm{L}) :- literal(\mathrm{L}).$$
$$holds(true,\mathrm{T}) :- time(\mathrm{T}).$$
$$holds(\mathrm{C},\mathrm{T}) :- conj(\mathrm{C}), time(\mathrm{T}), not\ action(\_,\_,\mathrm{C}), not\ auto(\_,\_,\mathrm{C}), holds(\mathrm{F},\mathrm{T}) : in(\mathrm{C},\mathrm{F}).$$
$$holds(\mathrm{D},\mathrm{T}) :- disj(\mathrm{D}), holds(\mathrm{F},\mathrm{T}), in(\mathrm{D},\mathrm{F}), not\ action(\_,\_,\mathrm{D}), not\ auto(\_,\_,\mathrm{D}).$$
$$event(\mathrm{A}) :- action(\mathrm{A},\mathrm{GD},\mathrm{E}ffect).$$
$$event(\mathrm{U}) :- auto(\mathrm{U},\mathrm{GD},\mathrm{E}ffect).$$

Programme $\pi_{\mathbb{C}}$

$$event(\mathrm{A},\mathrm{GD},\mathrm{E}ffect) :- action(\mathrm{A},\mathrm{GD},\mathrm{E}ffect).$$
$$event(\mathrm{U},\mathrm{GD},\mathrm{E}ffect) :- auto(\mathrm{U},\mathrm{GD},\mathrm{E}ffect).$$
$$inertia(h(\mathrm{L},\mathrm{T}),h(\mathrm{L},\mathrm{T}+1)) :- holds(\mathrm{L},\mathrm{T}), holds(\mathrm{L},\mathrm{T}+1), literal(\mathrm{L}).$$
$$r\_hh(h(\mathrm{GD\_L},\mathrm{T}),h(\mathrm{C},\mathrm{T})) :- conj(\mathrm{C}), holds(\mathrm{C},\mathrm{T}), in(\mathrm{C},\mathrm{GD\_L}).$$
$$r\_hh(h(\mathrm{GD\_L},\mathrm{T}),h(\mathrm{D},\mathrm{T})) :- disj(\mathrm{D}), holds(\mathrm{D},\mathrm{T}), in(\mathrm{D},\mathrm{GD\_L}), holds(\mathrm{GD\_L},\mathrm{T}).$$

NESS-causes directes

$$direct\_ness(o(ini(\mathrm{L}),-1),h(\mathrm{L},0)) :- initially(\mathrm{L}).$$
$$direct\_ness(o(\mathrm{E},\mathrm{T}),h(\mathrm{F},\mathrm{T}+1)) :- initiated(\mathrm{E},\mathrm{F},\mathrm{T}).$$
$$direct\_ness(o(\mathrm{E},\mathrm{T}),h(neg(\mathrm{F}),\mathrm{T}+1)) :- terminated(\mathrm{E},\mathrm{F},\mathrm{T}).$$
$$direct\_ness(\mathrm{E}vent,h(\mathrm{L},\mathrm{T}+1)) :- direct\_ness(\mathrm{E}vent,h(\mathrm{L},\mathrm{T})), inertia(h(\mathrm{L},\mathrm{T}),h(\mathrm{L},\mathrm{T}+1)).$$
$$direct\_ness(\mathrm{E}vent,h(\mathrm{GD},\mathrm{T})) :- direct\_ness(\mathrm{E}vent,h(\mathrm{GD\_L},\mathrm{T})), r\_hh(h(\mathrm{GD\_L},\mathrm{T}),h(\mathrm{GD},\mathrm{T})).$$

NESS-causes

$$ness(o(\mathrm{E1},\mathrm{T1}),h(\mathrm{GD\_L},\mathrm{T2})) :- direct\_ness(o(\mathrm{E1},\mathrm{T1}),h(\mathrm{GD\_L},\mathrm{T2})).$$
$$ness(o(\mathrm{E1},\mathrm{T1}),h(\mathrm{GD\_L},\mathrm{T3})) :- actual(o(\mathrm{E1},\mathrm{T1}),o(\mathrm{E2},\mathrm{T2})), ness(o(\mathrm{E2},\mathrm{T2}),h(\mathrm{GD\_L},\mathrm{T3})).$$

Causes effectives

$$actual(o(\mathrm{E1},\mathrm{T1}),o(\mathrm{E2},\mathrm{T2})) :- ness(o(\mathrm{E1},\mathrm{T1}),h(\mathrm{GD},\mathrm{T2})), happens(\mathrm{E2},\mathrm{GD},\mathrm{T2}), auto(\mathrm{E2},\mathrm{GD},\mathrm{E}ff).$$

Annexe C

Glossaire

acceptable un argument $a \in A$ est acceptable par un ensemble S, si S attaque tous les attaquants de a. 262

actions « conduites d'un humain (ou d'une entité anthropomorphisée) doté d'une raison d'agir (motif) et d'une intention » [Revaz, 2009]. 42, 161

admissible ensemble d'arguments S sans conflit et où tous ses éléments sont acceptables par S. 262

automatiser consiste à indiquer comment les processus sont reproduits par des algorithmes [Saint-Cyr et collab., 2014]. 41

bien une des trois notions de base en théorie morale. La notion de bien peut être rapprochée de la notion de valeur. Elle s'applique aux choses qui composent le monde. 18

But-for test ce test couramment utilisé en droit stipule que [Wright, 1985]: « an act was a cause of an injury if and only if, but for the act, the injury would not have occurred ». Il s'agit ici d'une simple reformulation de la dépendance causale. Il est aussi connu sous le nom de « Conditio Sine Qua Non » [Satoh et Tojo, 2006]. 66

cadre causal description d'un problème et d'une situation spécifique dans ce problème. Il s'agit du cadre précis dans lequel le raisonnement s'inscrit. 140, 167

causalité effective peut être vue comme la détermination des facteurs présents dans une situation donnée et qui ont produit une conséquence. Autrement dit, il s'agit de déterminer qu'elle partie d'une loi causale a été appliquée dans un cas précis. Cette capacité peut être assimilée à un raisonnement a posteriori. 58

causalité générale peut être vue comme la découverte de lois régissant le monde dans lequel nous vivons. C'est en quelque sorte ce que la science cherche à déterminer. La connaissance de ces lois ouvre la porte à un raisonnement a priori. 58

causalité négative partie de la causalité qui traite aussi bien des causes derrière le fait qu'un évènement ne se soit pas produit, que des conséquences que peut avoir la non occurrence d'un évènement. Nous parlons également de négation dans la relation causale. 78

causalité positive partie de la causalité qui traite des causes d'une partie de l'état du monde ou d'évènements se produisant. 78

causes effectives relation causale où la cause ainsi que la conséquence sont des occurrences d'évènements. 141, 186

causes effectives de non occurrence relation causale où la cause est une occurrence d'évènement et la conséquence une non occurrence d'évènement. 231

chemin causal chemin causal entre deux occurrences d'évènements qui est une séquence d'occurrences d'évènements où chaque évènement est une $\mathscr{F}$-cause des préconditions du suivant, si bien qu'il contribue au déclenchement du suivant. 142

chemin causal vers une non occurrence chemin causal entre une occurrence d'évènement et une non occurrence d'évènement qui est une séquence d'occurrences d'évènements où chaque évènement est une $\mathscr{F}$-cause du suivant, si bien qu'il contribue au déclenchement du suivant, et le dernier est une $\mathscr{F}$-cause de la négation des conditions de déclenchement de l'évènement qui ne se produit pas.. 232

code de conduite ensemble de normes morales qui dictent la façon juste d'agir. 21

cohérent un ensemble de littéraux de fluents est dit cohérent si $\forall l \in \mathrm{L}, \overline{l} \notin \mathrm{L}$. 43, 160

complet un ensemble de littéraux de fluents est dit complet si $\forall f \in \mathbb{F}, f \in \mathrm{L}$ ou $\neg f \in \mathrm{L}$. 43, 160

conditions de déclenchement toutes les conditions devant être satisfaites par l'état S pour que l'évènement puisse se produire. 162

conséquences tous les effets qu'une occurrence d'évènement a réellement, directs comme indirects. 165

conséquences négatives étant donné une relation binaire reliant une cause à une conséquence, il s'agit des cas où la conséquence est une non occurrence d'évènement. 218

conséquentialisme agent-neutre l'assignation de valeur à une conséquence est complètement indépendante de l'agent que cela affecte. 27

conséquentialisme direct consiste à évaluer les conséquences spécifiques de la réalisation d'une action dans un contexte donné. 28

conséquentialisme effectif ce sont les conséquences réelles qui doivent être prises en compte, celles qui ont effectivement eu lieu. 28

conséquentialisme impartial la valeur d'une conséquence qui affecte un agent sera la même quel que soit l'agent. 27

conséquentialisme maximisant considère qu'une action est juste si elle produit la plus grande quantité totale de valeur parmi toutes les alternatives. 29

conséquentialisme non priorisant l'appartenance d'un agent à un groupe n'a aucun impact sur la valeur qu'a une conséquence pour lui. 27

conséquentialisme universaliste considère que tous les individus impactés par les conséquences de l'action doivent être pris en compte dans son évaluation. 28

contexte description des éléments d'un problème comme les fluents qui décrivent le monde, les évènements qui décrivent les transitions, les temps de simulation, et les priorités entre les évènements. 140, 166

contrefactuel causalement si c est une cause effective de e dans un scénario, alors c doit pouvoir prendre une autre valeur dans un scénario contrefactuel où avec cette valeur c n'est pas suffisante causalement pour e. 67

correcte en cooccurrence (S, E, S') où $\nexists (e, e') \in E^2,\ e \succ_{\mathbb{E}} e'$, i.e. toutes les priorités entre les évènements ont été respectées, cela empêche que deux évènements interférents puissent avoir lieu en même temps. 164

correcte en déclenchement (S, E, S') où $\forall e \in \mathbb{E}$ tel que $S \vDash tri(e)$, $e \in E$ ou $\exists e' \in E,\ e' \succ_{\mathbb{E}} e$, i.e. tous les évènements qui devaient être dans E le sont bien. Si un évènement a ses conditions de déclenchement satisfaites par S, alors il doit être dans E, à moins qu'un autre évènement de E lui soit prioritaire. 164

dépendance causale si un évènement e est dépendant causalement d'un évènement c, alors l'occurrence de c implique l'occurrence de e. 67

désaccord causal fondamental désaccord sur les relations causales attendues étant donné un type de cas de surdétermination donné, la divergence se situe dans la définition même de ce qu'est la causalité. 84

désaccord causal non fondamental désaccord sur le type de cas de surdétermination qui est traité, la divergence est sur le problème qui est traité et la façon dont il est formalisé. 84

désaccord moral fondamental conflit entre les normes morales les plus basiques, même l'unanimité sur l'ensemble des faits non moraux concernant le conflit ne permettrait pas de le résoudre. 23

désaccord moral non fondamental conflit qui prend ses racines exclusivement dans les faits non moraux le concernant. Ce type de conflit peut être résolu de façon rationnelle en faisant appel à la science étant donné qu'il dépend exclusivement de faits non moraux. 23

dialogue séquence d'énoncés en langage naturel. Formellement, $\Delta = \{(a, o) | (a, o) \in A \times \mathbb{N}\}$, où chaque argument a est associé à son ordre d'énonciation o. 262

doctrine du double effet principe permettant de déterminer le statut déontique d'une action qui apporterait à la fois du bien et du mal dans la théorie du droit naturel. 33, 115

domaine de la causalité ensemble de la communauté pluridisciplinaire s'intéressant à la causalité. Cela inclut des philosophes, des juristes, des psychologues, des mathématiciens, des physiciens et des informaticiens. 58

effets changements de l'état du monde attendus si l'évènement se produit. 162

effets conditionnels effets qui ne sont pas appliqués dès que les préconditions sont satisfaites, il faut en plus que des conditions supplémentaires le soient. 206

effets effectifs étant donné un état partiel $L \subseteq Lit_{\mathbb{F}}$ et un ensemble d'évènements $E \subseteq \mathbb{E}$, les effets effectifs de E sont les conséquences qu'un ensemble d'évènements E se produisant à L a parmi ses effets intrinsèques. 165, 207

effets épistémiques effets « qui portent sur les croyances de l'agent » [Saint-Cyr et collab., 2014]) devant être distingués des effets ontiques. 42

effets intrinsèques effets avec lesquels un évènement a été formalisé $eff(e)$. 164

effets ontiques effets « qui portent sur le monde » [Saint-Cyr et collab., 2014] devant être distingués des effets épistémiques. 42

ensemble de décisions l'ensemble de décisions $\mathbb{D}$ est un ensemble d'ensembles d'actions défini comme : $\mathbb{D} = \{i : \mathbb{A} \to \{1,0\} \mid [\mathrm{C}]^i = 1\}$. 100

ensemble de scénarios ensemble des scénarios envisageables dans κ_c que nous notons $\sigma^{\mathbb{D}}$. 244

ensemble effectivement suffisant situation précise où tous les éléments de l'ensemble suffisant sont effectivement présents et donc nous pourrions dire que celui-ci a été effectivement suffisant à produire la conséquence. 64

ensemble suffisant de NESS-causes ensemble suffisant pour expliquer la véracité d'une formule vraie. 182

ensemble suffisant de NESS-causes directes ensemble suffisant pour expliquer la véracité d'une formule vraie composé uniquement de NESS-causes directes. 170

état ensemble de fluents décrivant le monde. Selon le STEE, des conditions différentes doivent être satisfaites pour que l'ensemble puisse être un état. 42, 139, 160

état argumentatif état dans lequel l'acceptabilité de tous les arguments présents après l'énonciation d'un nouvel argument est mise à jour. 263

état argumentatif final état argumentatif $\mathrm{S}^{\chi}(t)$ dans lequel $\forall x \in \mathrm{A}$, $\exists t' \in \mathbb{T}$ tel que $t' < t$ et $enunciate_x \in \mathrm{E}^{\chi}(t')$. 269

états partiels ensemble $\mathrm{L} \subseteq Lit_{\mathbb{F}}$ cohérent mais incomplet. Il s'agit de descriptions partielles du monde. 161

éthique branche de la philosophie s'intéressant à la morale. 17

éthique appliquée branche de l'éthique normative dont le but est d'étudier une pratique ou action spécifique. 17

éthique computationnelle domaine qui se préoccupe d'assurer que des systèmes informatiques qui automatisent au moins une partie d'un processus décisionnel prennent des décisions qui puissent être vues comme éthiques par les humains. C'est un sous-domaine de l'intelligence artificielle. 36

éthique computationnelle normative sous-domaine de l'éthique computationnelle qui cherche à atteindre l'objectif du domaine en formalisant des théories morales. 38

éthique normative branche de l'éthique regroupant deux branches qui cherchent à répondre à des questions morales. La première branche est la théorie morale qui s'intéresse à ce qu'est le juste et le bien. La deuxième branche est l'éthique appliquée dont le but est d'étudier une pratique ou action spécifique. 17

évènements variables décrivant les transitions. Le terme connote la possibilité d'actions sans agents [Russell et Norvig, 2010, chap 12]. 139, 160

évènements naturels « phénomènes se produisant dans la nature sous l'effet d'une cause » [Revaz, 2009]. 161

exécutable $(\mathrm{S}, \mathrm{E}, \mathrm{S}')$ où $\forall e \in \mathrm{E}$, $\mathrm{S} \vDash pre(e)$, i.e. tous les évènements dans l'ensemble E ont leurs préconditions satisfaites par l'état S. 164

expressif mesure d'à quel point un problème peut être représenté de façon concise dans un langage. 205

extrinsèquement mauvaise ou bonne qui est reliée a une autre chose qui a une valeur intrinsèque négative ou positive respectivement. 18

$\mathscr{F}-causes$ relation causale où la cause est une occurrence d'évènement et la conséquence est une formule d'état $\mathscr{F}$. 141

fluents variables représentant classiquement les propriétés du monde pouvant varier dans le temps. 42

force une non occurrence extension de la relation cause effective d'une non occurrence par assouplissement de la transitivité. 252

formalisation processus en trois étapes : modéliser, représenter et automatiser [SAINT-CYR et collab., 2014]. 41

formule d'effets conditionnels formule qui contient les informations reliées aux conditions des effets conditionnelles qui par leur véracité ont permis a une occurrence d'avoir ses effets effectifs et donc d'être une cause de la conséquence. 210

formules d'état formules de fluents construites en utilisant les opérateurs logiques classiques. 51, 161

hédoniste idée selon laquelle le bien-être d'un individu peut se résumer au plaisir et à la souffrance qu'il ressent. 26

interférents $e, e' \in \mathbb{E}^2$ sont interférents dans $\mathscr{S}_c$ si $\{l \in Lit_{\mathbb{F}} | l \in eff(e) \cup eff(e')\}$ n'est pas cohérent au sens de la définition 6.1. Ces évènements sont interférents dans $\mathscr{S}_c^+$ si l'ensemble $\{l \in Lit_{\mathbb{F}} | \exists \psi \in \mathscr{F},\ S \vDash \psi, [\psi] l \in eff(e) \cup eff(e')\}$ n'est pas cohérent au sens de la définition 6.1. 164, 207

interventionnisme opération permettant de fixer arbitrairement la valeur d'un ensemble de variables dans le scénario en respectant un ensemble de conditions. 67

intrinsèquement bonne le bien est propre à cette chose, elle est bonne en elle-même ou en tant que telle. 18

intrinsèquement mauvaise le mal est propre à cette chose, elle est mauvaise en elle-même ou en tant que telle. 18

intrinsèquement neutre qui n'est ni intrinsèquement mauvaise ni intrinsèquement bonne. 18

juste une des trois notions de base en théorie morale. La notion de juste peut être rapprochée de la notion de conduite juste ou de devoir. Elle s'applique aux actions ou décisions. 18

littéral de fluent représente le fluent $f \in \mathbb{F}$, ou sa négation $\neg f$. 42, 160

maxime modélisation d'un état psychologique d'un agent indiquant ce qu'il désire faire. Il s'agit d'une déclaration de la forme : je vais faire A, si certaines conditions sur l'état du monde S sont réunies, afin d'accomplir O. 25

méthaéthique branche de l'éthique regroupant différentes branches qui s'intéressent aux aspects non moraux autour de l'étude de la morale. Les questions abordées sont généralement classées en trois catégories distinctes : sémantiques, métaphysiques ou épistémologiques. 17

modéliser consiste à définir rigoureusement les concepts et les processus d'un point de vue mathématique [SAINT-CYR et collab., 2014]. 41

négation par l'échec règle d'inférence non monotone permettant de raisonner en l'absence d'information. 192

NESS-causes relation causale où la cause est une occurrence d'évènement et la conséquence est une formule d'état $\mathscr{F}$. 183, 210

NESS-causes directes type spécifique de NESS-cause où l'intérêt est porté uniquement aux conséquences directes de l'occurrence d'évènement dans le monde, i.e. les effets effectifs de l'évènement. Sa définition est basée sur le test NESS de WRIGHT [2011]. 171

non occurrence d'évènement couple $(\overline{e}, t)$ qui vérifie $e \notin \mathrm{E}^{\chi}(t)$. 219

norme morale déclaration morale au sujet d'un type d'action. C'est un type spécifique de règle morale. 21

objectif pratique consiste à procurer une procédure fiable de décision permettant aux agents rationnels et bien informés de produire des verdicts moraux corrects. 18

objectif théorique consiste à identifier quels aspects d'une action, d'une personne ou d'un état du monde font qu'ils peuvent être qualifiés de juste ou non juste, de vertueux ou non vertueux, ou de bons ou mauvais. 18

occurrences d'évènements retirables étant donné deux ensembles suffisants, un de NESS-causes directes C et l'autre de NESS-causes C′, un ensemble d'occurrences d'évènements retirables est défini comme $\mathrm{C_R} = \mathrm{C} \setminus \mathrm{C}'$ et $\mathrm{C_R} \subseteq (\mathbb{N} \setminus \mathrm{E}^{\chi}(-1))$. 182

occurrences d'évènements substituantes étant donné deux ensembles suffisants, un de NESS-causes directes C et l'autre de NESS-causes C′, un ensemble d'occurrences d'évènements substituantes est $\mathrm{C_S} = \mathrm{C}' \setminus \mathrm{C}$. 182

omission étant donné une relation binaire reliant une cause à une conséquence, il s'agit des cas où la cause est une non occurrence d'évènement. 218

omission d'agir décision consistant à ne réaliser aucune action. 100

omission d'une décision étant donné une décision $d \in \mathbb{D}$, omettre de réaliser d correspond à réaliser une décision $d' \in \overline{d}$, i.e. une alternative. 100

opérateur de mise à jour étant donné un état partiel $\mathrm{L} \subseteq Lit_{\mathbb{F}}$ et un ensemble d'évènements $\mathrm{E} \subseteq \mathbb{E}$, l'opérateur de mise à jour indique l'état partiel obtenu si E avait lieu dans L. 166

partition d'un ensemble Étant donné un ensemble X, $\mathrm{X}_1, \ldots, \mathrm{X}_k$ est une partition de X ssi : (i) $\forall i \in \{1, \ldots, k\}, \mathrm{X}_i \neq \varnothing$; (ii) $\bigcup_{i \in \{1,\ldots,k\}} \mathrm{X}_i = \mathrm{X}$; (iii) $\forall i, j \in \{1, \ldots, k\}^2, i \neq j \implies \mathrm{X}_i \cap \mathrm{X}_j = \varnothing$. 170

permet relation de responsabilité qui doit être comprise dans le sens de rendre possible quelque chose. Elle correspond en anglais à « enables ». 252

préconditions conditions qui doivent être satisfaites par l'état pour que l'évènement puisse se produire. 140, 162

principe moral règle générale posant les conditions selon lesquelles une action est juste, un état du monde est bon ou une personne est vertueuse. 20

problème de ramification problème de représentation car il faudrait pour être exhaustifs spécifier lorsqu'une action est faite tout ce qui change, directement et indirectement. Plus la représentation du monde est complexe, plus le fait de devoir décrire les effets indirects des actions devient tédieux. 43

problème du décor problème de représentation car il faudrait pour être exhaustifs spécifier lorsqu'une action est faite ce qui change et ce qui ne change pas. Plus la représentation du monde a de fluents, plus le fait de devoir décrire tout ce qui ne change pas devient tédieux. 43

programme de causalité effective étant donné un scénario $\pi_{sce}(\sigma)$, un contexte $\pi_{con}(\kappa_c)$, un langage de description d'action $\pi_{\mathbb{A}}$ et une définition de causalité effective $\pi_{\mathbb{C}}$, le programme de causalité effective est $\Pi(\chi) = \pi_{sce}(\sigma) \cup \pi_{con}(\kappa_c) \cup \pi_{\mathbb{A}} \cup \pi_{\mathbb{C}}$. 196

règle morale déclaration morale au sujet d'un type d'action, d'un état du monde ou d'un groupe d'agents. 20

relation causale relation binaire qui lie une cause à une conséquence. 58

représenter consiste à indiquer comment les concepts doivent être codés pour être traités par un ordinateur [SAINT-CYR et collab., 2014]. 41

responsabilité collective par omission version de la responsabilité par l'omission où le monde dans lequel se placer dans le raisonnement hypothétique est celui où tous les agents accomplissent leur devoir. 246

responsabilité crédule par omission version de la responsabilité par l'omission où n'importe lequel des monde où l'agent évalué accomplit son devoir peuvent être utilisés dans le raisonnement hypothétique. 246

responsabilité individuelle par omission version de la responsabilité par l'omission où le monde dans lequel se placer dans le raisonnement hypothétique est celui où seul l'agent évalué accomplit son devoir. 245

responsabilité sceptique par omission version de la responsabilité par l'omission où tous les monde où l'agent évalué accomplit son devoir doivent être utilisés dans le raisonnement hypothétique. 246

sans conflit ensemble d'arguments S où il n'y pas d'arguments $(a, b) \in S^2$ qui s'attaquent l'un l'autre. 261

scénario description d'une situation spécifique dans un contexte. Il s'agit d'une séquence d'actions qui doivent être réalisés à un temps donné. Les éléments du scénario est ce que nous cherchons à évaluer moralement. 167

séquence ensemble d'actions ordonnées. Formellement, $\varsigma \subseteq \mathbb{A} \times \mathbb{N}$. 264

signature d'action en géneral il s'agit du couple $\langle \mathbb{F}, \mathbb{A} \rangle$ où $\mathbb{F}$ est un ensemble de fluents et $\mathbb{A}$ est un ensemble d'actions. En Calcul des Évènements il s'agit du quadruplet $\langle \mathbb{F}, \mathbb{A}, \mathbb{T}, \leq \rangle$ où en plus nous avons $\mathbb{T}$ qui est l'ensemble de points temporels et $\leq$ est un ordre partiel sur l'ensemble $\mathbb{T}$. 42, 53

statut déontique le fait qu'une action soit juste ou non. Plus exactement, dire si une action est requise, optionnelle ou interdite, où une action requise ou optionnelle est une action juste et où une action interdite ne l'est pas. 18

support façon dont un évènement peut se produire, intrinsèquement reliée à ses préconditions. Plus précisément, un support est un impliquant premier de ses préconditions. 149, 171

surdétermination cas où, dans un cadre χ donné, plus d'une cause pourrait produire une occurrence d'évènement à elle seule. 145

surdétermination duplicative asynchrone cas de surdétermination où $\Omega^1 = \{\omega^1\}$, $\Omega^2 = \{\omega^2\}$, $W^1 \neq W^2$ et $t_1^1 < t_1^2$. 150

surdétermination duplicative synchrone cas de surdétermination où $\Omega^1 = \{\omega^1\}$, $\Omega^2 = \{\omega^2\}$, $W^1 \neq W^2$ et $t_1^1 = t_1^2$. 150

surdétermination imitative cas de surdétermination où $\Omega^1 = \{\omega^1\}$, $\Omega^2 = \{\omega^2\}$, $W^1 = W^2$ et $t_1^1 < t_1^2$. 151

surdétermination négative cas où, dans un cadre χ donné, plus d'une cause pourrait produire une non occurrence d'évènement à elle seule. 232

surdétermination préemptive précoce cas de surdétermination où $\Omega^1 = \{\omega^1\}$, $\Omega^2 = \emptyset$ et $\exists((e_i^1, t_i^1) \in \omega^1, (e_j^2, t_j^2) \in \omega^2), (e_i^1, t_i^1) \rightarrow (\neg pre(e_j^2), t_j^2)$. 149

surdétermination préemptive tardive cas de surdétermination où $\Omega^1 = \{\omega^1\}$, $\Omega^2 = \emptyset$ et $\exists(e_j^2, t_j^2) \in \omega^2, (e_\psi, t_\psi) \rightarrow (\neg pre(e_j^2), t_j^2)$. 149

surdétermination symétrique cas de surdétermination où $\Omega^1 = \{\omega^1\}$, $\Omega^2 = \{\omega^2\}$, $W^1 = W^2$ et $t_1^1 = t_1^2$. 150

système abstrait d'argumentation couple (A, R) où A est un ensemble fini d'arguments et R est une relation binaire sur $A \times A$. 260

système de transition d'états étiqueté modèle d'action et de changement. Un STEE peut être vu comme un graph orienté étiqueté où les états sont les nœuds et les arrêtes sont les actions. Dans ce graph, un triplet de deux nœuds reliés par une arrête est une relation étiquetée de transition entre états. 43

théories axées sur la valeur théories morales où le concept de valeur est nécessaire à la définition du concept de devoir. 19

théories axées sur la vertu théories morales où le concept de vertu est nécessaire à la définition du concept de devoir. 19

théories axées sur le devoir théories morales où le concept de devoir est au moins aussi fondamental que les concepts de valeur. 19

théorie de la valeur processus indéfini qui évalue les fluents dans $\mathbb{F}_e$. Formellement défini comme $\Upsilon : \mathbb{F}_e \rightarrow \mathbb{V}_\Upsilon$. 101

théorie de la vertu ensemble de principes moraux déterminant ce qui est considéré comme ayant de la valeur chez les individus. 19

théorie du juste ensemble de principes moraux déterminant le statut déontique des actions. 19

théorie morale branche de l'éthique normative qui s'intéresse à ce qu'est le juste et le bien. À ne pas confondre avec les différentes théories morales qui la composent dont nous étudions et modélisons certaines dans cette thèse. 17

théorie morale processus qui évalue les décisions dans $\mathbb{D}$. Formellement $\Theta : \mathbb{D} \times 2^{\mathbb{F}_e} \rightarrow \mathbb{V}_\Theta$. 102

traces d'évènements et d'états séquence d'évènements $E^\chi(-1), E^\chi(0), \ldots, E^\chi(N)$ et séquence d'états $S^\chi(-1), S^\chi(0), S^\chi(1), \ldots, S^\chi(N+1)$ vérifiant différentes conditions selon le STEE et décrivant un chemin dans ce STEE. 167, 265

utilité valeur des conséquences d'une action. 30

valides (S, E, S') qui sont exécutables, correctes en cooccurrence et correctes en déclenchement. 164

vertu une des trois notions de base en théorie morale. La notion de bien peut être rapprochée de la notion de valeur. Toutefois, il s'agit ici d'une valeur un peu particulière puisqu'il s'agit de la valeur attribuée à un agent. Il est question d'évaluer la valeur de la personne et son caractère. Elle s'applique aux agents. 18

welfarisme idée selon laquelle la seule chose qui a de la valeur intrinsèquement est le bien-être des individus. 26

Annexe D

Acronymes

ASP Answer Set Programming. 192

CS conséquentialisme satisfaisant. 32

EV éthique de la vertu. 35

FND forme normale disjonctive. 63, 180

GD goal descriptor. 193

HFS humanité comme fin en soi. 24

INUS Insufficient but Non-redundant part of an Unecessary but Sufficient Condition. 64

LU loi universelle. 25

NESS Necessary Element for the Sufficiency of a Set. 70

PDDL Planning Domain Description Language. 41

SEF cadre d'équations structurelles. 67
STEE Système de Transition d'États Étiqueté. 42
STIT Seeing To it That. 69
STRIPS The Stanford Research Institute Problem Solver. 41

TCD théorie du commandement divin. 22
TDN théorie du droit naturel. 33
TMK théorie morale de Kant. 25
TRM théorie du relativisme moral. 22

UA utilitarisme de l'acte. 31
UR utilitarisme de la règle. 31

Annexe E

Notations

$\mathbb{A}$ ensemble de variables décrivant des transitions entre états du monde, des actions. 42, 161

A ensemble fini d'arguments. 260

a_x indique l'acceptabilité de l'argument x. 262

$actual\mathrm{E}\boldsymbol{ff}(\mathrm{E},\mathrm{L})$ conséquences d'un ensemble d'évènements E se produisant à L parmi ses effets intrinsèques. Défini dans $\mathscr{S}_c$ comme $\{l \in Lit_{\mathbb{F}} | \exists e \in \mathrm{E}, l \in \mathit{eff}(e), \text{and } l \notin \mathrm{L}\}$ et dans $\mathscr{S}_c^+$ comme $\{l \in Lit_{\mathbb{F}} | \exists e \in \mathrm{E}, [\psi] l \in \mathit{eff}(e) \wedge \mathrm{L} \vDash \psi \wedge l \notin \mathrm{L}\}$. 165, 207

$after(\mathrm{E},\mathrm{L}_p,\mathrm{L}_m)$ prédicat permettant d'obtenir la formule d'effets conditionnels. 210

$\mathrm{A}tt_a$ ensemble des attaquants directs de a pour la relation R. Formellement, $\mathrm{A}tt_a = \{b \in \mathrm{A} | \mathrm{R}(b,a)\}$. 261

C ensemble d'occurrences d'évènements $\mathrm{C} \subseteq \mathbb{E} \times \mathbb{T}$. 170

$\mathrm{C}' \dashrightarrow (\psi, t_\psi)$ relation causale où la cause est une occurrence d'évènement et la conséquence est une formule d'état $\mathscr{F}$. 183

$\mathrm{C} \rightarrow (\psi, t_\psi)$ type spécifique de NESS-cause où l'intérêt est porté uniquement aux effets effectifs de l'évènement. Sa définition est basée sur le test NESS de Wright [2011]. 171

$\mathrm{C_R}$ ensemble d'occurrences d'évènements retirables $\mathrm{C_R} \subseteq (\mathbb{N} \setminus \mathrm{E}^\chi(-1))$. 183, 210

$\mathrm{C_S}$ ensemble d'occurrences d'évènements substituantes. 182

$c\mathrm{A}_{y,x}$ le fait que y peut attaquer x. 262

$\overline{d}$ ensemble des décisions alternatives possibles à la décision d, $\overline{d} = \mathbb{D} \setminus \{d\}$. 100

$\mathbb{D}$ ensemble de décisions. 100

dde doctrine du double effet. 115

Δ dialogue étant une séquence d'énoncés en langage naturel, définie formellement comme $\Delta = \{(a,o) | (a,o) \in \mathrm{A} \times \mathbb{N}\}$. 262

$\mathrm{E}^\chi(t)$ ensemble d'évènements appartenant à la trace d'évènements étant donné un cadre χ. Formellement, $\mathrm{E}^\chi(t) = \tau_\chi^e(t)$. 140, 168

$\boldsymbol{eff}$ fonction qui associe à chaque évènement ses effets, i.e. les littéraux qui deviennent vrais si l'évènement se produit. 162

$(e,t) \rightsquigarrow (e_\psi, t_\psi)$ relation causale où la cause ainsi que la conséquence sont des occurrences d'évènements. 186

$(e,t) \rightsquigarrow (\overline{e_\psi}, t_\psi)$ relation causale où la cause est une occurrence d'évènement et la conséquence est une non occurrence d'évènement. 231

$(\overline{e},t) \hookrightarrow (\overline{e_\psi}, t_\psi)$ relation de responsabilité. 245

(e,t) couple qui vérifie $e \in \mathrm{E}^{\chi}(t)$. Il s'agit de l'occurrence de l'évènement $e \in \mathbb{E}$. 141

$(\overline{e},t)$ couple qui vérifie $e \notin \mathrm{E}^{\chi}(t)$. Il s'agit de la non occurrence de l'évènement $e \in \mathbb{E}$. 219

$enunciate_x$ action consistant à énoncer un argument x. 262

$\mathbb{F}$ ensemble de variables décrivant l'état du monde. Ces variables représentent classiquement les propriétés du monde pouvant varier dans le temps, elles sont appelées des fluents. 42

$\mathscr{F}$ formules d'état, i.e. formules de fluents construites en utilisant les opérateurs logiques classiques. 51, 161

κ_c contexte dans $\mathscr{S}_c$, i.e. l'octuple $(\mathbb{E}, \mathbb{F}, pre, tri, eff, \mathrm{S}_0, \succ_{\mathbb{E}}, \mathbb{T})$. 166

κ_s contexte dans $\mathscr{S}_s$, i.e. l'heptuplet $(\mathbb{E}, \mathbb{F}, pre, eff+, eff-, \mathrm{S}_0, \mathbb{T})$. 140

$\overline{l}$ complément d'un littéral de fluent, défini comme $\overline{l} = \neg f$ si $l = f$, ou $\overline{l} = f$ si $l = \neg f$. 42, 160

$\overline{\mathrm{L}}$ complément d'un ensemble de littéraux, défini comme $\overline{\mathrm{L}} = \left\{\overline{l} \in Lit_{\mathbb{F}} | l \in \mathrm{L}\right\}$. 42, 160

$\boldsymbol{Lit}_{\mathbb{F}}$ ensemble des littéraux de fluents dans $\mathbb{F}$. 42, 160

$makesAcc_x$ évènement naturel ayant pour rôle de rendre acceptable l'argument non acceptable x si tous ses attaquants n'est acceptable. 264

$makesUnacc_{y,x}$ évènement naturel ayant pour rôle de rendre non acceptable l'argument acceptable x qui est attaqué par un autre argument acceptable y. 263

$\mathbb{N}$ ensemble des évènements naturels dans $\mathscr{S}_c$. 161

not négation par l'échec. 192

Θ théorie du juste d'une théorie morale, elle évalue les décisions dans $\mathbb{D}$. Formellement, $\Theta : \mathbb{D} \times 2^{\mathbb{F}_e} \rightarrow \mathbb{V}_\Theta$. 102

Θ_{cs} théorie du juste correspondant au conséquentialisme satisfaisant. 112

Θ_{hfs} théorie du juste correspondant à la première formulation de l'Impératif Catégorique incarnant la théorie morale de Kant. 106

Θ_{lu} théorie du juste correspondant à la seconde formulation de l'Impératif Catégorique incarnant la théorie morale de Kant. 108

Θ_{tcd} théorie du juste correspondant à la théorie du commandement divin. 104

Θ_{tdn} théorie du juste correspondant à la théorie du droit naturel. 116

Θ_{trm} théorie du juste correspondant à la théorie du relativisme moral. 105

Θ_{ua} théorie du juste correspondant à l'utilitarisme de l'acte. 109

Θ_{ue} théorie du juste correspondant à l'utilitarisme espéré. 110

Θ_{ur} théorie du juste correspondant à l'utilitarisme de la règle. 113

p_x indique si l'argument x est présent ou non dans le graphe. 262

$\Pi(\chi)$ programme de causalité effective $\Pi(\chi) = \pi_{sce}(\sigma) \cup \pi_{con}(\kappa_c) \cup \pi_{\mathbb{A}} \cup \pi_{\mathbb{C}}$. 196

$\pi_{con}(\kappa_c)$ programme en ASP obtenu en traduisant le contexte κ_c. 193

$\pi_{sce}(\sigma)$ programme en ASP obtenu en traduisant le scénario σ. 193

$\pi_{\mathbb{A}}$ programme en ASP obtenu en implémentant le langage de description d'action qui décrit $\mathscr{S}_c$. 192

$\pi_{\mathbb{C}}$ programme en ASP obtenu en implémentant notre approche de causalité effective adaptée à $\mathscr{S}_c$. 192

π_s politique dans $\mathscr{S}_s$, i.e. fonction qui associe à un couple état et temps l'ensemble d'évènements qui sont supposés se produire dans cet état à ce moment là. Formellement cela s'écrit : $\pi_s : 2^{\mathbb{F}} \times \mathbb{T} \to 2^{\mathbb{E}}$. 140

pre fonction qui associe à chaque évènement ses préconditions, i.e. conditions qui doivent être satisfaites par l'état pour que l'évènement puisse se produire. 140, 162

$[\psi]l$ notation des effets conditionnels indiquant que l est un effet d'un évènement qui se produit si la condition $\psi \in \mathscr{F}$ est satisfaite. 206

(ψ, t) couple qui vérifie $S^{\chi}(t) \vDash \psi$. Il s'agit d'une formule $\psi \in \mathscr{F}$ vraie dans un état. 141

R relation binaire sur A × A. 260

$R_{a}b$ relation d'attaque où nous considérons qu'un argument $a \in$ A attaque $b \in$ A. Cela peut également s'écrire $(a, b) \in$ R. 260

S état dans un STEE, i.e. ensemble de fluents décrivant le monde. 42, 160

$\mathscr{S}$ système de transition d'états étiqueté. 43

$\mathbb{S}$ ensemble d'états dans le STEE. 43

$S^{\chi}(t)$ état appartenant à la trace d'états étant donné un cadre χ. Formellement, $S^{\chi}(t) = \tau^{s}_{\chi}(t)$. 140, 168

$\mathscr{S}_{\mathscr{A}}$ système de transition d'états étiqueté correspondant au langage de description d'action $\mathscr{A}$. 49

$\mathscr{S}_{arg}$ système de transition d'états étiqueté pour l'argumentation abstraite. 265

$\mathscr{S}_{\mathscr{B}}$ système de transition d'états étiqueté correspondant au langage de description d'action $\mathscr{B}$. 50

$\mathscr{S}_{\mathscr{C}}$ système de transition d'états étiqueté correspondant au langage de description d'action $\mathscr{C}$. 51

$\mathscr{S}_c$ système de transition d'états étiqueté pour la causalité effective en éthique computationnelle. 164

$\mathscr{S}_c^{+}$ système de transition d'états étiqueté pour la causalité effective en éthique computationnelle. Version plus expressive que $\mathscr{S}_c$ par l'ajout d'effets conditionnels. 207

$\mathscr{S}_{\mathscr{E}}$ système de transition d'états étiqueté correspondant au langage de description d'action $\mathscr{E}$. 54

$\mathscr{S}_e$ système de transition d'états étiqueté pour l'éthique computationnelle. 98

$\mathscr{S}_s$ système de transition d'états étiqueté pour l'étude de la surdétermination. 139

$\mathscr{S}_{\text{SC}}$ système de transition d'états étiqueté correspondant au Calcul des Situations. 53

$\mathscr{S}_{\text{STRIPS}}$ système de transition d'états étiqueté correspondant à STRIPS. 48

σ scénario étant un ensemble d'actions couplées à un temps $\sigma \subseteq \mathbb{A} \times \mathbb{T}$. Il représente la volition des agents. 167

$\sigma^{\mathbb{D}}$ ensemble des scénarios envisageables dans κ_c. 244

ς séquence étant ensemble d'actions ordonnées.Formellement, $\varsigma \subseteq \mathbb{A} \times \mathbb{N}$. 264

$\mathbb{T}$ ensemble de points temporels de la formalisation. 53

τ ensemble des relations étiquetées de transition entre états notées (S, E, S') et devant satisfaire différentes conditions selon le STEE. Formellement $\tau \subseteq \mathbb{S} \times 2^{\mathbb{E}} \times \mathbb{S}$, où $\mathbb{S}$ sont les ensembles d'états dans le STEE et $\mathbb{E}$ l'ensemble des transitions possibles dans le STEE. 43, 164, 207, 265

τ^e_χ trace d'évènements qui est une séquence d'évènements $E^\chi(-1), E^\chi(0), \ldots, E^\chi(N)$ vérifiant différentes conditions selon le STEE. 167, 265

τ^s_χ trace d'états qui est une séquence d'états $S^\chi(-1), S^\chi(0), S^\chi(1), \ldots, S^\chi(N+1)$ vérifiant différentes conditions selon le STEE. 167, 265

tri fonction qui associe aux évènements leurs conditions de déclenchement, i.e. toutes les conditions devant être satisfaites par l'état S pour que l'évènement puisse se produire. 162

V fonction attribuant pour un état donné une valeur de vérité à chaque littéral de fluent, formellement $V : Lit_{\mathbb{F}} \times \mathbb{S} \rightarrow \{false, true\}$. 43

$\mathbb{V}_\Theta$ ensemble des évaluations possibles données par une théorie du juste. 102

$\mathbb{V}_\Upsilon$ ensemble des valuations possibles données par une théorie de la valeur. 101

ω chemin causal étant une séquence d'occurrences d'évènements $\omega = (e_n, t_n), \ldots, (e_1, t_1)$ reliant deux occurrences d'évènements (e_n, t_n), (e_ψ, t_ψ). 142

$\overline{\omega}$ chemin causal négatif étant une séquence d'occurrences d'évènements $\omega = (e_n, t_n), \ldots, (e_1, t_1)$ reliant une occurrence d'évènement (e_n, t_n) et une non occurrence d'évènement $(\overline{e_\psi}, t_\psi)$. 232

W support, i.e. une des façons dont un évènement peut se produire. Correspond à un impliquant premier de ses préconditions. 149, 171

χ cadre causal. Dans $\mathscr{S}_s$ il s'agit d'un couple (κ_s, π_s), où π_s une politique et κ_s le contexte. Dans $\mathscr{S}_c$ il s'agit d'un couple (κ_c, σ), où κ_c le contexte et σ un scénario. 140, 167

Υ théorie de la valeur qui évalue les fluents dans $\mathbb{F}_e$. Formellement $\Upsilon : \mathbb{F}_e \rightarrow \mathbb{V}_\Upsilon$. 101

$\triangleright$ l'opérateur de mise à jour indique l'état partiel obtenu si E avait lieu dans L. Formellement : $L \triangleright E = \left(L \setminus \overline{actualEff(E, L)}\right) \cup actualEff(E, L)$. 166

Remerciements

Chers membres du jury, tout d'abord, merci pour le temps et l'énergie consacrés à la lecture du manuscrit. Un remerciement spécial aux Dr. Dupin de Saint-Cyr et Dr. Bonnet pour leur rapport. Je vous remercie également d'être présents aujourd'hui pour cette soutenance et pour vos questions et retours bienveillants et enrichissants. Je tiens également à remercier la Dr. Beynier pour son accompagnement en tant que membre de comité de suivi durant ces trois années.

Jean Gabriel, je tiens à vous remercier sincèrement pour la confiance que vous m'avez témoignée, je peux vraiment dire que sans cette confiance je ne serais pas ici. Prendre un mécanicien/roboticien de formation sur un sujet entre l'informatique et l'éthique n'était pas un choix conventionnel, et je vous en remercie.

Gauvain, je te remercie sincèrement pour ton accompagnement tout au long de la thèse et pour nos innombrables heures de réunion de travail. Travailler avec toi a été extrêmement enrichissant intellectuellement. Je ne cesserai jamais d'être impressionné par la facilité que tu as à comprendre des choses techniques et complexes.

Marie-Jeanne et Isabelle, travailler avec vous a été une très belle expérience pour moi. J'admire votre implication aussi bien professionnelle qu'humaine dans votre travail de recherche et d'accompagnement des doctorants.

Un grand merci à tous les professeurs qui ont pris part à ma formation. Et un remerciement particulier à Hélène Dumontet, son implication dans le CMI et ses conseils sur mes choix d'orientation ont eu un impact très important dans ma vie.

Guillaume et Yann, travailler à vos côtés a été un vrai plaisir, ces périodes sont sans doute les moments les plus épanouissants de cette aventure. Je ne peux qu'encourager les doctorants à chercher des opportunités de collaborer entre eux, c'est une expérience enrichissante, aussi bien intellectuellement qu'humainement.

À tous mes collègues de couloir, de bureau et au delà de ces frontières, le CMI, l'ISIR, l'association SOphIA, et al., merci pour tous les moments partagés.

Mes amis de Colombie et notamment ceux avec qui j'ai traversé l'Atlantique pour venir faire nos études ici, un grand merci pour votre amitié durant toutes ces années. David, un remerciement spécial pour ton aide à la relecture d'un de mes articles.

Je tiens à remercier toute la famille de ma compagne de m'avoir pleinement intégré dans leur famille. Votre accueil chaleureux et tous les moments passés ont été indispensables pour en arriver là où je suis aujourd'hui. Un remerciement spécial à Bénou pour sa relecture et ses commentaires m'ayant permis d'améliorer le manuscrit.

Un grand merci à toute ma famille, nous avons très souvent vécu loin les un des autres, mais malgré cette distance je sais toujours que je peux compter sur votre amour.

À mes beaux parents, Sandra et Mathias, merci d'être rentrés dans ma vie pour y ouvrir de nouveaux mondes des possibles et pour l'amour que vous avez décidé de me donner.

À mon père et ma mère, je tiens à exprimer ma reconnaissance infinie pour les parents merveilleux que vous êtes. Vous m'avez chacun appris des choses très différentes, des façons très différentes de voir la vie. Si aujourd'hui je suis là, c'est grâce à la somme des choses merveilleuses que ces deux visions m'ont apportées.

Inès, je n'aurais pas pu trouver une plus belle âme pour partager ma vie. Merci pour tout l'amour, le bonheur et les rires que tu apportes au quotidien. Durant ces presque neuf années passées ensemble, nous avons vécu des aventures incroyables et pleines de sens, j'ai hâte de vivre les prochaines.

À Alloco et Hélène Dumontet.