Raisonner sur la manipulation dans les systèmes multi-agents : une approche fondée sur les logiques modales

Christopher Leturc

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité INFORMATIQUE

Préparée au sein de l'Université de Caen Normandie

Raisonner sur la manipulation dans les systèmes multi-agents : une approche fondée sur les logiques modales

Présentée et soutenue par
Christopher LETURC

Thèse soutenue publiquement le 02/12/2019 devant le jury composé de		
M. LAURENT PERRUSSEL	Professeur des universités, Université Toulouse 1 Capitole	Rapporteur du jury
M. HANS VAN DITMARSCH	Directeur de recherche, Loria	Rapporteur du jury
Mme CAROLE ADAM	Maître de conférences, Laboratoire LIG Grenoble	Membre du jury
M. FRANÇOIS SCHWARZENTRUBER	Maître de conférences HDR, Ecole normale supérieure de Rennes	Membre du jury
M. BRUNO ZANUTTINI	Professeur des universités, Université Caen Normandie	Président du jury
M. GREGORY BONNET	Maître de conférences HDR, Université Caen Normandie	Directeur de thèse

Thèse dirigée par GREGORY BONNET, Groupe de recherche en informatique, image, automatique et instrumentation

UNIVERSITÉ CAEN NORMANDIE

MIIS

GREYC

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité INFORMATIQUE

Préparée au sein de l'Université de Caen Normandie

Raisonner sur la manipulation dans les systèmes multi-agents : une approche fondée sur les logiques modales

Présentée et soutenue par
Christopher LETURC

Thèse soutenue publiquement le 02/12/2019 devant le jury composé de		
M. LAURENT PERRUSSEL	Professeur des universités, Université Toulouse 1 Capitole	Rapporteur du jury
M. HANS VAN DITMARSCH	Directeur de recherche, Loria	Rapporteur du jury
Mme CAROLE ADAM	Maître de conférences, Laboratoire LIG Grenoble	Membre du jury
M. FRANÇOIS SCHWARZENTRUBER	Maître de conférences HDR, Ecole normale supérieure de Rennes	Membre du jury
M. BRUNO ZANUTTINI	Professeur des universités, Université Caen Normandie	Président du jury
M. GREGORY BONNET	Maître de conférences HDR, Université Caen Normandie	Directeur de thèse

Thèse dirigée par GREGORY BONNET, Groupe de recherche en informatique, image, automatique et instrumentation

UNIVERSITÉ CAEN NORMANDIE

MIIS

GREYC

La logique sauve de l'ennui.
Arthur Conan Doyle

Remerciements

Cette thèse a été une expérience importante dans ma vie, autant sur le plan scientifique que personnel. Elle est la conséquence d'une collaboration entre plusieurs personnes qui sans eux, la thèse n'aurait pu aboutir. Ainsi, je tiens à remercier tous ceux qui m'ont permis de réaliser cette thèse.

La première personne à qui je tiens à adresser mes remerciements est Grégory Bonnet, Maître de conférence HDR à l'université de Caen, qui est mon directeur de thèse. Je le remercie pour sa gentillesse, sa disponibilité permanente, pour tous ses encouragements mais aussi pour m'avoir partagé ses brillantes intuitions, qui m'auront permis de réaliser dans les meilleures conditions cette thèse.

La seconde personne à qui je tiens à adresser mes remerciements est Bruno Zanuttini, Professeur des universités de Caen, pour avoir accepté d'être membre du jury mais aussi pour m'avoir permis de travailler sur cette thèse. Je le remercie également pour ses relectures d'articles et pour m'avoir poussé à continuer en thèse lorsque j'étais en master.

J'exprime ma profonde gratitude à Hans van Ditmarsch, Directeur de recherche CNRS au laboratoire LORIA de Nancy, mais aussi à Laurent Perussel, Professeur des universités de Toulouse et membre de l'IRIT, d'avoir accepté la charge de rapporteur. Je remercie également Carole Adam, Maître de conférence à l'université de Grenoble Alpes et François Schwarzentruber, Maître de conférence HDR à L'École Normale Supérieure de Rennes pour avoir accepté de faire parti des membres composant ce jury de thèse.

Pour Roger, Marc et Jacqueline Blandamour qui ne sont plus de ce monde aujourd'hui mais pour qui j'aurai toujours une pensée. Mais aussi pour Simone Bogdanski et Jean Leturc. Vous me manquez. Merci à toute ma famille pour m'avoir soutenu, merci à ma mère, mon père, mon oncle Philippe, ma sœur, mon frère ainsi que toutes les autres personnes de ma famille.

Merci à mes amis de longue date Benjamin Crétois et Gautier Delahaye qui m'ont toujours encouragé, soutenu et aussi pour m'avoir supporté leur parler pendant de longues heures de ma thèse.

Je remercie tous mes collègues au GREYC et plus particulièrement mes collègues de bureau Sebastien Gamblin et Josselin Guéneron pour m'avoir supporté pendant cette thèse et pour m'avoir fait part de leurs idées.

Pour toutes celles et ceux dont le nom n'apparaîtrait pas sur cette page de remerciements, veuillez m'en excuser et je tiens à vous dire aussi merci.

Merci à vous tous.

Table des matières

Table des figures

Liste des tableaux

Introduction

Ces dernières décennies, l'Intelligence Artificielle est devenue un domaine phare de la recherche en informatique. Le développement informatique est ainsi passé de la conception de logiciels individuels permettant de résoudre des problèmes simples à la conception de logiciels intelligents, autonomes et interagissant avec d'autres programmes pour résoudre des tâches beaucoup plus complexes. Ces logiciels intelligents sont appelés des *agent artificiels*. Un agent partage un environnement et interagit avec d'autres agents au sein d'un *système multi-agents*. Chaque agent est amené à prendre des décisions en vue d'atteindre des objectifs individuels ou collectifs et dispose de ressources pour les réaliser. Pour savoir comment les agents se répartissent les ressources, les agents doivent interagir entre eux.

Cependant, dans un tel système, des agents peuvent avoir des objectifs contraires avec les autres agents, ou avec des objectifs collectifs. Il est ainsi possible que certains agents soient amenés à tromper les autres agents pour pouvoir réaliser leurs objectifs et ce au détriment des autres agents. Un agent peut par exemple propager des fausses informations, usurper l'identité d'un autre agent, intercepter et altérer les communications entre plusieurs agents ou bien encore mettre en œuvre des stratégies plus complexes pour inciter les autres à prendre des décisions en sa faveur. Pour empêcher des agents malintentionnés d'interagir avec les autres à leur détriment, les systèmes multi-agents peuvent être dotés de règles, nommées *normes*, ou encore intégrer une notion de confiance afin d'inciter les agents à bien se comporter. Un exemple de norme consiste à obliger les agents à signaler auprès d'une autorité les agents qui ne transmettraient pas la même information à deux autres agents du système. Mais les normes ne suffisent généralement pas. En effet, certains agents malintentionnés peuvent contourner ces normes en décidant, soit de ne pas les respecter, soit en jouant sur ces mêmes normes pour inciter les autres à agir d'une certaine façon et ce sans qu'ils ne s'aperçoivent de cette stratégie. Dans le premier cas nous parlons d'*agents malhonnêtes*. Dans le second cas, nous parlons d'*agents manipulateurs*.

L'objectif de cette thèse est de proposer de nouveaux modèles permettant aux agents de raisonner dans un contexte où certains agents sont manipulateurs et peuvent, par exemple, abuser de la confiance entretenue entre les agents. Ainsi dans cette thèse, nous proposons dans un premier temps un modèle logique pour raisonner sur la manipulation et exprimons avec ce modèle certaines stratégies de manipulation usuelles comme le mensonge ou encore la tromperie. Dans un second temps, nous proposons un autre modèle permettant de représenter la notion de confiance en la sincérité entretenue entre deux agents.

Cette thèse propose trois contributions :

1. un système logique pour raisonner sur la manipulation, nommé le système KBE (Leturc and Bonnet, 2018b) ;
2. un système logique pour raisonner sur la confiance, nommé le système TB (Leturc and Bonnet, 2017; Leturc and Bonnet, 2018a) ;
3. une nouvelle méthode des tableaux pour raisonner avec les logiques multi-modales et appliquée pour résoudre le problème de satisfiabilité dans le système TB.

Dans le chapitre 1, nous présentons les systèmes multi-agents ainsi qu'un travail d'état de l'art permettant de comprendre et définir ce que nous appelons *manipulation*, puis nous présentons des travaux en informatique qui se sont intéressés à la détecter et à la combattre. Dans le chapitre 2, nous présentons un ensemble d'outils logiques portés sur la modélisation de la confiance et de notions proches à la manipulation comme l'influence, le mensonge, la malhonnêteté, la tromperie, mais aussi la prise de conscience. Dans les chapitres suivants, nous présentons nos contributions. Ainsi, le chapitre 3 propose un système logique nommé le *système KBE*, permettant de raisonner sur la notion de manipulation, le chapitre 4 présente le *système logique TB* permettant aux agents de raisonner sur la notion de confiance en la sincérité d'un agent. Ensuite, le chapitre 5 présente une méthode des tableaux pour résoudre le problème de satisfiabilité dans le système TB, puis nous implémentons un algorithme pour résoudre ce problème à partir de cette méthode. Enfin, le chapitre 6 présente le bilan de nos travaux et identifie plusieurs pistes de recherches futures.

Première partie

État de l'art sur la manipulation dans les systèmes multi-agents

Chapitre 1

La manipulation dans les systèmes multi-agents

Usuellement, lorsque nous parlons de manipulation, nous pensons à « l'action d'orienter la conduite de quelqu'un, d'un groupe dans le sens qu'on désire et sans qu'ils ne s'en rendent compte » (Larousse, 1867). Dans les sciences sociales, la manipulation est parfois considérée comme une forme de contrôle mental dans laquelle le manipulé n'est pas conscient de ce qui est en train de se passer ou ne peut tout simplement rien faire en conséquence (Van Dijk, 2006). Elle est aussi considérée comme une activité qui vise « à atteindre un but souhaité [...] par la tromperie, la coercition et la ruse, sans tenir compte des intérêts ou des besoins des personnes utilisées dans le processus » (Bowers, 2003). Dans un contexte politique et économique, la manipulation est une tentative par un ou plusieurs individus de structurer une situation de choix de groupe de manière à maximiser les chances d'une issue favorable ou à minimiser les chances d'une issue défavorable (Maoz, 1990). Dans les systèmes de vote, la manipulation désigne toute stratégie mise en œuvre par un agent et visant à fournir un faux profil de préférence afin de s'assurer d'un résultat préféré à celui qui serait obtenu en fournissant son véritable profil de préférence (Gibbard, 1973). Nous proposons dans ce mémoire de donner une définition générale de la manipulation et applicable aux agents artificiels.

Dans un premier temps, puisque la manipulation s'applique à des agents, la section 1.1 présente les *systèmes multi-agents* et s'intéresse davantage à une catégorie d'agents, appelés *agents cognitifs*, capables de raisonner sur leurs états mentaux [1], ainsi que les états mentaux des autres agents comme leurs croyances, leurs désirs, leurs préférences, ou encore leurs intentions. Dans un second temps, puisque des domaines comme les sciences sociales s'intéressent à la manipulation et apportent une certaine compréhension de la notion, la section 1.2 présente un état de l'art sur les différentes définitions de la manipulation du point de vue des sciences sociales. En se fondant sur cet état de l'art, nous retenons une définition de la manipulation dans le cadre de nos travaux. Les sections 1.3 et 1.4 présentent respectivement des méthodes informatiques per-

1. Dans notre cadre, un *état mental* désigne tout état de croyance, de connaissance, de désirs, de confiance, et tout état intentionnel, dans lequel un agent se trouve. Un état mental d'un agent est souvent associé à une proposition comme le fait de croire cette proposition, ou encore le fait de désirer cette proposition. La section 1.1.2 présente les différents états mentaux considérés dans les agents cognitifs.

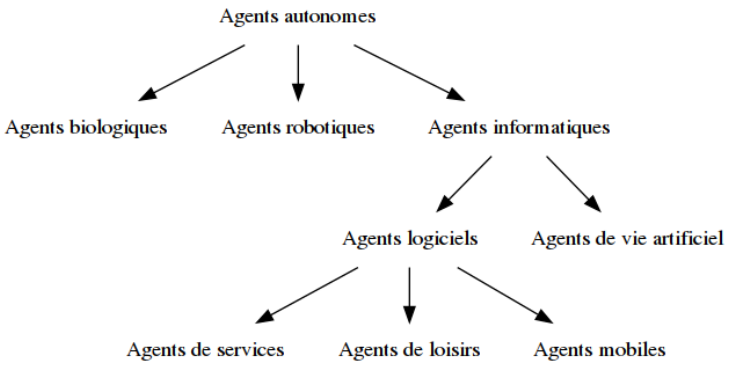

FIGURE 1.1 – Taxonomie des agents (Franklin and Graesser, 1996)

mettant de détecter les stratégies de manipulation, puis des méthodes pour les combattre dans les systèmes multi-agents. Enfin, la section 1.5 présente la problématique générale de cette thèse.

1.1 Des systèmes multi-agents

1.1.1 Définitions et taxonomies

Depuis les années 90 un intérêt considérable est porté sur les systèmes multi-agents (Wooldridge, 2009). Il existe de ce fait dans la littérature de nombreuses définitions d'un agent (Shoham, 1993; Wooldridge and Jennings, 1994; Jennings and Wooldridge, 1996; Franklin and Graesser, 1996; Panait and Luke, 2005). Pour (Panait and Luke, 2005), un agent est un programme informatique capable de prendre des décisions de manière autonome tout en étant capable d'effectuer des actions dans son environnement à partir des informations qu'il a pu en percevoir. (Jennings and Wooldridge, 1996) considèrent quant-à-eux qu'un agent est une entité autonome capable de contrôler son processus de décision et d'agir en vue d'atteindre ses objectifs, et ce, à partir de la perception qu'il a pu avoir de son environnement. Pour (Shoham, 1993), un agent est une entité qui fonctionne continuellement et de manière autonome dans un environnement où d'autres processus se déroulent et d'autres agents existent, tandis que (Franklin and Graesser, 1996) considèrent qu'un agent est une entité naturelle ou artificielle plongée au sein d'un environnement, capable de percevoir localement ce dernier à l'aide de capteurs et d'agir sur celui-ci à l'aide d'effecteurs.

Toutes ces définitions ont comme points communs qu'un agent est capable de :

— prendre des décisions de façon autonome ;

— percevoir son environnement ;

— agir dans et sur son environnement.

Il existe plusieurs types d'agents comme les agents logiciels, les agents mobiles, ou encore les agents physiques. La figure 1.1 représente une taxonomie des principaux types agents. Dans

la suite de ce travail, nous considérons la définition de (Russell and Norvig, 2016) car celle-ci considère chacun des points communs présentés précédemment et elle ne se restreint pas à un seul type d'agent.

Définition 1.1 - *Agents* : *Nous appelons* agent*, toute entité physique ou virtuelle dotée d'un processus de décision autonome, capable de percevoir et d'agir dans son environnement (Russell and Norvig, 2016).*

Lorsque le système est composé de plusieurs agents interagissant dans un même environnement, nous parlons de *système-multi-agents* (Wooldridge, 2009).

Définition 1.2 - *Systèmes Multi-Agents* : *Nous appelons* système multi-agents *(SMA), un système composé d'un ensemble d'agents formant une organisation. Les agents sont situés dans un environnement commun et partagé qu'ils perçoivent, et peuvent agir sur celui-ci. Ils disposent de ressources, d'objectifs individuels ou collectifs, de connaissances, de croyances, d'informations privées, et peuvent communiquer entre agents du système (Wooldridge, 2009).*

Il existe principalement deux types d'organisation pour les systèmes multi-agents : les systèmes multi-agents centralisés et décentralisés. Dans un SMA centralisé, les agents prennent leurs décisions en fonction d'un agent central qui décide à leur place. Dans un système multi-agent décentralisé, les agents prennent des décisions de façon complètement indépendante sans passer par l'intermédiaire d'un agent central. Dans les systèmes multi-agents décentralisés, les agents sont donc nécessairement amenés à interagir entre eux et il existe de ce fait trois types d' interactions possibles : la coopération, la neutralité et la compétition. Des agents peuvent coopérer lorsque les ressources disponibles ou les compétences sont insuffisantes dans le système mais que les agents ont des buts compatibles. Les agents vont donc s'entraider afin de réaliser leurs buts (Ferber and Weiss, 1999). Cependant des agents peuvent être en compétition lorsque leurs objectifs ne sont pas compatibles à cause des ressources et/ou de compétences insuffisantes.

Pour prendre des décisions, les agents sont dotés de processus internes. Ces processus peuvent être délibératifs ou réactifs. Un agent délibératif, que nous appelons aussi *agent cognitif*, raisonne sur la perception qu'il a de son environnement et des autres agents pour produire des actions à effectuer tandis qu'un agent réactif utilise une pile de comportements prédéfinis et réagit directement à sa perception du monde. Les agents réactifs sont fondés, par exemple, sur une architecture de subsomption (Brooks, 1986) telle qu'illustrée par la figure 1.2. Dans ce genre d'architecture, un agent possède une composante qui lui permet de percevoir son environnement, un ensemble de modules qui représentent les compétences de l'agent et une composante exécution qui représente les actions que l'agent met en œuvre. Contrairement aux agents cognitifs, le processus de prise de décision est rapide et reflète directement des perceptions de l'agent.

Cependant, les agents réactifs sont limités à leur simple perception de l'environnement et ne peuvent pas mettre en œuvre des stratégies qui nécessitent par exemple de se représenter les états de croyances des autres agents. Lorsque des agents ont des objectifs contraires avec les autres agents du système, certains peuvent avoir un intérêt à mentir, tromper ou influencer d'autres agents du système. Mais pour effectuer de telles actions, il est nécessaire de se représenter ce

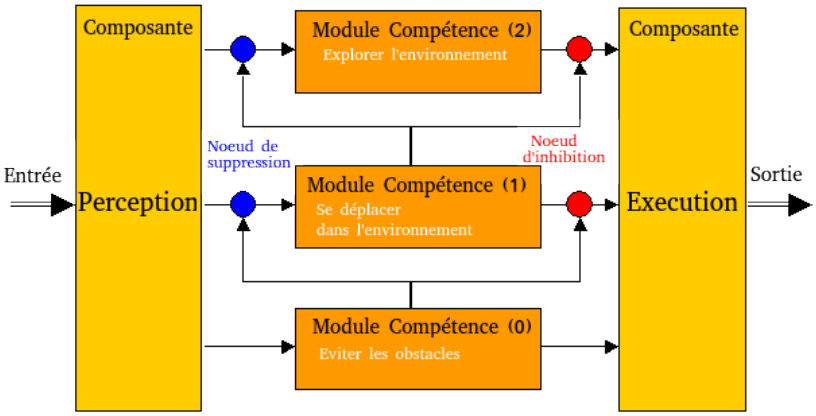

FIGURE 1.2 – Exemple d'architecture de subsomption (Florea et al., 2019)

que savent les autres agents et la manière dont les autres agents prennent leurs décisions. Pour mettre en œuvre de telles stratégies, il est donc nécessaire que ces agents se représentent l'état dans lequel ils se situent mais aussi l'état dans lequel les autres agents sont. Nous parlons d'*états mentaux* (cf. section 1.1.2). De tels agents capables de délibérer sur les états mentaux sont appelés *agents cognitifs*. Dans la suite de la thèse, nous considérons des agents cognitifs.

1.1.2 Agents cognitifs

Les *agents cognitifs* raisonnent sur le monde qui les entoure à travers leurs états mentaux. Un état mental représente une relation entre un agent et une proposition. Il existe plusieurs types d'états mentaux. Nous présentons de manière succincte les principaux états mentaux :

Les **connaissances** décrivent ce que sait l'agent d'une situation ou sur ce que savent les autres agents ;

Les **croyances** décrivent des incertitudes qu'un agent possède sur une situation ou sur ce que croient les autres agents ;

Les **choix** décrivent différentes alternatives futures possibles qu'un agent peut mettre en œuvre par des actions effectuées dans son environnement ;

Les **désirs** décrivent les volontés d'un agent à vouloir agir d'une certaine manière, les désirs sont souvent représentés par une relation de préférence sur des choix possibles d'un agent. Souvent les désirs sont considérés comme des objectifs qu'un agent cherche à satisfaire ;

Les **intentions** d'un agent décrivent les choix d'un agent à agir d'une certaine manière et avec un engagement de l'agent à réaliser ces choix ;

La **confiance** décrit la disposition d'un agent à se fier sur un autre agent par rapport à certains de ses actes ou de ses propos au regard d'une certaine propriété comme, par exemple, sa sincérité, sa fiabilité, sa disposition à agir, sa capacité à réaliser une tache qui lui est donnée, etc. ;

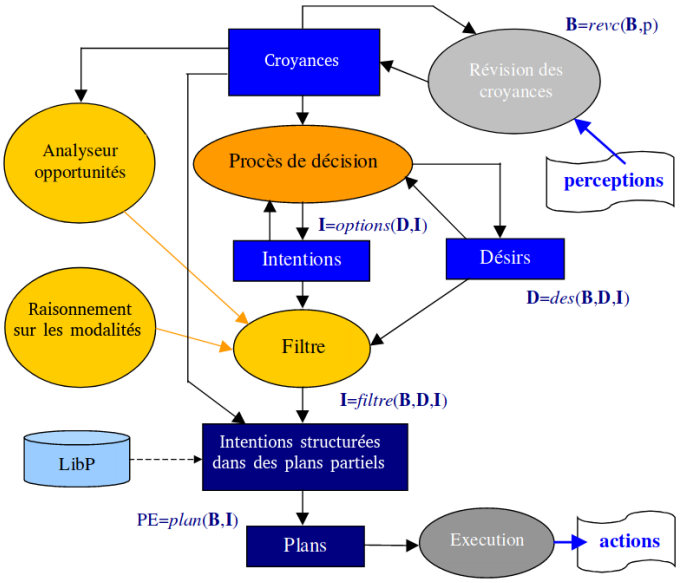

FIGURE 1.3 – Architecture BDI (Florea et al., 2019)

Les **émotions** d'un agent décrivent un ensemble d'états dans lequel un agent se trouve par rapport à une situation donnée. Par exemple, la joie, la vigilance, l'appréhension, la tristesse, l'aversion, la colère et la peur sont des états émotionnels.

Les processus internes des agents cognitifs se fondent souvent sur une architecture BDI[2] (Rao and Georgeff, 1991) telle qu'illustrée par la figure 1.3. Cette architecture BDI permet à un agent de raisonner en tenant compte de ses états mentaux mais aussi de ceux des autres agents du système. Elle est donc composée de plusieurs modules comme un module pour raisonner sur les croyances que l'agent possède sur son environnement, un module qui définit les désirs d'un agent, ainsi qu'un module permettant à l'agent d'inférer de nouvelles intentions pour réaliser ses désirs. Les logiques modales sont souvent utilisées pour exprimer et raisonner sur ces différents états mentaux. Elles introduisent des opérateurs appelés *modalités*, les modalités associent une proposition à un type d'état mental (Blackburn et al., 2002).

Par ailleurs, les différents types de processus internes aux agents, qu'ils soient réactifs ou délibératifs, ne sont pas mutuellement exclusifs. En effet, il existe des architectures pour des agents à la fois réactifs et délibératifs. Par exemple, l'architecture InteRRap pour Reactive-Deliberative Architectures ou Touring Machines (Lemaître and Verfaillie, 2007) est un exemple d'agents hybrides. Le module réactif décide des actions suivantes à effectuer en tenant compte des perceptions de cet agent mais en même temps, l'agent délibère sur les prochaines actions à effectuer sur le monde. L'action finale retenue est la conséquence de ces deux processus combinés.

Les architectures présentées précédemment ne sont pas les seules. Il existe d'autres architectures de processus internes permettant de définir d'autres types d'agents avec d'autres propriétés.

2. BDI correspond aux termes Beliefs Desires Intentions.

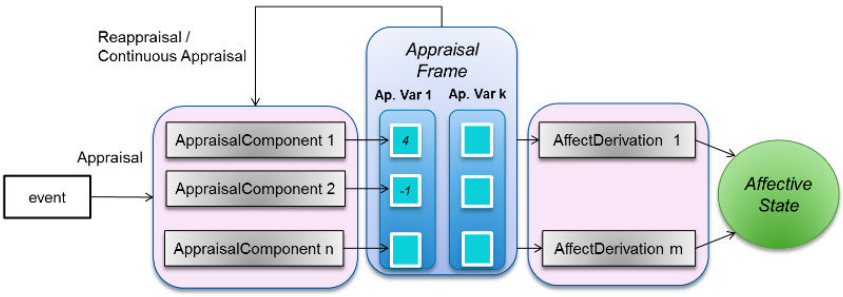

FIGURE 1.4 – Architecture FaTiMa (Dias et al., 2014)

Dans la suite, nous présentons les agents émotionnels, les agents intentionnels, les agents rationnels et les agents communicants.

Agents émotionnels Parmi les architectures hybrides pour les agents cognitifs, l'architecture FaTiMa, représentée par la figure 1.4, permet à des agents de raisonner sur les états émotionnels des autres agents. Par exemple, cette architecture a été utilisée par (Dias et al., 2014) pour simuler une théorie de l'esprit [3], et permettre à des agents de prendre des décisions en se fondant sur les états mentaux des autres agents du système. Elle est composée de plusieurs modules : un module permettant l'évaluation de différents composants liés aux états des agents comme par exemple « la désirabilité », « la désirabilité pour autrui », « le statut d'un but », ou même encore « ce qui est digne d'éloges ». Un autre module permet de statuer en fonction de ces composants, et des croyances sur une situation, quel type d'émotions va être associé à un agent, et ce, en se fondant sur la théorie OCC (Adam et al., 2006). La théorie OCC est une théorie psychologique des émotions, (Adam et al., 2006) s'appuient sur cette théorie et définissent un formalisme logique pour exprimer différents états émotionnels comme la peur, la joie, l'espoir, ou encore la colère. Par exemple, la joie y est définie comme le fait de croire une proposition comme vraie tout en désirant cette proposition. La peur est définie quant-à-elle comme le fait de s'attendre à une certaine conséquence non désirée. Par ailleurs, cette architecture a été utilisée par (Reis, 2012) pour permettre à un agent de délibérer de sorte à tromper d'autres agents du système en exploitant leurs états émotionnels. Un agent ayant peur d'une certaine conséquence sur le monde va être plus facilement influençable qu'un agent qui en est indifférent. Ainsi, un agent malveillant peut, par exemple, jouer sur les craintes d'un agent en le menaçant de les réaliser si l'agent ne veille pas à faire ce que l'agent malveillant attend de lui. Ce procédé de pression par la menace porte le nom de *coercition* en psychologie.

3. La théorie de l'esprit est un terme initialement apparu dans (Premack and Woodruff, 1978) où les auteurs définissent cette théorie comme la capacité d'un agent à inférer les états mentaux des autres agents comme leurs connaissances, leurs croyances, leurs désirs, et leurs intentions.

Agents intentionnels Cette volonté d'un agent à prendre des décisions en fonction de ses états mentaux et de s'engager à réaliser un choix possible traduit la notion même d'intention (Cohen and Levesque, 1990). Nous parlons alors d'*agents intentionnels*. Les choix effectués par les agents sont rarement le fruit du hasard. En effet, un agent va généralement chercher davantage à satisfaire ses désirs et ses buts plutôt que de veiller à amener une situation qui ne représente aucun intérêt pour l'agent. Cette caractéristique est à la base de la notion de *rationalité d'un agent*.

Agents rationnels Un agent est *rationnel* si cet agent prend ses décisions de telle sorte à toujours satisfaire un maximum de ses désirs et buts (Wooldridge, 2003). Les désirs et buts qu'un agent doit satisfaire sont souvent représentés soit par une fonction d'utilité, soit par une fonction de gain qui à chaque but accompli va associer une récompense (Binmore et al., 1998). Ainsi un agent rationnel va donc toujours chercher à résoudre un problème de maximisation de cette fonction d'utilité. Cette notion de rationalité est d'une grande importance en économie pour comprendre et anticiper le comportement des agents au sein du système. Un agent rationnel va donc chercher à atteindre ses buts jusqu'à ce que l'agent croit que son but est atteint ou que ce but est devenu inatteignable (Cohen and Levesque, 1990). Pour permettre aux agents de maximiser les buts et désirs accomplis, les agents doivent planifier en amont pour savoir quel choix ils doivent effectuer pour un instant donné.

Agents communicants Pour interagir, les agents doivent bien souvent communiquer, nous parlons alors d'*agents communicants*. Il existe trois manière pour les agents de communiquer : la première façon est l'introduction d'un protocole que doivent suivre les agents du système, la seconde peut passer par une communication via l'environnement comme la stigmergie [4] et enfin la dernière consiste en des actes de langages (Searle, 1969). Dans le cadre des agents cognitifs, nous nous intéressons seulement aux actes de langage. Un acte de langage est décrit par trois composantes principales :

Composante locutoire : correspond au fait qu'un énoncé doit suivre des règles syntaxiques ;

Composante illocutoire : représente le sens et la force de l'intention sous-jacente à l'énoncé ;

Composante perlocutoire : représente l'effet réalisé sur le destinataire de l'énoncé.

Un agent Anne dit à un autre agent Eve : « Peux-tu me dire si tu as confiance en cet agent Paul ? ». Dans cet énoncé, la composante locutoire est la question posée par l'agent Anne, la composante illocutoire est interrogative et la composante perlocutoire est la réponse donnée par l'agent Eve. Les actes de langage émis par un agent ont pour conséquence de transformer les représentations d'autrui, nous parlons alors d'effet performatif. Il existe plusieurs types performatifs associés aux actes de langages :

Assertif : le fait de dire, affirmer, informer ;

Directif : le fait de demander, réclamer, supplier ;

Commissif : le fait de promettre, garantir, refuser ;

4. La stigmergie est un mécanisme de coordination indirecte entre agents et consiste par exemple à déposer dans l'environnement certaines phéromones pour inciter le comportement des autres agents du système.

Déclaratif : le fait de demander, interroger, consulter ;

Expressif : le fait de féliciter, excuser, approuver, déplorer.

Ces actes de langage permettent donc aux agents de communiquer entre eux. Cependant, des agents cognitifs peuvent utiliser ces actes de langages pour communiquer des informations dans le but d'influencer, ou de tromper d'autres agents du système. De plus, certains de ces agents peuvent aussi ne pas être fiables dans les informations qu'ils transmettent mais pour autant, avoir été sincères. Ainsi, il est nécessaire d'avoir des systèmes multi-agents capables de réguler les interactions entre les agents du système en intégrant au système une notion de confiance entre agents ou de normes.

1.1.3 Réguler les interactions entre agents

Lorsque des agents peuvent être malintentionnés ou malhonnêtes [5] dans un système, il existe deux systèmes permettant de réguler les interactions entre agents : les *systèmes normatifs* et les *systèmes de confiance*. Les *systèmes normatifs* intègrent un ensemble de règles de bonne conduite à respecter, comme par exemple : « Il est interdit de mentir », ou encore « Tu dois rendre la pareille à tout agent te rendant un service ». Les agents sont alors sanctionnés par des pénalités s'ils ne respectent pas ces règles. Les *systèmes de confiance* intègrent la notion de confiance entre agents. Des agents reconnus comme malhonnêtes ou malintentionnés ne sont pas considérés de confiance et sont alors écartés du système. L'évaluation de la confiance peut se fonder sur l'expérience des interactions antérieures avec un agent, l'observation de son comportement dans le système, ou encore sur les témoignages transmis par d'autres agents du système. Ces témoignages lorsqu'ils sont agrégés fondent la notion de confiance collective appelée *réputation*. Nous parlons alors de *systèmes de réputation*.

Cependant, de tels systèmes ne suffisent pas pour empêcher des agents malintentionnés d'agir. En effet, certains agents peuvent exploiter les normes et la confiance entretenue entre agents pour en tirer un avantage, comme nous en discutons à la section 1.2.3.

Systèmes normatifs

Un système normatif peut être vu comme l'émergence d'un ensemble de comportements communs entre agents mais aussi il peut être vu comme un système d'obligations déontiques. Pour éviter les comportements non désirables dans les systèmes multi-agents, une manière consiste à intégrer explicitement un système de normes, obligeant ainsi les agents à bien se comporter. Pour (Boella et al., 2009), un système normatif est un système multi-agents associé à un ensemble de règles déontiques appelées *normes du système* et que les agents peuvent d'une part décider s'ils doivent suivre ou non une règle explicite, et d'autre part, les systèmes normatifs précisent dans quelle mesure les agents peuvent décider de comment modifier ces normes. Boella *et al.* précisent qu'un tel système est donc composé de plusieurs mécanismes pour représenter, communiquer,

5. Nous entendons par *agents malintentionnés*, un agent qui a de mauvaises intentions à l'égard du système. Par exemple, « Blesser ou endommager un autre agent du système ». Un agent est *malhonnête* s'il manque de probité à l'égard du système, i.e. s'il ne respecte par une norme. Par exemple, « Mentir dans le système alors que ce n'est pas permis ».

distribuer, détecter, créer, modifier et faire appliquer les normes ainsi que des mécanismes de raisonnement pour délibérer sur les normes.

Nous retrouvons donc une logique déontique dans un système normatif permettant de décrire les normes. Les logiques déontiques présentées historiquement par (Von Wright, 1951) permettent d'écrire avec un langage de logiques modales les normes du système. Par exemple, les logiques déontiques permettent de décrire formellement le fait qu'« il est nécessaire que les agents disent toujours la vérité dans le système ». Par ailleurs, nous pouvons distinguer deux types d'obligation (Meyer et al., 1988) : le « ce qui doit être » [6] qui imposent aux agents les états du système dans lequel celui-ci doit être et le « ce qui doit se faire » [7] qui contraint les agents à effectuer certaines actions obligatoirement.

Dans un système déontique, il existe généralement différents niveaux d'applications des normes. Pour (Dignum, 1999), il en existe trois : le niveau privé, le niveau contrat et le niveau convention à respecter entre agents. Le plus haut niveau est celui des conventions, ce niveau correspond par exemple au fait qu'« un agent doit être coopératif dans le système ». Ensuite, le niveau contrat désigne les obligations et les autorisations entre agents, une partie importante de ce niveau représente la description des répercussions en cas de violations de la norme. Enfin, la partie privée correspond aux normes qu'un agent va respecter ou non, à ce niveau, une norme d'obligation peut être le but qu'un agent s'oblige à tenir. Pour plus de détails sur les systèmes normatifs, le lecteur intéressé peut se référer à (Testerink, 2017).

Systèmes de confiance et de réputation

Une autre méthode pour inciter des agents à bien se comporter consiste à intégrer une notion de confiance dans le système. Nous parlons alors de *système de confiance*. Mais pour parler de confiance dans un système, il convient de définir ce que nous appelons la confiance. Ainsi, (Castelfranchi and Falcone, 2010) ont développé une théorie de la confiance dans laquelle ils mettent en lumière certaines composantes indispensables à la confiance.

La confiance est au cœur de nos interactions avec les autres. Elle construit l'ordre social et repose sur un cycle : observation - adhésion - convergence comportementale - régularité. La confiance a des répercussions locales (entre agents) et globales (au niveau des organisations d'agents). Elle peut prendre plusieurs formes comme, la confiance dans le contrat social, ou en l'autorité judiciaire, ou encore la confiance dans la négociation. De plus, la confiance n'est pas qu'une norme sociale ou un comportement, c'est aussi un état mental. Ainsi (Castelfranchi and Falcone, 2010) étudient la hiérarchie des différents composants fondamentaux de la confiance mais aussi les aspects dynamiques de la confiance, et plus particulièrement, le cadre de la décision, de la construction d'intentions, de l'acte de faire confiance en soi ou de s'autoriser à déléguer des actions. Ils fournissent tous les ingrédients nécessaires à la construction d'un modèle socio-cognitif pour représenter la confiance. Ils mettent en valeurs deux composantes fondamentales à la confiance : le risque et la délégation qui sont associés à l'action de faire confiance. Pour eux, la confiance n'a lieu d'être que parce qu'il existe un risque derrière l'objet sur lequel porte

6. Traduit « ought-to-be » en anglais.
7. Traduit « ought-to-do » en anglais.

la confiance. Par exemple, si une personne confie ses clés d'appartement à un voisin pour qu'il s'occupe de ses animaux en son absence, cette personne prend le risque de laisser son appartement à la merci de tout le monde. Dans cet exemple puisque la personne fait confiance, elle délègue l'action de nourrir ses animaux au voisin. Mais l'aspect de délégation n'est pas le seul aspect de la confiance, la personne évoquée dans cet exemple fait aussi confiance en l'honnêteté de son voisin pour ne rien voler dans l'appartement. Ainsi Castelfranchi et Falcone décrivent d'autres aspects de la confiance. En fonction d'un contexte, un agent X peut faire confiance en Y pour plusieurs raisons. La confiance peut alors prendre différentes formes comme :

- une intention que l'agent Y aura l'intention de réaliser quelque chose (confiance sur les intentions) ;
- une croyance que l'agent Y est fiable et sincère (confiance en la fiabilité et sincérité) ;
- une capacité (ou disposition) de l'agent Y à faire une certaine action (confiance dispositionnelle) ;
- une action de compter sur Y (confiance en la délégation) ;
- une conséquence sociale sur les interactions entre agents (confiance sociale) ;
- une évaluation ou prédiction sur le futur (confiance sur la prédiction).

De manière générale, nous constatons que la confiance se décrit comme le choix d'un agent de prendre le risque de considérer qu'une certaine propriété abstraite est respectée par un autre agent. Cette propriété abstraite peut correspondre à la délégation, la sincérité, l'honnêteté, ou encore à la fiabilité. La confiance peut être globale, c'est-à-dire que l'agent pour qui la confiance est destinée respecte une certaine propriété abstraite, ou la confiance peut être relative à une proposition lorsqu'elle s'applique à quelque chose en particulier. Une proposition désigne par exemple : « le garagiste X est fiable pour réparer les voitures », ou encore « M. Y est sincère lorsqu'il dit que le travail fourni est de bonne qualité ». Ainsi, en se fondant sur les travaux de (Castelfranchi and Falcone, 2010), nous définissons la confiance de la façon suivante :

Définition 1.3 - *Confiance abstraite* : *Nous appelons* confiance abstraite, *le choix d'un agent de considérer qu'un autre agent respecte une certaine propriété abstraite. La confiance abstraite est dite* relative à une proposition, *si le choix de l'agent qui adresse sa confiance, porte sur une proposition respectant la propriété abstraite de l'agent pour qui la confiance est adressée.*

De plus, remarquons que la confiance peut aussi avoir des aspects quantitatifs comme le soulignent Castelfranchi et Falcone. Elle peut être mesurable en degrés de croyance. Par exemple, un agent peut se fixer un seuil de confiance selon lequel il va accepter de faire confiance. Ce degré de croyance peut et doit aussi reposer sur l'évaluation du risque encouru à faire confiance à quelqu'un. Ils proposent aussi de considérer un degré de méfiance sur la prise de décision de faire confiance ou non à un autre agent.

Cet aspect lié à la méfiance a aussi été une partie étudiée par Castelfranchi et Falcone. Ils étudient la frontière entre le manque de confiance et la méfiance mais aussi étudient le lien étroit entre confiance et certains états émotionnels comme la peur ou la surprise. De plus, la

confiance peut varier au cours du temps. Ainsi, un autre aspect étudié est celui de la dynamique de la confiance : comment peut-on influencer la confiance, ou encore, comment diffuser de la confiance et créer une atmosphère de confiance, ou encore comment prédire qu'un agent va faire confiance à un autre agent ? Castelfranchi et Falcone s'intéressent alors par exemple à la diffusion de la confiance par la présence d'une autorité de confiance, ou encore à la confiance inspirée par la ressemblance entre deux agents (comportements imitatifs), à la pseudo-transitivité de la confiance, mais aussi à la confiance dérivée de l'expérience directe ou indirecte, et les liens entre confiance, contrôle et autonomie sont aussi étudiés. Par exemple, un moyen pour un agent Anne d'accorder sa confiance à un agent Eve consiste à regarder pour quelles tâches similaires, l'agent Anne a déjà fait confiance à Eve.

Cependant Castelfranchi et Falcone mettent en garde sur un point : la confiance peut aussi être un outil de manipulation. Ils affirment que la réelle stratégie pour manipuler quelqu'un consiste avant tout à jouer sur la confiance. Ainsi s'intéressent-ils à la manière dont un agent peut donner l'image d'un agent de confiance et cela passe par ce qu'ils appellent le capital relationnel et le capital social. Ils proposent alors de définir une notion de réseau de dépendances et caractérisent le fait que ce capital relationnel peut être manipulé par des fausses croyances.

Enfin, ils distinguent une notion de confiance relationnelle (ou encore confiance individuelle), d'une notion de confiance globale appelée : *la réputation*. Les systèmes de réputation permettent d'évaluer la réputation des agents qu'ils doivent accorder à un produit, à un vendeur ou à un service (Ruan and Durresi, 2016). Pour évaluer la *réputation*, les agents s'échangent des témoignages des interactions passées avec d'autres agents, puis par un mécanisme d'agrégations sur l'ensemble des témoignages obtenus, les agents évaluent le degré de confiance qu'ils vont accorder à un autre agent. Un tel système va être essentiellement composé de trois mécanismes :

1. un mécanisme de notation propre à chaque agent : notion de confiance individuelle ;
2. un mécanisme de transmission et d'agrégation des témoignages : flots de témoignages ;
3. un mécanisme de construction de la réputation : fonction de réputation.

Par conséquent, la valeur de réputation d'un agent va directement être influencée par les témoignages envoyés par les autres agents du système. Ainsi, un agent malintentionné peut avoir un intérêt de tromper les autres agents (Hoffman et al., 2009) : que ce soit pour diffuser des malwares dans les réseaux pair à pair, mentir sur la qualité d'un produit, ou diffamer des concurrents et ce, en usant de la réputation qu'ils peuvent avoir dans le système.

Dans la littérature, il existe de nombreuses implémentations de la confiance et de la réputation (Josang and Ismail, 2002; Kamvar et al., 2003; Page et al., 1999)(cf. 1.1.3). Certains de ces systèmes peuvent reposer sur des logiques floues (Falcone et al., 2002; Kant and Bharadwaj, 2013; Wang and Huang, 2007) et introduisent une notion subjective de degré de confiance et de vraisemblance d'une source d'information. D'autres travaux s'intéressent à une confiance modélisée par des logiques modales (Dundua and Uridia, 2010; Herzig et al., 2010; Singh, 2011; Smith et al., 2011) [8]. Ces logiques modales permettent d'exprimer la confiance comme la combinaison d'une ou plusieurs modalités d'intentions, de croyances, d'actions ou de buts qu'un agent

8. Nous revenons plus en détails sur les logiques modales représentant la confiance en section 2.2

possède, ou encore comme une modalité spécifique. Si ces approches permettent d'exprimer facilement certains aspects de la confiance comme *la délégation*, elles s'intéressent principalement à la confiance dans l'action des autres agents. Or, dans le cadre des systèmes de réputation, les agents sont amenés à communiquer avec d'autres pour les informer, par exemple, de la qualité des services proposés par des agents tiers. Si certains travaux sont plutôt consacrés aux aspects de révision des connaissances fondés sur la confiance (Lorini et al., 2014; Fan and Liau, 2016), d'autres sont consacrés à la modélisation de la confiance qu'exprime un agent envers le discours d'un autre agent (Christianson and Harbison, 1997; Demolombe, 2004; Liau, 2003; Dastani et al., 2004).

Nous avons présenté trois systèmes multi-agents permettant de réguler les interactions entre les agents : *les systèmes normatifs*, *les systèmes de confiance* et *les systèmes de réputation*. Ces systèmes sont fondés sur l'hypothèse qu'il est possible que dans des systèmes multi-agents, il existe des agents malintentionnés ou malhonnêtes. Cependant de tels systèmes peuvent être exploités par certains agents cognitifs pour mieux influencer, tromper ou manipuler les agents du système. Dans la section suivante, nous définissons la notion de manipulation, puis nous présentons certaines stratégies de manipulation au sein de tels systèmes.

1.2 Comprendre et définir la manipulation

La notion de manipulation est présente dans certains domaines de l'informatique comme en théorie du choix social, ou encore dans les systèmes de réputation où elle est définie comme une stratégie permettant à un agent d'influencer et de contrôler le processus de décision individuel (ou collectif) d'un ensemble d'agents à l'aide de fausses informations afin que ces derniers prennent une décision favorable à l'agent manipulateur (Vallée, 2015). Cependant, cette notion de manipulation n'est définie que par rapport à des applications données, dans des contextes spécifiques comme les systèmes de vote ou les systèmes de réputation. Cependant, ces définitions ne peuvent s'appliquer dans d'autres systèmes comme les réseaux sociaux, où des humains peuvent appliquer des techniques de manipulation, s'appuyant sur des travaux de la psychologie sociale, pour manipuler d'autres humains. Un manipulateur peut alors utiliser ces mécanismes psychologiques pour influencer les décisions d'un autre être humain sans qu'il ne s'en rende compte. L'agent manipulateur n'a alors pas nécessairement recourt à la transmission de fausses informations pour manipuler. (Joule et al., 2002) présentent un ensemble de mécanismes d'influence sur les être humains dans une théorie qu'ils nomment *la théorie de la soumission librement consentie*. Par exemple, une stratégie de manipulation est le *pied dans la porte* qui consiste à obtenir un premier comportement préparatoire de la victime pour l'amener plus facilement à prendre une autre décision en faveur du manipulateur. Un manipulateur peut aussi utiliser les normes sociales d'un système pour influencer les décisions d'autres humains. (Cialdini, 2012) explique que de nombreuses techniques d'influence sont fondées sur les normes sociales comme la *preuve sociale* qui consiste à montrer à un agent victime un ensemble de comportements que d'autres agents exhibent pour l'inciter à imiter ces agents. Nous constatons donc que les définitions de (Gibbard, 1973) et de (Vallée, 2015) ne peuvent pas être appliquées dans un tel contexte puisque

leur définition de manipulation n'y est définie que par rapport à certaines informations révélées par un manipulateur aux autres agents du système. Or, comme nous venons de l'expliquer, une manipulation ne repose pas nécessairement sur une information transmise, elle peut aussi reposer sur un comportement spécifique des agents.

Par conséquent, pour donner une définition suffisamment générale à la manipulation, nous cherchons tout d'abord à comprendre par les sciences sociales ce que les chercheurs définissent par « manipulation ». C'est pourquoi la section 1.2.1 présente un état de l'art sur la manipulation du point de vue de ces sciences. Cet état de l'art nous permet alors de donner une définition à ce terme en section 1.2.2. Enfin, la section 1.2.3 présente des stratégies de manipulation dans les systèmes multi-agents.

1.2.1 Qu'est-ce que la manipulation en sciences humaines ?

Les chercheurs en sciences sociales sont souvent en désaccords quant à la définition de la manipulation. (Kligman and Culver, 1992) soulignent le fait que le terme « manipulation » est souvent employé en psychiatrie mais rarement défini ou discuté. Pour certains psychiatres, la manipulation est considérée comme l'acte d'altérer le jugement d'individus en le privant d'une partie de leurs choix délibérés (Clair, 1966; Kligman and Culver, 1992; Sunstein, 2015). Cependant, une telle définition amènerait à considérer comme de la manipulation la persuasion rationnelle, la tromperie ou encore la coercition, alors que comme le souligne (Rudinow, 1978), « la plupart des gens distingueraient la manipulation de la persuasion, d'une part, et de la coercition, d'autre part » . Ainsi (Handelman, 2009) considère que « la manipulation n'est pas exactement de la coercition, ni de la persuasion et n'est pas tout à fait semblable à la tromperie ».

Ainsi dans la littérature, il existe quatre visions de la manipulation : ceux qui considèrent la manipulation comme un concept vague, ceux qui considèrent la manipulation comme un exercice du pouvoir sur autrui, ceux qui considèrent que la manipulation est une altération du jugement d'autrui et enfin ceux qui considèrent la manipulation comme l'exercice invisible d'un pouvoir. Nous passons en revue chacune de ces écoles et retenons une définition de la manipulation sur laquelle la suite de cette thèse reposera.

La manipulation vue comme un concept vague

(Ackerman, 1995) passe en revue plusieurs situations dans lesquelles le terme « manipulation » est fréquemment employé mais donne des contre-exemples à chacune d'entre elles. Elle suggère alors qu'il n'est pas vraiment possible de donner une définition au concept de manipulation car, selon elle, la manipulation est un concept combinatoirement vague. Par « concept combinatoirement vague », elle désigne le fait qu'il y a une variété de conditions dans lesquelles le terme est fréquemment utilisé. Mais ces conditions ne sont jamais suffisantes pour permettre de discriminer les situations dans lesquelles nous parlons de manipulation, des autres situations dans lesquelles ces conditions ne sont pas suffisantes et/ou nécessaires pour parler de manipulation. Cependant, Ackerman identifie un certain nombre de conditions qui peuvent constituer une manipulation. Ces conditions peuvent être liées à une influence ; une perspicacité du manipulateur ; une déviance ; une utilisation de moyens indirects, astucieux et subtils ; à une inhibition

de la délibération rationnelle du manipulé ; à une falsification ou omission d'informations ; à un jeu sur des impulsions non rationnelles du manipulé ; à de la tromperie ; à des motifs cachés ; à inciter quelqu'un à faire quelque chose différemment de ce qu'il aurait fait sans la manipulation ; à un manque d'éthique ; à une inhibition de l'action, de la croyance, de l'émotion que le manipulé trouve naturel ou approprié ; à de la pression où, le manipulateur cherche à rendre gênant pour le manipulé de dire non. Ackerman met alors en garde contre le fait que nous ne pouvons pas savoir quelles combinaisons de ces conditions constituent réellement une manipulation. Mais cette vision de la manipulation est loin de faire l'unanimité chez les chercheurs.

La manipulation vue comme l'exercice d'un pouvoir sur autrui

Pour beaucoup de chercheurs la manipulation peut être considérée comme une forme d'exercice du pouvoir sur autrui (Kligman and Culver, 1992; Maoz, 1990; Abell, 1977; Goodin, 1980; Todd, 2013).

L'intention de modifier quelque chose de l'environnement (Kligman and Culver, 1992) définissent la manipulation comme l'intention de modifier quelque chose dans son environnement. Par exemple, un manipulateur peut délibérément retenir ou présenter sélectivement, certaines informations et en omettre d'autres, exploiter l'ignorance ou les croyances de sa victime afin de pouvoir garder le contrôle sur ses options qu'elle perçoit et de l'orienter dans la direction souhaitée par le manipulateur. Pour (Maoz, 1990), dans un contexte politique, une manipulation politique est une tentative par un ou plusieurs individus de structurer une situation de choix de groupe de manière à maximiser les chances d'une issue favorable ou à minimiser les chances d'une issue défavorable. (Abell, 1977) pense que la manipulation est un processus impliquant un agent manipulateur A et un manipulé B et consistant pour l'agent A à contrôler les préférences de l'agent B en réduisant sa compréhension de la situation ou en réduisant les moyens ouverts à lui. De manière intéressante, ces définitions de la manipulation rejoignent celles utilisées par les économistes en théorie des jeux et du choix social. Par exemple, (Gibbard, 1973; Gärdenfors, 1976) considèrent qu'un individu manipule un système de vote si, en fournissant un faux profil de préférence, il s'assure d'un résultat qu'il préfère à celui normalement obtenu s'il avait fourni son véritable profil de préférence.

Une influence qui est souvent contre l'intérêt de l'agent manipulé Lorsqu'un agent manipule d'autres agents dans un système de vote, il le fait pour son intérêt propre, et cela va souvent à l'encontre des intérêts propres des autres agents du système. (Goodin, 1980) ajoute que la manipulation est avant tout une influence trompeuse contraignant à agir contre sa volonté. Pour (Barnhill, 2014) la manipulation est aussi le fait d'influencer intentionnellement certains traits de caractères ou dispositions psychologiques dans le but d'amener la victime dans des idéaux de croyances, de désirs ou d'émotions et d'une manière qui ne va généralement pas dans son intérêt personnel ou qui n'est probablement pas dans son intérêt personnel dans le contexte actuel. Toutefois, la manipulation ne va pas toujours contre l'intérêt propre de l'agent manipulé. Par exemple, l'effet placebo est parfois utilisé en médecine pour faire croire à un patient qu'il va guérir

avec un médicament inefficace. Il s'agit d'une manipulation dans l'intérêt du patient (Turner et al., 1994).

Perte d'autonomie La manipulation est parfois caractérisée comme une perte d'autonomie. Par exemple, lorsqu'un parasite prend le contrôle de son hôte (Poulin, 2010). Cependant, (Todd, 2013) présente une expérience de pensée dans laquelle des neuroscientifiques implantent une puce dans la tête d'un patient leur permettant de contrôler tous ses faits et gestes telle une marionnette. Même si ces scientifiques manipulent le patient comme le ferait un ingénieur manipulant sa machine, Todd affirme qu'il serait complètement incongru de dire qu'un ingénieur est un parfait manipulateur et donc que ces scientifiques sont des manipulateurs. Pour lui, même si la manipulation se caractérise comme une perte d'autonomie, cela ne suffit pas pour définir la manipulation et il est donc nécessaire de distinguer la manipulation « *per se* » comme manipuler un objet, de la manipulation « par cas » et selon laquelle des agents agissent de façon manipulatrice. Selon Todd, c'est sur ce dernier sens que les philosophes doivent porter toute leur attention. Par conséquent, même si la perte d'autonomie ne suffit pas à elle seule pour définir la manipulation, elle est nécessaire pour que la manipulation ait lieu. Cette perte d'autonomie peut aussi être caractérisée par le fait qu'un manipulateur contrôle, par influence les croyances de la victime, mais aussi ses désirs et ses émotions.

La manipulation est une intention délibérée (Noggle, 1996) considère que la manipulation est l'intention d'amener une personne sur certains idéaux de croyances, de désirs, et d'émotions, mais qui sont gouvernés du point de vue du manipulateur. Il affirme que la manipulation est avant tout une intention d'agir sur un autre agent sous la forme d'une influence. Cette influence est alors dite manipulatrice si elle est délibérée à l'avance par le manipulateur car l'influence n'est pas une action manipulatrice tant qu'elle est sincère, c'est-à-dire en accord avec ce que l'influenceur prend pour vrai, pertinent et approprié. Noggle n'est pas le seul à insister sur le fait que la manipulation est une intention délibérée. Cela rejoint la vision des psychiatres Kligman et Culver qui soulignent le fait que la manipulation est nécessairement intentionnelle (Kligman and Culver, 1992). En effet, lorsque nous sommes en train de manipuler, l'acte a été sciemment délibéré. L'utilisation du concept de « manipulation non intentionnelle » est fortement rejeté. Pour eux, une personne appliquant un mécanisme d'influence, sans le savoir, et donc sans en avoir eu l'intention, ne peut être considéré comme une manipulation.

L'instrumentalisation de l'autre dans la manipulation Comme le rappellent (Kligman and Culver, 1992) le terme « manipulation » est parfois utilisé comme un simple synonyme d'utiliser quelqu'un pour quelque chose dans son intérêt propre, ou comme un stratagème ayant pour intention d'extraire une réponse précise d'un individu. Mais pour (Clair, 1966), la manipulation se définit avant tout comme une manœuvre psychosociale qui utilise l'agressivité, l'intelligence et la tromperie pour influencer quelqu'un dans le but qu'il accomplisse un désir du manipulateur. Pour qu'il y ait une manipulation, il est nécessaire qu'il y ait une incompatibilité entre les désirs du manipulateur et ceux du manipulé. Mais cette définition de la manipulation ne fait pas encore l'unanimité chez les psychiatres. En effet, pour le psychiatre Bowers la manipulation est avant

tout une activité qui vise « à atteindre un but souhaité (pervers, normal, symbolique ou réel) par la tromperie, la coercition et la ruse, sans tenir compte des intérêts ou des besoins des personnes utilisées dans le processus » (Bowers, 2003). Il considère que la manipulation est l'activité qui vise à réaliser un but désiré en usant de la tromperie, de la coercition ou des ruses sans tenir compte des intérêts ou des besoins de ceux impliqués dans ce processus de manipulation. Cependant pour ces chercheurs, l'instrumentalisation est un point fondamental associé à la manipulation.

La manipulation vue comme l'altération du jugement d'autrui

Cette instrumentalisation peut se dérouler par l'exécution de méthodes psychologiques permettant d'altérer le jugement de la victime. C'est pourquoi certains chercheurs considèrent la manipulation comme une méthode consistant à court-circuiter des processus cognitifs, d'autres la considèrent plutôt comme de la persuasion non-rationnelle, ou comme l'exploitation d'une faiblesse dans les agents manipulés, et enfin, les chercheurs qui considèrent la manipulation comme une distorsion de la réalité.

Court-circuiter des processus cognitifs Une des méthodes psychologiques pour altérer le jugement de la victime vise à court-circuiter ses processus cognitifs. (Mills, 1995) énonce que la manipulation est le but d'effectuer un changement dans les états internes de l'autre ; la manipulation s'opère sur nos croyances et nos désirs - mais surtout sur nos émotions - et ce, d'une manière indirecte. La manipulation fait qu'une personne a des désirs ou des croyances qui ne sont pas venues de manière naturelle par le biais de ses croyances et désirs antérieurs ; mais qui ont été produits d'une manière à contourner ses processus cognitifs ordinaires ainsi que ses processus affectifs. Par exemple (Wilkinson, 2013) décrit la manipulation comme une sorte d'influence qui contourne ou subvertit les capacités rationnelles de la victime. De plus, selon lui, la manipulation est une intention de parvenir à influencer quelqu'un en utilisant des méthodes qui pervertissent son choix. En linguistique, cette vision de court-circuiter des processus cognitifs dans la manipulation est aussi présente. Pour (Maillat and Oswald, 2009), un discours manipulateur est une forme de communication qui met le destinataire dans une situation où il sera amené à faire des hypothèses contextuelles superficielles. Par hypothèses contextuelles superficielles, celui-ci entend que la stratégie manipulatrice repose alors sur le déplacement de l'attention du destinataire, en disant A (ce qui est vrai), le manipulateur va chercher à cacher la vérité d'une proposition B qui affaiblit ou contredit la vision entretenue par le présentateur. La manipulation constitue alors pour lui une forme de contrainte cognitive sur le choix de ces hypothèses contextuelles.

La manipulation peut être une forme de persuasion non-rationnelle Plutôt que de voir la manipulation comme le fait de court-circuiter des processus cognitifs, certains chercheurs préfèrent voir la manipulation comme une forme de persuasion non-rationnelle qui amène, par exemple, une victime à prendre une décision de façon réactive et qui n'est pas nécessairement dans son propre intérêt comme nous avons pu le voir précédemment. En réalité, chez des agents humains, (Kahneman, 2011) explique qu'il existe deux types de systèmes cognitifs. Le système 1

qui fonctionne rapidement, est une sorte de pilote automatique. Celui-ci est par exemple animé par les habitudes d'une personne, ou par ses instincts. Ce système est facile à manipuler pour un agent averti. Ce système est un ordonnateur pas un planificateur. Pour faire simple, ce système permet à l'agent d'agir de façon réactive. En revanche, le système 2 est réfléchi et délibératif. Ce système évalue la situation et délibère sur une situation donnée. Ce système est lent, contrairement au système 1 qui lui est réactif, mais ce système est plus difficilement manipulable et demande pour un manipulateur de considérer les états mentaux dans lequel le manipulé doit se trouver.

Exploitation d'une faiblesse chez l'autre agent Les systèmes 1 et 2 peuvent être exploités par des agents malintentionnés pour manipuler d'autres agents humains, notamment en exploitant les biais cognitifs. Les biais cognitifs sont des distorsions de l'information au moment de leur traitement cognitif et correspondent à une déviation systématique de la pensée logique et rationnelle par rapport à la réalité. Ils peuvent être perçus comme une réelle faiblesse dans le raisonnement humain (Kahneman, 2011). Ainsi dans le langage courant, un agent exploite un état d'ignorance ou de vulnérabilité psychique comme les biais cognitifs, et ce pour amener un autre agent à prendre des engagements dont il est incapable de voir l'importance, est considéré comme un abus de faiblesse. Cet abus de faiblesse est parfois défini comme de la manipulation. Pour (Rudinow, 1978) la manipulation se décrit comme le fait qu'un agent « A tente de manipuler [un agent] S si, et seulement si, A influence les motivations complexes du comportement de S, par le biais de la tromperie, ou en jouant sur une faiblesse supposée de S ». Par ailleurs, Rudinow fait la distinction entre une tentative de manipulation sans réussite et le fait de ressentir d'avoir été manipulé sans pour autant avoir réellement été manipulé. La manipulation est cette tentative réussie de manipuler quelqu'un et ce, en dépit du fait de ressentir d'avoir été manipulé ou d'avoir la croyance de l'avoir été ou encore d'être, à l'instant, l'objet d'une tentative réussie de manipulation.

Une distorsion de la réalité La manipulation est parfois considérée comme intention de déformer la réalité. Pour le linguiste Eddo Rigotti un message est manipulateur s'il déforme la vision du monde, aussi bien physique que social ou humain, réel ou virtuel dans l'esprit de l'adressé, de sorte qu'il ou elle est empêché(e) d'avoir une attitude saine envers sa propre décision (c'est-à-dire à adopter une attitude répondant à son intérêt même), et poursuit le but du manipulateur dans l'illusion de poursuivre son propre but. Il est nécessaire que le manipulateur lui dissimule ses réelles intentions (Rigotti, 2005). Pour (Faden and Beauchamp, 1986), la manipulation est tout acte réussi et intentionnel d'influencer les croyances d'un agent ou son comportement en provoquant des changements dans les processus mentaux autres que ceux impliqués dans leur compréhension. De plus, ils identifient trois types de stratégies d'influences manipulatrices :

1. Influencer en modifiant les options disponibles dans l'environnement : en augmentant ou diminuant les options disponibles pour les autres agents ;
2. Influencer en offrant une récompense ou en menaçant d'une punition l'agent ciblé ;
3. Influencer directement les états mentaux : croyances, connaissances, désirs, intentions.

(Baron, 2003) ajoute à cet ensemble de stratégies d'influence l'application de la pression sociale et manipulation de la situation afin de limiter artificiellement les options de l'agent manipulé. Cependant pour (Raz, 1986), qui n'est pas d'accord cela suggère plutôt que la manipulation, contrairement à la coercition, n'interfère pas avec les options de la personne. À la place, elle pervertit la manière dont la personne va prendre ses décisions, va former ses préférence ou adopter des buts. La manipulation porte donc atteinte à l'autonomie de la victime en subvertissant et en insultant son pouvoir de décision. Cette vision de la manipulation rejoint alors la vision de (Wilkinson, 2013) pour qui la manipulation est une influence qui contourne ou subvertit les capacités rationnelles de la victime et ce en utilisant des méthodes qui pervertissent son choix.

La manipulation vue comme l'exercice invisible d'un pouvoir

(Van Dijk, 2006) définit le concept de manipulation comme une forme de contrôle mental dans laquelle le manipulé n'est pas conscient de ce qui est en train de se passer ou ne peut tout simplement rien faire en conséquence. Pour Van Dijk, une stratégie de manipulation doit impérativement échapper à la conscience du manipulé. Dans le cadre clinique, la manipulation est alors associée aux efforts employés par un patient d'utiliser des moyens cachés afin d'obtenir le contrôle ou le support de personnes significatives (Gunderson, 1984). Gunderson fait alors allusion aux plaintes somatiques, aux actions provocatrices ou aux messages trompeurs, ainsi qu'aux actes autodestructeurs que certains patients peuvent avoir envers le personnel soignant. Le linguiste Akopova a ensuite repris ces définitions en considérant que la manipulation est une intention négative du locuteur et avec un caractère d'influence caché (Akopova, 2013). Une manipulation est réalisée quand l'interlocuteur ne peut plus voir les intentions du locuteur derrière ce qu'il affirme. À cela (Handelman, 2009) ajoute que la signification pratique de la manipulation est bien que la cible soit soumise à une influence cachée et que celle-ci croit que ses choix sont faits librement et indépendamment. La manipulation vise donc à motiver la cible à opérer sous une forme qui, dans des conditions normales, l'aurait amené probablement à résister ou rejeter l'interaction. Handelman insiste alors sur le fait que cette ingérence doit être avant tout indirecte, invisible et secrète pour pouvoir s'opérer. Pour (de Saussure and Schulz, 2005), elle peut passer par le fait de retenir des informations ou bien, au contraire, de donner certaines informations correctes dans le but que le destinataire conclut qu'il devrait agir dans la direction qui est en faveur du manipulateur, et ce sans que celui-ci soit conscient de cette stratégie. Pour (Ware, 1981), il est clair que la manipulation se définit comme une sorte d'influence cachée. Alan Ware explique que manipuler quelqu'un revient à structurer son environnement avec l'intention d'en changer ses choix, et pour y parvenir, sans que la victime ne sache ou ne comprenne ce qui est en train de se produire. Par ailleurs, il distingue d'une part la manipulation qui cible une personne ou influence directement la personne ; et d'autre part, de la manipulation qui consiste à modifier une situation ou changer les options disponibles pour une personne. De la même manière, (Sunstein, 2015) considère qu'une déclaration ou une action est manipulatrice dans la mesure où elle ne fait pas appel à la capacité de réflexion et de délibération des manipulés.

1.2.2 Vers une définition générale de la manipulation

Toutefois, même si nous avons mis en lumière des divergences notables entre les chercheurs, il semblerait que ces chercheurs soient en accord sur certains points. Ainsi, dans la suite, nous utilisons ces points communs pour considérer une définition générale de la manipulation. Nous avons donc pu constater que dans de nombreuses approches des sciences humaines de la manipulation, une des caractéristiques principales de la manipulation est qu'elle est d'une part l'exercice d'un pouvoir sur autrui et que d'autre part, cet exercice du pouvoir est dissimulé aux agents manipulés.

La manipulation ne peut pas être simplement réduite à de la tromperie ou à de l'influence. De plus, celle-ci est bien distincte de la persuasion et de la coercition. La manipulation est une intention délibérée d'influencer autrui tout en veillant à lui dissimuler cette intention.

Pour l'ensemble des chercheurs, la manipulation n'est pas exactement de la coercition, ni vraiment de la persuasion, ni complètement similaire à la tromperie (Handelman, 2009; Kligman and Culver, 1992). Elle est quelque chose qui arrive de façon complètement invisible, et au moment où nous commençons à parler de manipulation, l'acte a déjà été commis (Goodin, 1980). Ainsi, quand nous parlons de manipulation, que ce soit au passé (« j'ai été manipulé ») ou à la seconde personne (« tu as été manipulé ! »), nous sommes bien en train de déclarer quelque chose que la victime ne savait pas. Pour (Handelman, 2009), la conclusion inévitable est que la cible est nécessairement dans l'incapacité d'identifier qu'elle a été sujette à une influence manipulatrice.

Ainsi, pour de nombreux chercheurs (Handelman, 2009; Akopova, 2013; Todd, 2013; Cohen, 2017), la manipulation est tout d'abord une intention volontaire du manipulateur d'utiliser la victime pour accomplir quelque chose. Nous parlons alors *d'instrumentalisation*[9] de la victime. De plus, cette instrumentalisation est dissimulée à la victime. Le manipulateur, au cours de la manipulation a vraiment l'intention que la victime ait l'impression de choisir librement et indépendamment ses actions mais que le manipulateur en sait toujours bien plus sur la vraie raison de ses croyances, de ses désirs et de ses intentions (Handelman, 2009). La manipulation est nécessairement intentionnelle et l'utilisation du concept de « manipulation non intentionnelle » est discutée et fortement refusée (Kligman and Culver, 1992). En effet, nous ne pouvons parler de manipulation si le soit disant manipulateur n'a pas délibéré sur le fait de manipuler une victime. Ainsi, un agent qui influence à son insu un autre agent sans qu'il en soit conscient ne peut pas être en train de manipuler. En effet, nous constatons par exemple que l'influence est omniprésente dans chacune des interactions inter-agents et il n'est pas pertinent de considérer que tout est manipulation. Ainsi pour faire cette distinction entre la simple influence, dans laquelle l'influencé n'a pas non plus conscience d'être influencé, il est nécessaire que dans un cas de manipulation, le manipulateur ait eu cette véritable intention d'influencer l'autre sans qu'il ne s'en rende compte. Nous parlons *d'intentions délibérées*.

9. Remarquons que contrairement à l'influence qui peut porter sur des croyances, des connaissances ou même des intentions, nous faisons le choix de considérer l'instrumentalisation comme une influence ne portant que sur les intentions des agents victimes ou le fait que la victime amène quelque chose de vrai à son insu.

En résumé, la manipulation possède donc trois caractéristiques fondamentales :

— la manipulation est *une intention délibérée du manipulateur* ;

— la manipulation est *une instrumentalisation exercée sur une victime* ;

— la manipulation est *toujours dissimulée à la victime.*

Ces trois caractéristiques fondent la définition suivante de la manipulation :

Définition 1.4 - *Manipulation* : *Nous appelons* manipulation*, l'intention délibérée d'un agent d'instrumentaliser un autre agent, tout en veillant à lui dissimuler cette intention de l'instrumentaliser.*

Si cette définition de la manipulation n'a encore jamais été considérée dans les systèmes multi-agents, des travaux ont étudié des notions proches de la manipulation comme la tromperie, ou encore sur la construction de stratégies de manipulation dans les SMA.

1.2.3 Stratégies de manipulation dans les systèmes multi-agents

La manipulation est donc définie comme l'intention délibérée d'instrumentaliser un agent tout en veillant à lui dissimuler cette intention de l'instrumentaliser. Cependant, une stratégie de manipulation est, quant-à-elle, définie comme un ensemble d'actions réalisées par un ou plusieurs agents amenant à la réalisation d'une manipulation [10]. Par exemple, le fait de cacher une information dans le but de garder le contrôle sur une personne est une stratégie de manipulation qui repose sur les connaissances de cette personne. Cette section présente des travaux qui ont mis en lumière, d'une part, plusieurs taxonomies de stratégies de manipulation en lien avec la tromperie, et d'autre part, nous présentons un ensemble d'exemples de stratégies de manipulation dans différents SMA comme dans les systèmes à base de jeux, les systèmes de vote, les systèmes d'enchères, ou encore les systèmes de réputation, puis, nous présentons une vision de l'ensemble des bases de stratégies de manipulation pouvant exister dans les SMA.

Taxonomies de la tromperie dans les SMA

Dans les systèmes multi-agents, la tromperie se définit comme l'intention d'induire en erreur en dissimulant des informations ou en révélant de fausses informations à d'autres agents. Elle se caractérise donc comme une intention de dissimuler plutôt que d'instrumentaliser. Dans les réseaux pair-à-pair, des agents peuvent tromper les autres agents du système dans le but de propager des malwares au sein du réseau sans pour autant chercher à prendre le contrôle des agents du système (Fox, 2001). Dans la tromperie, il n'est donc pas nécessairement question d'instrumentalisation, et donc, la tromperie n'implique donc pas de la manipulation. En revanche puisque la manipulation est une intention d'instrumentaliser tout en veillant à dissimuler cette

10. Dans ce manuscrit, nous parlons souvent de « manipulations » au pluriel pour évoquer les stratégies de manipulation tandis que nous parlons de « la » manipulation pour désigner « l'intention délibérée d'instrumentaliser l'autre agent sans qu'il ne s'en rende compte ». Il y a alors une (et unique) manipulation au singulier mais des stratégies de manipulation. Ainsi, l'utilisation au pluriel de la manipulation désigne en réalité les stratégies de manipulation et non la manipulation.

Dimensions	**Catégories**	**Descriptions**
Objet de l'interaction	Tromperie Robot-Humain (H)	Le robot trompe ses partenaires humains
	Tromperie Robot-Non humain (N)	Le robot trompe d'autres robots, des animaux, etc.
But de l'interaction	Orienté vers soi (S)	Le robot trompe par bénéfices personnel
	Orienté vers les autres (O)	Le robot trompe pour les bénéfices de l'agent trompé
Type de l'interaction	Physique (P)	La tromperie passe par l'aspect physique du robot, sans aucune nécessité d'utiliser les processus cognitives élevés
	Comportementale (B)	La tromperie passe par les représentations des états mentaux, le comportement et des processus cognitifs élevés

Tableau 1.1 – Dimensions de la tromperie pour les agents artificiels (Shim and Arkin, 2013)

instrumentalisation, la manipulation implique nécessairement de la tromperie. Un autre exemple de stratégie de manipulation dans les systèmes de réputation, qui permettent d'évaluer la qualité d'un produit en tenant compte des témoignages des différents agents, un vendeur peut tromper sur la qualité de son produit en intégrant de fausses identités et s'auto-promouvoir afin d'inciter de nouveaux clients à venir acheter ses produits (Douceur, 2002). L'intention du vendeur est bien d'influencer ses clients en les amenant à acheter son produit sans même qu'ils ne se rendent compte du stratagème. Le vendeur manipule donc les autres agents.

(Shim and Arkin, 2013) soulignent le fait que puisque la tromperie est si fréquente dans le comportement des humains, il n'y a aucune raison pour lesquelles nous ne pourrions pas aussi la rencontrer chez des agents artificiels. De ce fait, ils proposent alors une taxonomie sur la tromperie entre humains-robots et robots-robots en distinguant trois dimensions dans la tromperie illustrée par la figure 1.1 :

L'objet de l'interaction : le type d'agent dans l'interaction, qu'il soit artificiel ou humain ;

Le but de l'interaction : les raisons pour lesquelles un agent peut désirer tromper, qu'elles soient pour lui ou pour l'autre agent ;

Le type de l'interaction : un agent peut tromper par son physique (ex. le camouflage) ou par des processus cognitifs.

Pour (Whaley, 1982), il existe deux catégories de tromperie : les tromperies de type dissimulatrices, i.e. celles qui cachent le réel ; de celles qui simulent en montrant le faux. Chacune de ces catégories se divisent en trois sous-catégories résumées dans la figure 1.2. Dans la mise en place d'une tromperie, certaines de ces techniques peuvent être combinées entre elles, par exemple utiliser l'éblouissement et faire passer cela pour un leurre.

Catégories	**Techniques**	**Descriptions**
Dissimulation	Masquage	Consiste à éviter la détection en jouant un double rôle, ou en cherchant à rendre invisible le motif de la tromperie
	Reconditionnement	Consiste à cacher le réel en faisant transparaître quelque chose pour autre chose
	Éblouissement	Consiste à attirer l'attention sur un ensemble de choses afin de dissimuler la tromperie
Simulation	Mimétisme	Consiste à attirer la victime en simulant un certain comportement
	Invention	Consiste à inventer une fausse information
	Leurre	Consiste à mettre en place un leurre afin de créer une diversion

Tableau 1.2 – Techniques de la tromperie (Whaley, 1982)

Ces différentes taxonomies précisent ce que nous considérons comme de la tromperie dans les SMA et nous éclairent sur différentes manières dont un agent peut s'y prendre pour tromper les autres agents du système. Par ailleurs, si des auteurs comme Shim et Arkin ou Bell et Whaley établissent une taxonomie de la tromperie ; d'autres auteurs présentent des stratégies de manipulation dans certains SMA comme c'est le cas dans les systèmes à base de jeux, les systèmes de vote, ou encore les systèmes de réputation.

Exemples de stratégies de manipulation dans les SMA

Dans cette section, nous présentons des travaux qui ont étudié la construction de stratégies de manipulation dans des systèmes multi-agents comme, par exemple, les systèmes à base de jeux, les systèmes de vote, ou encore les systèmes de réputation [11].

Systèmes à base de jeux La theorie des jeux s'intéresse aux interactions entre agents et se fonde sur l'hypothèse que les agents sont rationnels, c'est-à-dire que les agents vont toujours chercher à maximiser leur fonction de récompense personnelle ou collective en fonction des décisions qu'ils peuvent prendre. À partir de la théorie des jeux, Ettinger et Jehiel fondent une théorie de la tromperie (Ettinger and Jehiel, 2010). Un agent va chercher à tromper les autres agents s'il a le plus d'intérêts à le faire et ce en utilisant un ensemble de connaissances qu'il possède sur la manière dont les autres agents raisonnent. (Wagner and Arkin, 2009; Wagner and Arkin, 2011) utilisent la théorie des jeux et plus particulièrement, le dilemme du prisonnier pour représenter une situation dans laquelle un agent artificiel a tout intérêt à tromper un autre agent. Dans un dilemme du prisonnier, deux agents ont deux actions possibles : dire la vérité ou mentir. Si les deux agents mentent alors ils gagnent une certaine récompense, si l'un ment et l'autre dit la

11. Par soucis de concision, nous n'avons pas la prétention d'être exhaustif et ne présentons que quelques exemples de systèmes multi-agents vulnérables aux stratégies de manipulation. En effet, d'autres systèmes sont vulnérables comme les systèmes de recommandation (Bhaumik et al., 2006; Zhang and Zhou, 2012).

vérité celui qui ment perd, si les deux agents disent la vérité, ils perdent moins que si un agent disait la vérité et l'autre mentait. Un agent peut donc avoir un intérêt rationnel de mentir s'il sait qu'il va maximiser ses gains.

Systèmes de vote et d'enchères La théorie du choix social s'intéresse à la représentation formelle des systèmes de vote pour permettre à des agents d'effectuer des prises de décisions collectives. Elle s'intéresse aussi aux stratégies de manipulation dans les systèmes de vote et à leur robustesse (Gibbard, 1973). Dans les enchères de Vickrey (Sanghvi and Parkes, 2004; Parkes and Ungar, 2000), qui consistent pour chaque agent de proposer une valeur à un objet dont chacun des agents du système désirent, un agent va remporter l'objet si celui-ci propose la valeur la plus élevée, mais ce dernier paiera la somme de la valeur du deuxième agent de l'enchère. Une stratégie de manipulation de ces enchères consiste à fonder une collusion entre agents où tous les agents révèlent aux autres agents leur prix annoncé pour un objet (Robinson, 1985). Une fois leurs prix révélés, les agents conservent l'ordre donné. Ainsi, chaque agent peut baisser sa valuation afin de garantir à l'agent gagnant le prix le plus bas pour l'objet. Une autre stratégie de manipulation pour un vendeur est de créer une fausse identité afin d'enchérir sur les autres offrants afin de maximiser le bénéfice réalisé sur l'objet (Ausubel et al., 2006).

Systèmes de réputation Les systèmes de réputation présentés à la section 1.1.3 peuvent être utilisés à des fins malhonnêtes par des agents. Certains de ces agents peuvent jouer sur la confiance que leur accordent les autres agents pour les instrumentaliser sans qu'ils ne s'en rendent compte. (Vallée, 2015) propose une taxonomie, représentée par la figure 1.5, sur les différentes stratégies de manipulation dans ces systèmes. Il distingue de ce fait deux classes de stratégies de manipulation dans les systèmes de réputation. Les stratégies de manipulation explicites consistent à partager explicitement des informations avec la victime afin de l'induire en erreur. Tandis que les stratégies de manipulation implicites consistent à interagir avec le système afin que les autres agents déduisent de leurs observations de fausses connaissances. Parmi les stratégies de manipulation explicites, celui-ci distingue encore deux catégories : les stratégies qui vont reposer sur les informations privées qu'un agent va transmettre aux autres, que ce soit son profil de préférences ou ses témoignages sur les autres agents ; des stratégies qui reposent sur les informations publiques (identités) comme par exemple le fait de mentir sur son identité (usurpation) ou le fait de recommencer avec une nouvelle identité dans le système (blanchiment). Les manipulations implicites incluent deux catégories de stratégies de manipulation : les stratégies comportementales et les stratégies qui reposent sur un positionnement stratégique. Par exemple, lorsqu'un agent oscille entre un comportement malhonnête et un comportement honnête pour induire en erreur les autres agents du système (oscillation), il s'agit d'une stratégie comportementale. La traîtrise est aussi une stratégie comportementale qui consiste tout d'abord à obtenir la confiance des autres agents du système puis à agir de façon malhonnête avec eux. Enfin, il considère les manipulations par positionnements stratégiques, les attaques comme les dénis de service [12] pour faire croire aux autres agents qu'un service proposé n'est pas fiable, ou encore les

12. Une attaque par déni de service ou attaque DoS (Wood and Stankovic, 2002; Specht and Lee, 2004) consiste à surcharger un serveur par un nombre très important de requêtes. Ces attaques peuvent avoir pour finalité de

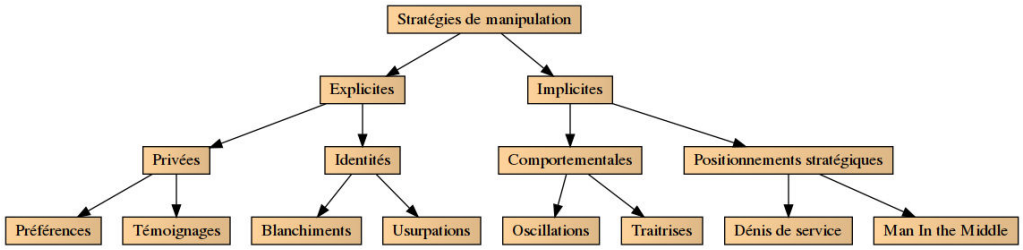

FIGURE 1.5 – Taxonomie des manipulations dans les SR (Vallée, 2015)

attaques de l'homme du milieu où un agent va par exemple intercepter des témoignages entre deux agents et falsifier ces derniers.

De manière générale, dans les systèmes multi-agents, nous pouvons regrouper ces stratégies de manipulation en six catégories présentées dans le tableau de la figure 1.3. Dans la suite de cette section, nous présentons ces différentes catégories.

Exploiter une faille ou une faiblesse

Lorsqu'un agent manifeste un certain comportement, un manipulateur peut exploiter ce comportement à ses propres fins. En psychologie sociale, Joule et Beauvois ont écrit *Un petit traité de manipulation à l'usage des honnêtes gens* (Joule et al., 2002) dans lequel ils présentent la théorie de la soumission librement consentie. Cette théorie décrit un ensemble de techniques de manipulations comportementales. Parmi celles-ci, l'amorçage consiste à demander à une victime un service simple, comme par exemple, demander l'heure. Il s'avère qu'après avoir obtenu ce premier service, il est plus facile de demander un service plus contraignant.

Cette faille dans le comportement humain trouve aussi son analogue en informatique, comme par exemple, lorsqu'un pirate informatique exploite un débordement de tampon dans une application logicielle et injecte du code pour prendre la main sur le système. (Erickson, 2008) présente un ensemble de failles dans les applications logicielles qu'un pirate peut exploiter et utiliser comme stratégies de manipulation du système. Le pirate informatique détourne le comportement d'un agent logiciel pour lui faire réaliser une fonctionnalité que l'agent logiciel ne désirait pas. Il parvient à contrôler le système d'une manière initialement non prévue sans même que le système ne se rende compte de cette anomalie. Nous pouvons donc parler de stratégie de manipulation.

Détourner une norme

Un agent manipulateur peut détourner une norme sociale (cf. section 1.1.3) à son profit. En psychologie sociale, (Cialdini, 2001) présente un ensemble de normes sur lesquelles un manipulateur peut s'appuyer pour pousser un autre agent à réaliser son désir. Par exemple, la norme de *réciprocité* consiste à toujours rendre la pareille aux agents nous offrant un service. Remarquons qu'une norme de réciprocité est fondamentale dans les réseaux pair-à-pair (Fox, 2001) puisque

faire tomber un service.

Bases de la stratégie	**Descriptions**	**Exemples dans les SMA**
Exploiter une faille ou une faiblesse	L'agent manipulateur exploite un certain comportement d'un agent pour l'instrumentaliser.	Un pirate informatique qui exploite une faille système pour prendre le contrôle du système (Erickson, 2008)
Détourner une norme	L'agent utilise un comportement collectif pour instrumentaliser un agent	Offrir gratuitement un service dans un réseau pour obtenir un avantage (Fox, 2001)
Abuser de la confiance	L'agent joue sur la confiance des autres agents pour les instrumentaliser	Falsifier son identité pour se promouvoir dans un système de réputation (Ruan and Durresi, 2016)
Profiter de la rationalité	L'agent utilise la rationalité des agents pour les instrumentaliser	Fournir un faux profil de préférence permet à un agent d'influencer une situation en sa faveur lorsqu'il considère les autres agents comme rationnels(Vallée et al., 2014)
S'appuyer sur les émotions	L'agent s'appuie sur les émotions d'un autre agent pour le contrôler	Imiter les émotions permet à des agents d'augmenter leur chance d'influencer un humain (Zawieska, 2015)
Jouer sur les connaissances et croyances	L'agent utilise ses connaissances ou les croyances de l'autre agent pour lui dissimuler sa stratégie	La tromperie, le mensonge ou le baratinage peuvent être des stratégies de manipulation dans lesquelles un agent joue sur les croyances (Sakama et al., 2015).

Tableau 1.3 – Résumé des différentes bases de stratégies de manipulation.

ces systèmes reposent sur le partage des ressources. Les resquilleurs[13], qui sont des agents qui ne rendent pas la pareille et ne partagent pas les ressources qu'ils récupèrent des autres agents, doivent donc être exclus du système. A contrario, un agent manipulateur peut détourner cette norme en offrant volontairement un service ou une ressource à un agent dans le but que celui-ci lui rende la pareille, en lui offrant le service demandé par l'agent manipulateur.

Cependant, même si un grand nombre de chercheurs se sont intéressés à la formalisation de systèmes normatifs en intelligence artificielle (Castelfranchi et al., 1998; Ågotnes et al., 2007; Herzig et al., 2011; Knobbout and Dastani, 2012), il n'existe que très peu de travaux dans les SMA consacrés à l'utilisation des normes dans le but de manipuler les autres agents du système. Les seuls travaux s'en rapprochant sont issus de la théorie des jeux et modélisent la norme de réciprocité présentée précédemment en cherchant à évaluer l'intérêt d'un agent à rendre ou ne pas rendre la pareille (Cox, 2004; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006).

Abuser de la confiance

Pour (Castelfranchi and Falcone, 2010), la confiance est le pilier de la société (cf. section 1.1.3). C'est aussi un outil de manipulation. En effet, une stratégie de manipulation consiste à construire de la confiance afin de mieux contrôler les autres agents du système. En marketing, la confiance est un outil pour amener un client à acheter un produit dont il n'a pas besoin et sans qu'il n'en ait conscience (Calo, 2013). Par exemple, des vendeurs peuvent utiliser un principe dit de *rareté* (Cialdini, 2001) pour donner l'illusion à un client qu'un produit est de qualité parce qu'il n'en reste plus beaucoup dans les stocks.

Dans les systèmes multi-agents, une des techniques privilégiée pour déterminer si un agent est de confiance ou non est l'utilisation d'un système de réputation (Ruan and Durresi, 2016) comme présentée à la section 1.1.3. Dans ce type de système, un agent manipulateur peut, par exemple, profiter d'une haute réputation pour flouer les autres agents, ce qui est appelée *traîtrise* (Such, 2013). La stratégie de manipulation dans la traîtrise consiste pour le manipulateur à construire de la confiance pour amener les agents à réaliser son but sans que ces agents ne se doutent que l'agent manipulateur a l'intention de les trahir à la fin de son plan. La manipulation n'a donc pas lieu au moment où les agents réalisent qu'ils ont été trompés mais au moment même où le manipulateur met en place sa stratégie.

Jouer sur les connaissances et les croyances

Puisque dans toute manipulation, il y a une dissimulation, il est donc nécessaire d'adopter une stratégie empêchant l'autre agent de croire ou de savoir qu'il est manipulé . Une manipulation doit donc nécessairement jouer sur les connaissances et les croyances des agents. Le linguiste Eddo Rigotti présente une typologie des principaux processus permettant à un agent manipulateur d'induire en erreur un autre agent en jouant sur ses croyances et ses connaissances (Rigotti, 2005). Par exemple, le sophisme est une technique qui consiste à jouer sur le fait qu'un agent croit que A implique B. Le manipulateur sait que A implique B est faux, mais fait croire que A

13. *Free-riders* en anglais.

est vraie afin que l'agent manipulé déduise B. L'agent manipulateur exploite de façon délibérée l'erreur de l'agent manipulé de croire que A implique B et amène donc l'agent manipulé à croire B tout en veillant à ce que cet agent ne sache pas la stratégie du manipulateur d'utiliser son erreur de raisonnement.

Le mensonge est représenté dans les systèmes multi-agents comme l'intention d'un agent, appelé menteur, de faire croire, en communiquant à un autre agent, une information que l'agent menteur ne croit pas comme vraie (Sakama et al., 2015). Le mensonge peut être vu comme une stratégie de manipulation dès lors que l'agent menteur a pour intention de contrôler l'autre agent par le mensonge et ce, sans que l'agent victime ne s'en aperçoive. Il repose donc sur les croyances et les connaissances de l'agent victime, et l'empêche d'accéder à la vérité.

Profiter de la rationalité

La rationalité est un comportement des agents à vouloir toujours maximiser ce qu'ils peuvent espérer gagner sur une situation donnée. Elle peut être utilisée pour influencer des décisions sans même que l'agent soit conscient de cette influence. Même si cette notion de rationalité est débattue dans le cadre de modèles économiques appliqués pour modéliser le comportement des humains (Kahneman, 2011). Elle peut être utilisée, par exemple, par des gouvernements pour effectuer des choix géopolitiques et stratégiques. (Handelman, 2009) évoque la guerre d'Irak annoncée en 2003 par le président W. Bush, comme un cas de guerre rationnelle avec des intérêts géopolitiques et stratégiques, permettant au président de gagner en popularité par l'instrumentalisation de cette guerre. La rationalité permet donc à des agents de maintenir le contrôle sur une situation ou une population dans un système sans même que ces agents ne sachent que cette stratégie a pour finalité de maintenir un contrôle.

Par ailleurs, cette hypothèse de rationalité est utilisée en théorie des jeux pour influencer les décisions des autres agents comme nous l'avons présenté en section 1.2.3. Par exemple, dans le cadre des jeux hédoniques qui sont des jeux dans lesquels les agents ont chacun un profil de préférences sur un ensemble de coalitions qu'ils désirent former. (Vallée et al., 2014) étudie le cas de certaines stratégies de manipulation, appelées *attaques Sybil*. Dans ces systèmes, les attaques Sybil consistent à intégrer de faux agents dans le système pour permettre à un agent manipulateur d'obtenir une meilleure coalition que celle qu'il pourrait initialement obtenir sans cette stratégie. Nous parlons aussi de stratégie de manipulation puisqu'un manipulateur veille de façon délibérée à ne jamais révéler sa stratégie pour influencer la décision collective retenue par le système.

S'appuyer sur les émotions

Les manipulations entre humains peuvent reposer sur une base émotionnelle. Il s'agit pour un agent manipulateur de simuler certains états émotionnels comme par exemple lorsqu' un enfant pleure pour obtenir un nouveau jouet auprès de ses parents. L'enfant veille de façon délibérée à se mettre dans cet état pour affecter l'état émotionnel des parents afin d'obtenir ce qu'il désire. En psychiatrie, cette stratégie de manipulation basée sur les émotions trouve de nombreux exemples comme lorsqu'une personne menace de se suicider. Celle-ci le fait toujours

dans le but d'obtenir quelque chose d'autrui : une écoute, un service, espérer se remettre en couple avec une personne (Gunderson, 1984).

Pour des agents artificiels, il est aussi envisageable que des stratégies de manipulation puissent reposer sur les émotions. En effet, (Zawieska, 2015) affirme qu'un robot augmente sa crédibilité lorsque celui-ci imite les traits des humains comme les émotions. Une stratégie de manipulation d'un robot envers un humain peut donc reposer sur l'imitation des émotions afin d'influencer le comportement d'un humain dans le sens du robot manipulateur. Ce type de stratégie de manipulation peut aussi s'appliquer aux interactions entre agents artificiels. En effet, si un agent artificiel est capable de reconnaître les émotions d'un autre agent et réponde à cette émotion d'une façon particulière, un agent manipulateur peut simuler des émotions afin que cet agent artificiel se comporte de la façon désirée par le manipulateur.

Conclusion

Il existe donc de nombreuses bases de stratégies de manipulation comme le résume la figure 1.3. Toutes ces stratégies de manipulation ont trois caractéristiques fondamentales d'une manipulation :

1. l'agent manipulateur délibère sur la stratégie qu'il est en train de mettre en place dans le système pour manipuler les agents ou le système ;
2. il veille à amener les autres agents, ou le système, à une certaine configuration ;
3. cet agent veille toujours à dissimuler sa stratégie.

Remarquons que chaque stratégie est par nature dissimulée à l'autre agent. Sans cela, elles ne pourraient être considérées comme des stratégies de manipulation. Toutes ces stratégies de manipulation ont toujours pour finalité d'amener un agent manipulé à faire ou ne pas faire quelque chose. Mais elles se distinguent par deux choses : la base de la stratégie, et les conséquences impliquées par la stratégie de manipulation. Les différentes bases de stratégies peuvent soit être fondées sur les comportements, les normes, sur les croyances ou les connaissances, sur la rationalité ou les émotions, ou peuvent être fondées sur la confiance entretenue entre agents. Quant-aux conséquences immédiates, nous distinguons deux types de conséquences : celles qui consistent à amener directement l'agent à l'objet de la manipulation sans qu'il ne s'en rende compte, nous parlons alors de *stratégies directes de manipulation* ; sinon nous parlons de *stratégies indirectes de manipulation* lorsqu'un agent manipulateur utilise l'agent manipulé comme substitut à sa manipulation. L'agent manipulateur va tout d'abord chercher à faire croire à l'agent manipulé quelque chose, soit lui faire désirer quelque chose, ou bien chercher à lui faire faire confiance sur un autre agent afin que l'agent victime ou un tiers-agent soit amené finalement à accomplir les volontés manipulatrices du manipulateur.

1.3 Détecter la manipulation

Si la manipulation est par nature dissimulée aux agents victimes de la manipulation, il est toutefois possible de la détecter soit après qu'elle ait eu lieu, soit pendant ou avant mais dans ce cas, uniquement par un agent extérieur à la manipulation. Ainsi, il existe deux approches permettant de détecter une manipulation : les approches statistiques qui consistent à reconnaître des comportements manipulateurs, et les méthodes de preuves automatiques qui consistent à prouver l'existence d'un comportement manipulateur. Cette section présente ces deux approches.

1.3.1 Approches statistiques

Parmi les approches statistiques pour détecter la tromperie, il existe principalement deux façons d'opérer (Santos and Johnson, 2004) : les approches qui s'intéressent à détecter le processus lié à la tromperie, c'est-à-dire reconnaître les motivations d'un agent à tromper, ou reconnaître les attributs de l'environnement perpétrés par une manipulation ; des approches qui visent plutôt la détection d'informations trompeuses. Pour détecter des informations trompeuses, (Santos and Johnson, 2004) identifient quatre types de stratégies de détection :

1. Une méthode fondée sur la reconnaissance passée ou communes d'informations trompeuses ;
2. Une méthode fondée sur des inférences sur les données collectées jusqu'à présent dans un contexte d'informations incomplètes, et ce, jusqu'à ce que l'analyse de l'ensemble des données soit complet et consistant ;
3. Une méthode fondée sur la présomption et consiste à n'accepter de nouvelles données qu'après vérification et validation de celles-ci ;
4. Une méthode fondée sur les intentions et consiste à détecter les croyances, les connaissances et les buts des agents permettant de mieux prédire des informations trompeuses.

(Santos and Johnson, 2004) proposent un modèle pour détecter la tromperie. Ils distinguent deux formes de transmissions de fausses informations : la désinformation réalisée de façon involontaire[14], et la désinformation réalisée de façon volontaire[15]. Santos et Johnson s'intéressent alors à la détection de la désinformation volontaire et présentent un modèle, initialement proposé dans un contexte de détections de fraudes (Johnson et al., 1993). Leur modèle est constitué de quatre étapes :

1. L'étape de l'**activation** consiste à reconnaître certains attributs de l'environnement qui n'étaient pas ceux attendus. Il est possible d'être la cible d'une tentative de manipulation ;
2. La cible de la manipulation émet alors certaines hypothèses sur l'objet de la manipulation : c'est l'étape de **détection** ;
3. La cible de la manipulation **édite** sa représentation du monde en rapport avec l'hypothèse émise à l'étape précédente et les informations de confiance, ainsi qu'en effaçant de sa représentation les informations douteuses ;

14. Nommée en anglais comme la *misinformation*.
15. Nommée en anglais *disinformation*.

4. La dernière étape est la **réévaluation** de la situation, la cible de la manipulation réévalue la situation en révisant sa base de connaissance sur l'environnement.

Pour chacune de ces quatre étapes, les auteurs proposent différentes techniques. Par exemple pour l'étape d'activation, ils proposent d'effectuer une comparaison entre les différentes opinions des agents, de vérifier une opinion attendue pour un certain agent, ou encore de reconnaître les intentions manipulatrices. Ainsi, le lecteur intéressé peut se référer à (Santos and Johnson, 2004) pour davantage de détails sur ces différentes techniques.

Les méthodes statistiques pour la détection des manipulations sont aussi utilisées dans des domaines comme les systèmes de réputation et de recommandation (Vallée and Bonnet, 2015). Ils utilisent la mesure de Kullback-Leibler pour mesurer la dissimilarité entre deux mesures de distributions de probabilités (Kullback and Leibler, 1951), l'une représentant la distribution sur les témoignages honnêtes des agents, l'autre représentant le témoignage d'un agent quelconque. Ainsi, plus la mesure de Kullback-Leibler est importante, plus il est possible qu'un agent soit en train de fournir un faux témoignage.

Nous retrouvons des approches très similaires dans les systèmes de recommandation pour la détection de techniques de manipulation (Hurley et al., 2009; Zhang and Zhou, 2012; Cao et al., 2013) [16]. Par exemple, (Hurley et al., 2009) utilise une analyse en composante principale afin de mettre en évidence au sein de groupes (clusters) les utilisateurs qui trompent le système.

1.3.2 Détecter la tromperie par le raisonnement

Si les approches statistiques sont très efficaces pour détecter la tromperie, elles ne fournissent aucune description de ce qui caractérise la nature de la tromperie. Une dernière méthode consiste à prouver par déductions logiques qu'un utilisateur est en train de tromper le système. Ce type d'approche consiste à définir des mécanismes de raisonnements logiques, ainsi que des prédicats pour caractériser formellement une tromperie dans un système spécifique (Muller and Vercouter, 2004; Burrows et al., 1989; Van Ditmarsch et al., 2012; Sakama et al., 2015).

Par exemple, (Muller and Vercouter, 2004) considèrent la détection d'agents menteurs dans un système où les agents s'échangent des informations. Pour ce faire, ils contraignent les agents du système à respecter une norme de bonne conduite consistant à ne jamais accepter deux témoignages qui se contredisent et provenant d'un même agent. En revanche, si l'autorité centrale détecte que cette norme de consistance est violée pour un agent alors elle déduit qu'il existe un agent menteur dans le système. Une méthode consiste à appliquer la détection dirigée par l'observation. Cette méthode consiste à observer les propositions émises par un agent et de vérifier si l'ensemble de ces propositions émises est cohérent.

D'autres auteurs comme (Van Ditmarsch et al., 2012; Sakama et al., 2015) définissent des logiques modales pour exprimer, avec des prédicats, le mensonge, la tromperie, ou encore, le baratinage. Le mensonge est par exemple définit comme le fait qu'un agent communique une

16. Les systèmes de recommandation sont des systèmes relativement proches des systèmes de recommandation. A la différence, les systèmes de recommandation ont pour finalité de renvoyer un ensemble d'objets pour des utilisateurs. Une technique très utilisée est le filtrage collaboratif qui vise à regrouper les objets en fonctions de groupes d'utilisateurs similaires.

information qu'il ne croit pas lui même et avec l'intention que l'interlocuteur croit cette information. Dans le chapitre 2, nous nous intéressons à ces méthodes de preuves formelles et nous présentons de façon plus approfondie ces méthodes permettant de prouver l'existence de telles stratégies de manipulation dans un système.

1.4 Combattre la manipulation

Les stratégies de manipulation sont par nature dissimulées et sont donc souvent difficiles à détecter. Il existe des méthodes pour la combattre directement dans les systèmes et permettant d'assurer une certaine robustesse du système face à certaines stratégies de manipulation. Si une première méthode consiste à intégrer un ensemble de normes contraignant les agents à bien se comporter ou une notion de confiance entre agents, ces deux systèmes amènent de nouvelles stratégies de manipulation comme par exemple exploiter la norme réciprocité ou encore d'abuser de la confiance. Ces méthodes ne suffisent donc pas pour combattre véritablement la manipulation bien qu'elles puissent inciter certains agents à bien se comporter. Il existe trois types de méthodes permettant de construire des stratégies de défense face aux manipulations (Vallée, 2015). Ces méthodes passent par :

1. Axiomatiser le système de sorte qu'il soit robuste à certaines manipulations ;
2. Jouer sur la complexité à la mise en place de manipulations ;
3. Intégrer des mécanismes d'incitation à bien se comporter.

1.4.1 Axiomatiser un système robuste aux manipulations

Pour empêcher l'existence de certaines stratégies de manipulation dans un système, une première méthode consiste à penser ce système de telle sorte qu'il soit prouvé robuste à la manipulation si certaines propriétés nommées *axiomes* sont vérifiées pour le système.

Dans les systèmes de vote, la robustesse aux *choix stratégiques*[17] se passe par la construction d'une certaine fonction de choix social[18] robuste[19] (Gibbard, 1973; Satterthwaite, 1975; Barbera, 2001; Nehring and Puppe, 2007; Guo and Conitzer, 2010). De manière intéressante, (Gibbard, 1973; Satterthwaite, 1975) ont démontré qu'il était impossible de construire une règle de vote qui puisse satisfaire la propriété de *non-dictature*[20] et la propriété de *robustesse aux choix stratégiques*. Ce résultat rejoint et complète le théorème d'impossibilité de Arrow (Arrow, 2012), c'est-à-dire qu'il n'existe aucune fonction d'agrégation définie sur l'ensemble des profils de

17. Nous parlons de *choix stratégiques* pour désigner la stratgie de manipulation visant à fournir un faux profil de préférence afin de s'assurer d'un résultat préférable à celui obtenu en donnant son véritable profil de préférence (Gibbard, 1973).

18. Nous appelons *fonction de choix social*, une fonction qui pour tout ensemble de relations d'ordre total associe une relation d'ordre total. Chaque relation d'ordre représente un profil de préférences pour un agent sur un ensemble d'issues possibles dans un vote. La valeur de retour représente la décision finale du vote.

19. Une fonction de choix social est dite robuste aux choix stratégiques si, et seulement si, il n'existe pas d'agent dans le système tel que si l'agent fournit un profil de préférence différent de celui qu'il aurait proposé initialement, celui-ci obtient une issue qui lui soit préférable par cette fonction de choix social.

20. La propriété de non-dictature signifie que le résultat d'un vote ne dépend pas du profil de préférences d'un agent en particulier.

préférences du système et satisfaisant la propriété d'unanimité[21], d'indifférence aux options non pertinentes[22] et de non-dictature.

Concernant les systèmes de réputations, (Tennenholtz, 2004; Altman and Tennenholtz, 2005; Altman and Tennenholtz, 2010) ont étudié la manière dont un système de réputation ou de recommandation pouvait être axiomatisé pour rendre impossible l'existence de certaines stratégies de manipulation permettant à un agent d'obtenir à résultat qui lui soit favorable en manipulant la fonction de recommandation ou la fonction d'agrégation dans le cas des systèmes de réputation. (Altman and Tennenholtz, 2007) ont étudié différentes combinaisons d'axiomes comme *la généralité*[23], *la transitivité*[24], *la monotonie*[25], *l'indépendance* et *l'incitation à la vérité*[26] et *l'indépendance des alternatives non pertinentes*. En particulier, ils ont montré qu'il était impossible de construire un système de réputation qui respecte à la fois la généralité, la transitivité, la monotonie et l'incitation à la vérité.

D'autres chercheurs comme (Cheng and Friedman, 2005) ont étudié la robustesse de systèmes de réputation face à certaines stratégies de manipulation comme l'attaque Sybil. L'attaque Sybil (Douceur, 2002) consiste pour un agent manipulateur à introduire de faux agents pour obtenir une issue finale plus favorable. Si la fonction de réputation possède certaines propriétés comme la symétrie[27], celle-ci ne peut être robuste à la stratégie d'attaque Sybil.

1.4.2 Empêcher certaines stratégies par la complexité

Bien que des auteurs comme (Altman and Tennenholtz, 2007) ont montré que dans certains systèmes, il était impossible de les construire robustes à certaines stratégies de manipulation, une autre manière pour les combattre consiste à construire des systèmes complexes à manipuler. C'est pourquoi certains domaines comme la théorie du choix social computationnel s'intéressent à la complexité des stratégies de manipulation (Conitzer and Sandholm, 2004; Bartholdi et al., 1989)

Par exemple, dans les systèmes de vote, (Bartholdi et al., 1989; Bartholdi and Orlin, 1991) montrent que construire une stratégie de manipulation pour le vote majoritaire est un problème

21. L'unanimité signifie que si tous les agents préfèrent une issue plutôt qu'une autre alors l'issue finale est bien celle que tous les agents préfèrent.

22. Une fonction de choix social est dite indifférente aux options non pertinentes si pour tout sous-ensemble d'issues possibles, l'ordre des préférences donné sur ce sous-ensemble par la fonction de choix social reste inchangé par rapport à celui donné pour toutes les issues possibles.

23. La généralité signifie que la fonction qui évalue la confiance doit être définie pour tout graphe de confiance. Un graphe de confiance est un graphe où chaque nœud représente un agent et les arêtes la confiance.

24. La transitivité dans ce contexte signifie que si l'ensemble des agents G_i qui ont confiance en un agent i et G_j l'ensemble des agents qui ont confiance en j et tel que G_i a un meilleur rang (réputation) que les agents de G_j alors l'agent i a un meilleur rang (réputation) que l'agent j.

25. La monotonie traduit que si un agent i a un meilleur rang qu'un agent j alors il existe un agent k qui a confiance en i et tel que k a un meilleur rang que tout agent de G_j où G_j est le groupe d'agents qui fait confiance à j.

26. L'incitation à la vérité traduit la propriété qu'il est impossible pour un agent d'augmenter son rang en fournissant un faux profil.

27. Une fonction de réputation f est symétrique si et seulement si pour tout graphe de confiance $G_1 = (V_1, E_1), G_2 = (V_2, E_2)$ isomorphiques par σ l'ordre donné par la fonction de réputation est le même i.e $f_i(G_1) = f_{\sigma(i)}(G_2)$ pour tout agent $i \in V_1$

polynomial, donc facile à résoudre. (Conitzer and Sandholm, 2004) montrent qu'il est NP-difficile de construire une stratégie de manipulation, dans un problème de formation de coalitions, lorsque la répartition des gains se fait en fonction de la valeur de Shapley. La complexité des manipulations diffère donc selon le type de manipulation et le système considéré comme le montrent (Conitzer and Sandholm, 2002; Conitzer and Sandholm, 2006).

Cependant, la plupart des approches étudient la complexité dans le pire cas. Or, (Walsh, 2009) montre qu'une règle de vote qui soit difficile à manipuler dans le pire cas peut être réalisée en temps polynomial dans de nombreux cas pratiques. Par conséquent, si la complexité algorithmique peut être considérée comme un frein à la construction de stratégies de manipulation, cela ne garantit en rien la robustesse d'un système face à la construction de telles stratégies dans le cas général.

1.4.3 Intégrer des mécanismes d'incitation à bien se comporter

Après avoir constaté que l'axiomatisation du système ou que la complexité ne permettaient pas d'empêcher complètement les agents de construire des stratégies de manipulation, une dernière méthode consiste à inciter les agents à bien se comporter. Ces approches basées sur la théories des jeux étudient des mécanismes, appelés *mécanismes d'incitation*, et cherchent à montrer que dans certaines configurations, un agent rationnel a davantage intérêt à bien se comporter plutôt que le contraire.

Par exemple, les systèmes de réputation et les systèmes normatifs ont cela pour objectif. Ainsi un système normatif va avoir pour rôle de punir des agents malhonnêtes, en leur affectant des pénalités (Elster, 1989; Boella et al., 2009), et les agents honnêtes pourront obtenir certaines récompenses. Du point de vue de la théorie des jeux, ces pénalités représentent un coût pour l'agent. Ainsi, un agent rationnel peut déduire que dans un système normatif, celui-ci n'a aucun intérêt à se comporter de façon malhonnête.

De la même manière, un système de réputation peut permettre d'inciter des agents à bien se comporter en intégrant certains protocoles au système. Par exemple, (Bonnet, 2012) propose un protocole de transmissions de témoignages incitatif, si le coût de création d'un agent Sybil est non-nul.

1.5 Problématique générale

Dans ce chapitre, nous avons donc tout d'abord présenté en section 1.1 les *systèmes multi-agents* et les *agents cognitifs* qui sont des agents capables de raisonner sur leurs états mentaux ainsi que les états mentaux des autres agents comme, par exemple, leurs connaissances, leurs croyances, leurs intentions, leurs désirs, ou encore leurs buts. Puisque certains agents du système peuvent être malhonnêtes, non fiables, ou manipulateurs, nous avons ensuite présenté des systèmes capables de réguler les interactions entre agents comme *les systèmes normatifs* et *les systèmes de confiance*.

En section 1.2, après avoir présenté un état de l'art sur la notion de manipulation, nous sommes parvenus à la conclusion que la manipulation pouvait se définir comme « l'intention

délibérée d'instrumentaliser un agent, tout en veillant à lui dissimuler cette intention de les instrumentaliser », puis nous avons présenté dans les systèmes multi-agents des travaux en lien avec cette notion de manipulation, ainsi qu'un ensemble de bases de stratégies de manipulation comme l'exploitation d'une faille, le détournement d'une norme, l'abus de la confiance, ou encore profiter de la rationalité des agents.

Cependant, même si la manipulation est par définition toujours dissimulée à la victime, nous avons présenté en section 1.3 des méthodes permettant de la détecter, puis à la section 1.4, nous avons présenté différentes approches pour se défendre face à certaines stratégies de manipulation. Parmi ces approches, nous retrouvons l'axiomatisation d'un système robuste aux choix stratégiques, l'étude de la complexité des manipulations, ou encore l'étude de mécanismes d'incitation. En revanche, aucun travaux n'a été porté sur la représentation formelle de la manipulation telle que définie dans ce manuscrit.

Or, nous pensons que mieux comprendre et définir formellement la manipulation permettrait de **mieux concevoir des systèmes informatiques robustes à la manipulation** ou, dans le cas où il est impossible de construire de tels systèmes, cela aiderait lors de la **détection des configurations du système dans lesquelles il est possible qu'un agent soit en train de manipuler le système**.

Si certains travaux se sont penchés sur la détection automatique de la manipulation avec des approches statistiques, ces travaux ne fournissent aucune description permettant de mieux comprendre la manipulation. Ainsi, nous nous intéressons dans le chapitre 2 aux approches logiques en lien avec des notions permettant de caractériser la manipulation comme l'influence, la tromperie, la malhonnêteté ou encore la prise de conscience. De plus, puisque certaines stratégies de manipulation peuvent être fondées sur la confiance, les normes, les émotions, ou la rationalité, nous présentons des approches logiques permettant de les caractériser. Dans la suite de cette thèse, nous proposons en chapitre 3, un système logique pour raisonner sur la manipulation, puis, en chapitre 4, puisque la confiance naît du constat que certains agents peuvent être non fiables, malhonnêtes, voire manipulateurs, nous proposons un système logique pour raisonner sur la notion de confiance en la sincérité d'un agent sur une certaine proposition. Le chapitre 5 présente une nouvelle méthode des tableaux pour raisonner dans le système logique de la confiance en la sincérité. Enfin, le chapitre 6 conclut cette thèse en présentant une synthèse des contributions ainsi que les perspectives de recherches futures.

Chapitre 2

Manipulation et confiance vues par les logiques formelles

Dans le chapitre 1, nous avons défini la manipulation comme « l'intention délibérée d'un agent d'instrumentaliser un autre agent, tout en veillant à lui dissimuler cette intention de l'instrumentaliser ». Nous avons alors présenté des travaux dans les systèmes multi-agents qui proposaient une taxonomie associée à la tromperie, puis d'autres travaux sur la construction de stratégies de manipulation, ainsi que des méthodes permettant de détecter des stratégies de manipulations, enfin, nous avons présenté des méthodes pour s'en protéger. Cependant, aucun travaux présenté jusqu'alors n'a décrit formellement la manipulation. Notre définition met en lumière le fait que différents états mentaux sont nécessaires pour établir une manipulation :

1. la manipulation est une *intention délibérée* d'un agent manipulateur, c'est-à-dire, elle est préméditée par cet agent comme une stratégie préalablement calculée ;
2. la manipulation est une *instrumentalisation* d'un agent victime, c'est-à-dire, une influence exercée par le manipulateur sur les intentions d'une victime pour l'amener à réaliser, à son insu ou de façon délibérée, les intentions du manipulateur ;
3. la manipulation est *toujours dissimulée* à la victime, c'est-à-dire, elle peut être soit un manque de connaissance, soit un manque de conscience, soit une non croyance de la victime sur les intentions du manipulateur.

Or, les approches logiques permettent de donner une description formelle à des objets abstraits comme les connaissances, les croyances ou encore les intentions. En particulier, les logiques modales permettent d'exprimer explicitement les états mentaux des agents. Ce chapitre présente des logiques modales permettant d'exprimer des états mentaux pouvant être impliqués dans la manipulation, ou encore dans des stratégies de manipulation. Ces états mentaux sont : les intentions ; les connaissances et les croyances ; la prise de conscience ; la confiance ; les émotions ; les normes.

Dans un premier temps, la section 2.1 fait des rappels sur les logiques et présente les logiques modales. Cette section termine en présentant un ensemble d'exemples de logiques modales qui modélisent les croyances, les connaissances, les normes, les intentions et les actions, mais aussi

les désirs et les émotions des agents. Dans un second temps, si la notion de confiance naît du constat que certains agents peuvent être malintentionnés, la section 2.2 présente des logiques qui modélisent la confiance. Enfin, la section 2.3 présente les logiques en lien avec la manipulation et qui modélisent l'influence, la malhonnêteté, le mensonge ou encore, la prise de conscience.

2.1 Des logiques pour agents cognitifs

Il existe de nombreuses logiques dans la littérature : les logiques propositionnelles, les logiques de prédicats, les logiques floues, ou encore, les logiques non monotones. De façon générale, les logiques peuvent être classées en quatre catégories distinctes nommées : *les logiques d'ordre zéro* qui n'emploient aucun quantificateurs universels ($\forall$), ni existentiels ($\exists$) ; *les logiques du premier ordre* qui peuvent utiliser les quantificateurs universels et existentiels ; *les logiques d'ordre supérieur* qui autorisent l'utilisation des quantificateurs sur les prédicats ; et enfin, *les logiques non classiques* qui peuvent, par exemple, s'affranchir de certaines règles de la logique classique comme la règle de la monotonie [1]. Dans le cadre de ce travail, nous faisons tout d'abord des rappels sur les logiques propositionnelles en section 2.1.1, puis nous présentons en section 2.1.2 les logiques modales. Le lecteur familier peut se référer directement à la section 2.1.3 pour une présentation des logiques modales classiques en lien avec la problématique de cette thèse sur la modélisation de la manipulation.

2.1.1 Rappels sur les logiques

Logiques propositionnelles

La logique propositionnelle est une logique d'ordre zéro. Le lecteur intéressé peut se référer à (Givant and Halmos, 2008). Cette logique est décrite par un langage $\mathcal{L}_p$ sur un ensemble de lettres propositionnelles, noté $\mathcal{A}(\mathcal{L}_p) = \{p, q, r, s, \ldots\}$ et généré par une grammaire sous sa forme de Backus-Naur, pour tout atome $p \in \mathcal{A}(\mathcal{L}_p)$:

$$\phi ::= p|\phi_1 \wedge \phi_2|\phi_1 \vee \phi_2|\phi_1 \Rightarrow \phi_2|\phi_1 \Leftrightarrow \phi_2|\neg\phi|\top|\bot$$

Par exemple, la formule $p \wedge (q \Rightarrow r)$ est une formule du langage propositionnel. Soit $\nu : \mathcal{A}(\mathcal{L}_p) \rightarrow \{\top, \bot\}$ une fonction qui pour chaque atome du langage associe sa valeur de vérité où $\top$ représente le vrai, et $\bot$ représente ce qui est faux. Nous appelons ν une *fonction d'assignation.*

Définition 2.1 - *Fonction d'interprétation* : *Soit* $\nu : \mathcal{A}(\mathcal{L}_p) \rightarrow \{\top, \bot\}$ *une fonction d'assignation. Nous appelons* fonction d'interprétation *associée à l'assignation* ν*, une fonction* $I_\nu : \mathcal{L}_p \rightarrow \{\top, \bot\}$ *telle que :*

1. $I_\nu(\top) = \top$ *et* $I_\nu(\bot) = \bot$
2. $\forall p \in \mathcal{A}(\mathcal{L}_p) : I_\nu(p) = \nu(p)$

1. La monotonie consiste à dire que si une propriété Q est une conséquence logique de P alors ajouter une nouvelle hypothèse R n'empêche pas la déduction logique de Q.

3. $\forall \phi \in \mathcal{L}_p : I_\nu(\neg\phi) = \neg I_\nu(\phi)$
4. $\forall \phi, \psi \in \mathcal{L}_p : I_\nu(\phi \wedge \psi) = \top$ *si, et seulement si,* $I_\nu(\phi) = \top$ *et* $I_\nu(\psi) = \top$
5. $\forall \phi, \psi \in \mathcal{L}_p : I_\nu(\phi \vee \psi) = \neg I_\nu(\neg\phi \wedge \neg\psi)$
6. $\forall \phi, \psi \in \mathcal{L}_p : I_\nu(\phi \Rightarrow \psi) = I_\nu(\neg\phi \vee \psi)$
7. $\forall \phi, \psi \in \mathcal{L}_p : I_\nu(\phi \Leftrightarrow \psi) = I_\nu((\phi \Rightarrow \psi) \wedge (\psi \Rightarrow \phi))$

La fonction d'interprétation permet alors d'associer une sémantique au langage $\mathcal{L}_p$.

Définition 2.2 - ***Vocabulaire*** **:**

1. *Toute formule de* $\mathcal{L}_p$ *est qualifiée de* formule bien formée *;*
2. *Une formule* ϕ *est* valide *ou une* tautologie *si, et seulement si, quelque soit les assignations* ν, $I_\nu(\phi) = \top$ *sinon elle est dite* invalide *ou* falsifiable. *Une formule valide est notée* $\models \phi$ *;*
3. *Une formule* ϕ *est* satisfiable *si, et seulement si, il existe une assignation* ν *telle que* $I_\nu(\phi) = \top$ *sinon elle est dite* contradictoire *ou* insatisfiable. *S'il existe* ν *tel que* $I_\nu = \top$ *alors nous appelons la fonction d'interprétation* I_ν *un* modèle de la formule ϕ *et nous notons* $I_\nu \models \phi$, *ou encore* $\models_{I_\nu} \phi$ *;*
4. *Si* Γ *est un ensemble de formules de* $\mathcal{L}_p$, I_ν *est un* modèle *pour* Γ *si, et seulement si,* $\forall \phi \in \Gamma$, I_ν *est un modèle pour* ϕ *;*
5. *Si* $\Gamma = \{\phi_0, \phi_1, \ldots\}$ *est un ensemble de formules de* $\mathcal{L}_p$, ϕ *est une* conséquence sémantique *de* Γ, *et noté* $\Gamma \models \phi$ *si, et seulement si, tout modèle de* Γ *est un modèle de* ϕ.

Systèmes de preuves syntaxiques

De nombreux systèmes syntaxiques de preuves existent comme le calcul des séquents, les systèmes de déduction naturels, les méthodes des tableaux ou encore les systèmes à la Hilbert.

Définition 2.3 - ***Preuve syntaxique*** **:** *Soit le triplet* $\mathcal{S} = (\mathcal{L}_p, \mathcal{X}, \mathcal{D})$ *sur* $\mathcal{L}_p$ *avec* $\mathcal{X}$ *un ensemble de formules de* $\mathcal{L}_p$ *nommées* axiomes *et* $\mathcal{D}$ *un ensemble de règles de déductions, i.e. un ensemble de fonctions* $f : 2^{\mathcal{L}_p} \to \mathcal{L}_p$ *qui à un ensemble de formules appelées* prémisses *associe une formule, appelée* conclusion. *Nous appelons* $\mathcal{S}$, un système de preuve. *Une preuve d'une formule* ϕ *est une liste finie de formules* $(\phi_i)_{i \in [|0,n|]}$ *avec* $n \in \mathbb{N}$ *et telles que :*

1. $\phi_n = \phi$ *;*
2. ϕ_0 *est un axiome de* $\mathcal{X}$ *;*
3. $\forall i \in [|0, n|]$, ϕ_i *est soit un axiome du système de preuves* $\mathcal{S}$, *ou si* $i > 0$, *il existe* $f \in \mathcal{D}$ *et* $P \subseteq \{\phi_0, \phi_1, \ldots, \phi_{i-1}\}$ *tel que* $f(P) = \phi_i$.

Si pour une formule ϕ, *il existe une preuve de* ϕ *dans le système* $\mathcal{S}$, *alors* ϕ *est une* formule prouvée *du système* $\mathcal{S}$ *ou encore* ϕ *est un* théorème *de* $\mathcal{S}$, *noté* $\vdash_{\mathcal{S}} \phi$.

Par exemple, un système à la Hilbert $\mathcal{S}$ considère une liste d'axiomes comme ceux donnés dans la figure 2.1 et considère deux règles de déduction : le *modus ponens* et la *substitution*

$$\vdash \phi \Rightarrow (\psi \Rightarrow \phi) \quad (R_1)$$
$$\vdash (\phi \Rightarrow (\psi \Rightarrow \theta)) \Rightarrow ((\phi \Rightarrow \psi) \Rightarrow (\phi \Rightarrow \theta)) \quad (R_2)$$
$$\vdash (\neg\psi \Rightarrow \neg\phi) \Rightarrow (\phi \Rightarrow \psi) \quad (R_3)$$
$$\vdash (\neg\phi \Rightarrow \neg\psi) \Rightarrow ((\neg\phi \Rightarrow \psi) \Rightarrow \phi) \quad (R_4)$$
$$\vdash \phi \Rightarrow (\phi \vee \psi), \psi \Rightarrow (\phi \vee \psi) \quad (R_5)$$
$$\vdash (\phi \vee \psi) \Rightarrow ((\phi \Rightarrow \theta) \Rightarrow ((\psi \Rightarrow \theta) \Rightarrow \theta)) \quad (R_6)$$
$$\vdash \phi \Rightarrow (\psi \Rightarrow \phi \wedge \psi) \quad (R_7)$$
$$\vdash \phi \wedge \psi \Rightarrow \phi, \phi \wedge \psi \Rightarrow \psi \quad (R_8)$$

FIGURE 2.1 – Liste des axiomes du Calcul Propositionnel (CP).

uniforme. Le *modus ponens* est une règle de déduction naturelle qui consiste à énoncer que si ϕ est prouvée dans le système $\mathcal{S}$ et que $\phi \Rightarrow \psi$ est aussi prouvée dans $\mathcal{S}$, alors la conclusion ψ est prouvée dans $\mathcal{S}$. Cette règle est alors notée, formellement, de la façon suivante :

$$\begin{array}{ll} (P1) & \vdash \quad \phi \\ (P2) & \vdash \quad \phi \Rightarrow \psi \\ \hline (C1) & \vdash \quad \psi \end{array}$$

Un système de preuves à la Hilbert considère une autre règle qui est la *substitution uniforme*.

Définition 2.4 - *Substitution* : *Nous appelons* substitution, *une fonction* $\sigma : \mathcal{A}(\mathcal{L}_p) \to \mathcal{L}_p$ *qui associe à chaque variable propositionnelle une formule du langage* $\mathcal{L}_p$. *La fonction* $(.)^\sigma : \mathcal{L}_p \to \mathcal{L}_p$ *est la* substitution uniforme induite *de* σ *si, et seulement si,* σ *est une substitution et :*

1. $(\top)^\sigma = \top$, $(\bot)^\sigma = \bot$
2. $\forall p \in \mathcal{A}(\mathcal{L}_p) : (p)^\sigma = \sigma(p)$
3. $\forall \phi \in \mathcal{L}_p : (\neg\phi)^\sigma = \neg\phi^\sigma$
4. $\forall \phi, \psi \in \mathcal{L}_p : (\phi \wedge \psi)^\sigma = \phi^\sigma \wedge \psi^\sigma$
5. $\forall \phi, \psi \in \mathcal{L}_p : (\phi \vee \psi)^\sigma = \phi^\sigma \vee \psi^\sigma$
6. $\forall \phi, \psi \in \mathcal{L}_p : (\phi \Rightarrow \psi)^\sigma = \phi^\sigma \Rightarrow \psi^\sigma$
7. $\forall \phi, \psi \in \mathcal{L}_p : (\phi \Leftrightarrow \psi)^\sigma = \phi^\sigma \Leftrightarrow \psi^\sigma$

Nous notons $Sub_{\mathcal{L}_p}$ *l'ensemble des substitutions définies sur* $\mathcal{L}_p$.

Nous appelons *substitution uniforme*, la règle définie comme si $\vdash \phi$ est prouvée dans un système de preuves $\mathcal{S}$, alors pour toute substitution uniforme d'une lettre propositionnelle p par une formule ψ, le résultat de cette substitution, notée $\vdash (\phi)[p\backslash\psi]$ est aussi prouvé dans $\mathcal{S}$. La règle de la substitution uniforme est alors donnée par :

$$\frac{(P1) \quad \vdash \quad \phi}{(C1) \quad \vdash \quad (\phi)[p\backslash\psi]}$$

Dans la logique propositionnelle, la substitution uniforme, ainsi que le modus ponens préservent la validité. Cela signifie que si une formule ϕ est une tautologie du calcul propositionnel et, si après l'application d'une de ces deux règles nous déduisons une formule ψ, alors la formule ψ est aussi une tautologie du calcul propositionnel.

Lemme 2.1 : *Les propriétés de préservation :*

1. *Le modus ponens préserve la validité i.e. :*
   si $\models \phi$ *et* $\models \phi \Rightarrow \psi$ *sont valides, alors* $\models \psi$ *est valide ;*
2. *La substitution uniforme préserve la validité i.e. :*
   si $\models \phi$ *est valide, alors pour toute substitution* σ, $\models (\phi)^{\sigma}$ *est aussi valide.*

Exemple 2.1 : *Voici un exemple d'application des systèmes de preuves à la Hilbert.*

1. $\vdash (\phi \Rightarrow ((\psi \Rightarrow \phi) \rightarrow \phi)$ *par (Sub,* $\sigma(\psi) = \psi \Rightarrow \phi$*) dans* (R_1)
2. $\vdash (\phi \Rightarrow ((\psi \Rightarrow \phi) \rightarrow \phi) \Rightarrow ((\phi \Rightarrow (\psi \Rightarrow \phi)) \rightarrow (\phi \Rightarrow \phi)$ *(Sub,* $\sigma(\psi) = \psi \Rightarrow \phi, \sigma(\theta) = \phi$*) dans* (R_2)
3. $\vdash ((\phi \Rightarrow (\psi \Rightarrow \phi)) \Rightarrow (\phi \Rightarrow \phi)$ *(MP de 1 et 2)*
4. $\vdash \phi \Rightarrow (\psi \Rightarrow \phi)$ (R_1)
5. $\vdash \phi \Rightarrow \phi$ *(MP de 3 et 4)*

Ces lemmes permettent de prouver que des systèmes de preuves qui utilisent ces règles sont corrects, c'est-à-dire que toutes les formules qui sont prouvées par un tel système sont bien valides du point de vue de la sémantique. Cette propriété est nommée *propriété de correction* ou encore *propriété d'adéquation*. La réciproque est nommée *propriété de complétude*. Elle énonce que toute formule $\models \phi$ valide peut alors être prouvée $\vdash \phi$ dans un système de preuves $\mathcal{S}$ complet.

Théorème 2.1 - *Correction et complétude* : *Soient* $\mathcal{S}$ *un système de preuves à la Hilbert muni du modus ponens, de la substitution uniforme et des axiomes donnés par la figure 2.1 et* ϕ *une formule.*

Ce système $\mathcal{S}$ *est* correct *et* complet*, c'est-à-dire vérifie les théorèmes suivants :*

1. *Théorème de correction : si* $\vdash_{\mathcal{S}} \phi$ *alors* $\models \phi$
2. *Théorème de complétude : si* $\models \phi$ *alors* $\vdash_{\mathcal{S}} \phi$

La preuve du théorème de complétude repose sur la notion d'ensembles maximums $\mathcal{S}$-consistants. Nous rappelons ci-dessous des résultats bien connus sur les ensembles maximaux $\mathcal{S}$-consistants. Ces résultats sont étendus aux logiques modales et nous serviront pour prouver la complétude des systèmes logiques que nous proposons aux chapitres 3 et 4.

Ensembles de formules $\mathcal{S}$-consistants

Considérons un système de preuves $\mathcal{S}$.

Définition 2.5 - *Ensembles $\mathcal{S}$-consistants* : *Soit un ensemble Σ de formules,*

— Σ *est* $\mathcal{S}$-inconsistant *si, et seulement si,* $\exists \psi_1, ..., \psi_n \in \Sigma : \vdash \neg \bigwedge_{i=1}^{n} \psi_i$;

— Σ *est* $\mathcal{S}$-consistant *si, et seulement si,* Σ *n'est pas* $\mathcal{S}$*-inconsistant.*

Tout ensemble de formules $\mathcal{S}$-consistant peut être étendu à un ensemble maximal $\mathcal{S}$-consistant. Un ensemble maximal $\mathcal{S}$-consistant est donc un ensemble de formules $\mathcal{S}$-consistant tel qu'il n'existe aucun sur-ensemble strict de formules qui soit aussi $\mathcal{S}$-consistant.

Définition 2.6 - *Ensembles maximaux $\mathcal{S}$-consistants* : *Un ensemble Γ de formules est* maximal $\mathcal{S}$-consistant *si, et seulement si,* $\nexists \Gamma' : \Gamma \subsetneq \Gamma'$ *tel que* Γ' *est* $\mathcal{S}$*-consistant.*

Les ensembles $\mathcal{S}$-consistants ont pour propriété fondamentale que pour tout ensemble Γ de formules $\mathcal{S}$-consistant, il existe un sur-ensemble de formules maximal $\mathcal{S}$-consistant contenant Γ. Cette propriété porte le nom de *lemme de Lindenbaum* et servira aux chapitres 3 et 4 afin de montrer la complétudes des systèmes logiques proposés dans ces chapitres.

Théorème 2.2 - *Lemme de Lindenbaum* :
Pour tout ensemble Γ de formules $\mathcal{S}$-consistant, il existe un ensemble de formules Γ' tel que $\Gamma \subseteq \Gamma'$ et Γ' est maximal $\mathcal{S}$-consistant.

Les ensembles maximaux $\mathcal{S}$-consistants possèdent plusieurs propriétés notées de *MC1* à *MC5*.

Théorème 2.3 - : *Pour tout ensemble Γ maximal $\mathcal{S}$-consistant et ϕ, ψ deux formules.*

1. *MC1 :* $\Gamma \vdash_{\mathcal{S}} \phi \Longrightarrow \phi \in \Gamma$
2. *MC2 :* $(\phi \in \Gamma \vee \neg\phi \in \Gamma) \wedge \neg(\phi \in \Gamma \wedge \neg\phi \in \Gamma)$
3. *MC3 :* $(\phi \vee \psi \in \Gamma) \Longrightarrow \phi \in \Gamma$ *ou* $\psi \in \Gamma$
4. *MC3' :* $(\phi \wedge \psi \in \Gamma) \Longrightarrow \phi \in \Gamma$ *et* $\psi \in \Gamma$
5. *MC4 :* $[(\phi \Rightarrow \psi \in \Gamma) \wedge (\phi \in \Gamma)] \Longrightarrow \psi \in \Gamma$
6. *MC5 :* $\vdash_{\mathcal{S}} \phi$ *ssi* $\forall \Gamma$ *maximal* $\mathcal{S}$*-consistant,* $\phi \in \Gamma$

Définition de la $\mathcal{S}$-déductibilité

Pour tout système de preuves $\mathcal{S}$, nous parlons de $\mathcal{S}$-déductibilité lorsqu'une formule ϕ peut être déduite à partir d'un ensemble de formules Σ dans ce système de preuves.

Définition 2.7 - *$\mathcal{S}$-déductibilité* : *Soient $\mathcal{S}$ un système de preuves, Σ un ensemble de formules prouvées dans $\mathcal{S}$ et ϕ une formule. Nous appelons* $\mathcal{S}$-déduction de ϕ à partir de l'ensemble Σ dans le système $\mathcal{S}$ *et notons $\Sigma \vdash_{\mathcal{S}} \phi$ si, et seulement si :*

1. *Si $\Sigma = \emptyset$, alors $\vdash_{\mathcal{S}} \phi$;*
2. *Sinon $\exists \psi_1, \ldots, \psi_n \in \Sigma$ avec $n \in \mathbb{N}^*$ et $\vdash_{\mathcal{S}} \psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \phi$.*

Si $\Sigma \vdash_{\mathcal{S}} \phi$ est vérifié, alors ϕ est dite $\mathcal{S}$-déductible.

La $\mathcal{S}$-déductibilité possède différentes propriétés fondamentales comme la réflexivité, la transitivité, ou encore l'affaiblissement à gauche.

Proposition 2.1 : *Soient $\mathcal{S}$ un système de preuves, Σ, Γ deux ensembles de formules prouvées dans $\mathcal{S}$, ϕ et ψ deux théorèmes.*

Réflexivité *: si $\phi \in \Sigma$ alors $\Sigma \vdash_{\mathcal{S}} \phi$*

Transitivité *: si $\Sigma \vdash_{\mathcal{S}} \phi$ et $\{\psi\} \vdash_{\mathcal{S}} \phi$ alors $\Sigma \vdash_{\mathcal{S}} \phi$*

Affaiblissement à gauche *: si $\Sigma \vdash_{\mathcal{S}} \phi$ et $\Sigma \subseteq \Gamma$ alors $\Gamma \vdash_{\mathcal{S}} \phi$*

Ces définitions de systèmes de preuves et de $\mathcal{S}$-déductibilité nous donnent les théorèmes suivants.

Théorème 2.4 - *Théorèmes de déduction* : *Soient $\mathcal{S}$ un système de preuves, Σ un ensemble de formules prouvées dans $\mathcal{S}$, ϕ et ψ deux théorèmes.*

$$\Sigma \cup \{\psi\} \vdash_{\mathcal{S}} \phi \text{ si, et seulement si, } \Sigma \vdash_{\mathcal{S}} \psi \Rightarrow \phi$$

De manière équivalente, nous avons le théorème de déduction sous sa forme sémantique :

$$\Sigma \cup \{\psi\} \models \phi \text{ si, et seulement si, } \Sigma \models \psi \Rightarrow \phi$$

Les théorèmes de déduction nous permettent de démontrer les versions fortes des théorèmes de correction et de complétude.

Théorème 2.5 - *Correction et complétude forte* : *Soient $\mathcal{S}$ un système de preuves à la Hilbert muni du modus ponens, de la substitution uniforme et des axiomes donnés par la figure 2.1, ϕ une formule et Σ un ensemble de formules prouvées dans $\mathcal{S}$.*

Ce système $\mathcal{S}$ est fortement correct *et* fortement complet*, c'est-à-dire vérifie les théorèmes :*

1. *Théorème de correction forte : si $\Sigma \vdash_{\mathcal{S}} \phi$ alors $\Sigma \models \phi$*
2. *Théorème de complétude forte : si $\Sigma \models \phi$ alors $\Sigma \vdash_{\mathcal{S}} \phi$*

Fermeture d'ensembles de formules

Dans le chapitre 3 et le chapitre 5 nous utilisons les notions d'ensembles fermés et de fermeture d'un ensemble. Ainsi, nous en profitons dans cette section pour rappeler ces quelques définitions.

Définition 2.8 - *Ensembles fermés* : *Un ensemble de formules* Σ *est dit* fermé *si, et seulement si, toutes ces propriétés sont respectées :*

1. Σ *est* fermé sur les sous-formules, *c'est-à-dire :*
   - *si* $\phi \vee \psi \in \Sigma$ *alors* $\phi \in \Sigma$ *et* $\psi \in \Sigma$
   - *si* $\phi \wedge \psi \in \Sigma$ *alors* $\phi \in \Sigma$ *et* $\psi \in \Sigma$
   - *si* $\phi \Rightarrow \psi \in \Sigma$ *alors* $\phi \in \Sigma$ *et* $\psi \in \Sigma$
   - *si* $\phi \Leftrightarrow \psi \in \Sigma$ *alors* $\phi \in \Sigma$ *et* $\psi \in \Sigma$
   - *si* $\neg\phi \in \Sigma$ *alors* $\phi \in \Sigma$
   - *si* $\Box\phi \in \Sigma$ *ou* $\Diamond\phi \in \Sigma$ *alors* $\phi \in \Sigma$
2. Σ *est* fermé sur la négation simple, *i.e. si* $\sigma \in \Sigma$ *et* σ *n'est pas de la forme* $\neg\theta$ *alors* $\neg\sigma \in \Sigma$.

Définition 2.9 - *Fermeture* : *Soit* Γ *un ensemble de formules. L'ensemble de formules* $Cl(\Gamma)$ *est appelé* la fermeture de Γ *si, et seulement si,* $Cl(\Gamma)$ *est le plus petit ensemble fermé contenant* Γ.

L'exemple 2.2 présente un ensemble de formules Γ et sa fermeture $\Sigma = Cl(\Sigma)$.

Exemple 2.2 : *Soit* $\Gamma = \{p \Rightarrow (q \wedge \neg r), r \wedge (p \vee q)\}$ *un ensemble de formules où* p, q, *et* r *sont des variables propositionnelles. Si* $\Sigma = Cl(\Gamma)$ *est la fermeture de* Γ, *alors* $\Sigma = \{p \Rightarrow (q \wedge \neg r), r \wedge (p \vee q), p, \neg p, q, \neg q, r, \neg r, \neg(p \Rightarrow (q \wedge \neg r)), \neg(r \wedge (p \vee q)), q \wedge \neg r, \neg(q \wedge \neg r), p \vee q, \neg(p \vee q)\}$ *est le plus petit ensemble fermé contenant* Γ.

Nous avons donc rappelé les principaux résultats sur les systèmes logiques. Ces résultats sont étendus et vérifiés dans les logiques modales. Ainsi, dans la suite de cette section nous présentons des logiques modales et leurs systèmes de preuves à la Hilbert corrects et complets.

2.1.2 Les logiques modales

Les logiques modales sont des extensions des logiques propositionnelles dans lesquelles des opérateurs sont ajoutés au langage. Ces opérateurs, appelés *modalités*, décrivent les notions comme le nécessaire, la possibilité ou encore, le déontique. Le lecteur intéressé peut se référer à la référence (Blackburn et al., 2002). De plus, ces modalités permettent de représenter différents états mentaux des agents comme la connaissance, les croyances ou encore les intentions. La logique modale standard est décrite par le langage $\mathcal{L}_{\Box,\Diamond}$ sur un ensemble de lettres propositionnelles, noté $\mathcal{A}(\mathcal{L}_{\Box,\Diamond}) = \{p, q, r, s, \ldots\}$ et généré par la grammaire sous sa forme de Backus-Naur, pour tout atome $p \in \mathcal{A}(\mathcal{L}_p)$:

$$\phi ::= p|\phi_1 \wedge \phi_2|\phi_1 \vee \phi_2|\phi_1 \Rightarrow \phi_2|\phi_1 \Leftrightarrow \phi_2|\neg\phi|\top|\bot|\Box\phi|\Diamond\phi$$

Dans les logiques modales standards, nous retrouvons deux types de modalités : une modalité de type nécessaire, notée $\Box$; et une modalité de type possible $\Diamond$. Ces deux modalités sont souvent définies comme le dual de l'une par rapport à l'autre. Par exemple, lorsqu'il est nécessaire que

« demain il fasse beau » ($\Box\phi$) alors il est impossible (i.e. non possible) qu'« il ne fasse pas beau » ($\neg\Diamond\neg\phi$). Ainsi utiliser $\Box\phi$ est équivalent à $\neg\Diamond\neg\phi$.

Une modalité va différer d'une autre, par le sens qui lui est conféré. Par exemple, les modalités dites *aléthiques* représentent le nécessaire et le possible ; les modalités dites *épistémiques* représentent la connaissance ; les modalités dites *doxastiques* représentent les croyances ; et les modalités dites *déontiques* décrivent la notion d'obligations et de permissions. Enfin, il existe de nombreuses autres logiques modales comme les logiques temporelles, ou encore les logiques dynamiques. Certaines de ces logiques combinent plusieurs de ces modalités pour former un même système logique, nous parlons alors de logiques *multi-modales*. De manière générale, la sémantique la plus utilisée est celle de Kripke (Kripke, 1959) mais d'autres sémantiques existent comme la sémantique de voisinage (Pacuit, 2017).

Sémantique de Kripke

La sémantique de Kripke introduite par (Kripke, 1959) associe un graphe pour interpréter chaque modalité du langage.

Définition 2.10 - *Cadre de Kripke* : *Nous appelons* cadre de Kripke, *le couple* $\mathcal{C} = (\mathcal{W}, \mathcal{R})$ *dans lequel* $\mathcal{W}$ *représente un ensemble d'éléments, appelés* mondes possibles, *et une relation binaire* $\mathcal{R}$ *définie sur* $\mathcal{W}$, *appelée* relation d'accessibilité *ou encore,* relation d'indiscernabilité.

L'ensemble $\mathcal{W}$ des mondes possibles représente toutes les configurations qu'un agent considère comme possible tandis que la relation d'indiscernabilité associe pour chaque monde possible $w \in \mathcal{W}$, tous les mondes possibles qu'un agent ne peut discerner à partir de ce monde w. Une logique fondée sur un cadre de Kripke est dite *logique modale normale*.

Comme pour les logiques propositionnelles, nous définissons la notion de modèle, de satisfiabilité et de validité d'une formule écrite dans un langage de logique modale.

Définition 2.11 - *Modèle de Kripke* : *Nous appelons* modèle de Kripke $\mathcal{M} = (\mathcal{W}, \mathcal{R}, V)$ *avec* $V : \mathcal{A}(\mathcal{L}_{\Box,\Diamond}) \to 2^{\mathcal{W}}$ *une fonction appelée* fonction de valuation. *Pour tout monde* $w \in \mathcal{W}$, *pour toute formule* $\phi, \psi \in \mathcal{L}_{\Box,\Diamond}$ *et pour tout atome* $p \in \mathcal{A}(\mathcal{L}_{\Box,\Diamond})$.

0. $\mathcal{M}, w \models \top$
1. $\mathcal{M}, w \not\models \bot$
2. $\mathcal{M}, w \models p$ *ssi* $\mathcal{M}, w \in i(p)$
3. $\mathcal{M}, w \models \neg\phi$ *ssi* $\mathcal{M}, w \not\models \phi$
4. $\mathcal{M}, w \models \phi \vee \psi$ *ssi* $\mathcal{M}, w \models \phi$ *ou* $\mathcal{M}, w \models \psi$
5. $\mathcal{M}, w \models \phi \wedge \psi$ *ssi* $\mathcal{M}, w \models \phi$ *et* $\mathcal{M}, w \models \psi$
6. $\mathcal{M}, w \models \phi \Rightarrow \psi$ *ssi* $\mathcal{M}, w \models \neg\phi$ *ou* $\mathcal{M}, w \models \psi$
7. $\mathcal{M}, w \models \Box\phi$ *ssi* $\forall v \in \mathcal{W}, w\mathcal{R}v : \mathcal{M}, v \models \phi$
8. $\mathcal{M}, w \models \Diamond\phi$ *ssi* $\exists v \in \mathcal{W}, w\mathcal{R}v : \mathcal{M}, v \models \phi$

L'exemple 2.3 illustre la notion de cadre de Kripke et de modèle dans un contexte où nous considérons une modalité associée à la connaissance d'un agent.

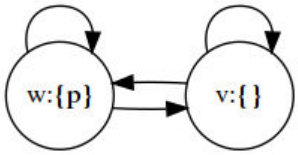

FIGURE 2.2 – Exemple de structure de Kripke

Exemple 2.3 : *Considérons une variable propositionnelle p qui représente le fait : « il fait beau dehors ». Supposons que l'agent raisonnant ne sait pas s'il fait beau dehors. Pour l'agent, il existe alors deux mondes possibles : un monde dans lequel la variable p est vraie et noté w, i.e. le monde où « il fait beau dehors » ; un monde possible dans lequel p est fausse et noté v, i.e. le monde où « il ne fait pas beau dehors ». Pour représenter l'état mental de non connaissance de l'agent, nous considérons l'ensemble de mondes possibles $\mathcal{W} = \{w, v\}$ et la fonction de valuation V est telle que $V(p) = \{w\}$. Dans cet exemple, nous représentons $w = \{p\}$ et $v = \{\}$. De plus, puisque l'agent ne sait pas s'il beau dehors, l'agent ne peut pas discerner entre le monde dans lequel « il fait beau » (i.e. w) du monde dans lequel « il ne fait pas beau » (i.e. v). Ainsi cet état de connaissance est représenté par la relation $\mathcal{R} = \{(w, v), (w, w), (v, w), (v, v)\}$ qui est illustrée par la figure 2.2. L'agent ne peut donc pas discerner entre le monde réel qui est soit w, soit v ; de l'autre monde possible.*

Dans un modèle de Kripke $\mathcal{M}$, une formule est vraie pour un monde possible w, noté $\mathcal{M}, w \models \phi$. Contrairement aux logiques propositionnelles qui ne nécessitaient qu'une fonction d'interprétation pour évaluer la vérité d'une formule, un cadre de Kripke nécessite de préciser, en plus, le monde possible par rapport auquel nous nous plaçons. Nous parlons alors de *monde pointé* w dans un modèle $\mathcal{M}$. En effet, une formule ϕ est vraie n'a de sens que par rapport au monde dans lequel nous nous sommes placés. Ainsi cette formule ϕ peut être vraie dans un monde w, i.e. $\mathcal{M}, w \models \phi$ mais fausse dans un autre monde v, $\mathcal{M}, v \not\models \phi$. Les notions de satisfiabilité et de validité sont distinctes de celles des logiques propositionnelles.

Définition 2.12 - *Satisfiabilité et validité* :

1. *Une formule ϕ est* satisfiable *si, et seulement si, il existe un modèle $\mathcal{M}$ et un monde w tel que $\mathcal{M}, w \models \phi$ est vrai ;*
2. *Une formule ϕ est* valide dans un modèle $\mathcal{M}$ *si, et seulement si, pour tout monde $w \in \mathcal{W}$ du modèle $\mathcal{M}$, $\mathcal{M}, w \models \phi$ est vrai ;*
3. *Une formule ϕ est* valide dans un cadre de Kripke $\mathcal{C}$, *noté $\models_{\mathcal{C}} \phi$ ou $\mathcal{C} \models \phi$ si, et seulement si, ϕ est valide dans tout modèle $\mathcal{M}$ du cadre $\mathcal{C}$.*

Remarquons que les logiques modales sont un cas particulier de logiques du premier ordre. En effet, puisque le $\Box$ est interprété comme un pour tout $\forall$, si $\Box\phi$ est vraie dans un monde w, cela signifie que $\forall v$, sur l'univers du discours $\mathcal{W}$, si le prédicat $\mathcal{R}(w, v) = w\mathcal{R}v$ est vérifié alors la formule ϕ est nécessairement vraie dans v.

Les notions de conséquence sémantique (cf. définition 2.2) et de $\mathcal{S}$-déductibilité (cf. définition 2.7) présentées précédemment sont inchangées dans les logiques modales. De plus, dans tout cadre de Kripke $\mathcal{C}$, le *modus ponens* et la *substitution* préservent la validité. Ces règles peuvent donc être intégrées comme des règles de déduction dans un système de preuves fondé sur le langage modal. Cependant, contrairement aux logiques propositionnelles, de nouvelles règles de déductions doivent être considérés dans de tels systèmes comme la règle dite de *nécessitation* qui décrit le fait que si $\vdash \phi$ est un théorème du système alors $\vdash \Box\phi$ est aussi un théorème.

Définition 2.13 - *Règle de nécessitation ou de généralisation* : *Nous appelons* règle de nécessitation *ou encore* règle de généralisation *la règle de déduction suivante :*

$$\frac{(P1) \quad \vdash \quad \phi}{(C1) \quad \vdash \quad \Box\phi}$$

Si $\vdash \phi$ est un théorème prouvé du système, alors $\vdash \Box\phi$ est un théorème prouvé du système.

Remarquons que puisque la définition de la $\mathcal{S}$-déductibilité d'une formule ϕ à partir d'un ensemble Σ est restreinte à des théorèmes prouvés de $\mathcal{S}$, la règle de généralisation peut s'appliquer sans problème sur toutes les formules $\mathcal{S}$-déductibles.

En plus des axiomes du calcul propositionnel donnés en figure 2.1, les logiques modales doivent considérer un nouvel axiome représenté par la formule valide $\models \Box(\phi \Rightarrow \psi) \Rightarrow \Box\phi \Rightarrow \Box\psi$. Cette formule est appelée *axiome K*. Elle traduit que s'il est nécessaire que $\phi \Rightarrow \psi$ et qu'il est nécessaire que ϕ alors il est aussi nécessaire que ψ.

Nous avons donc en résumé que dans tout cadre de Kripke $\mathcal{C}$, le modus ponens, la substitution, la nécessitation et l'axiome K préservent la validité. Un système fondé sur ces règles de déduction et les axiomes du calcul propositionnel en ajoutant l'axiome K est nommé le *système K*. Ce système est le plus petit système logique fondé sur un cadre de Kripke.

Théorème 2.6 - *Correction du système K* :

1. *Le modus ponens et la substitution préservent la validité pour tout cadre de Kripke $\mathcal{C}$;*
2. *La nécessitation préserve la validité pour tout cadre de Kripke $\mathcal{C}$, c'est-à-dire :*

$$\textit{si } \models_{\mathcal{C}} \phi \textit{ est valide, alors } \models_{\mathcal{C}} \Box\phi \textit{ est aussi valide ;}$$

3. *L'axiome K est valide dans tout cadre de Kripke $\mathcal{C}$, c'est-à-dire :*

$$\models_{\mathcal{C}} \Box(\phi \Rightarrow \psi) \Rightarrow \Box\phi \Rightarrow \Box\psi \quad (K)$$

Le système K est correct.

L'exemple 2.4, présente une preuve avec un système à la Hilbert. Les théorèmes démontrés dans cet exemple sont des tautologies qui sont vérifiées dans tout cadre de Kripke. Ces théorèmes seront appliqués au chapitre 4.

Exemple 2.4 :

Pour tout cadre de Kripke $\mathcal{C}$, les formules suivantes sont des théorèmes :

1. $\vdash \Box\phi \wedge \Box\psi \equiv \Box(\phi \wedge \psi)$
2. $\vdash (\Box\phi \vee \Box\psi) \Rightarrow \Box(\phi \vee \psi)$

Démonstration. Le sens ($\Rightarrow$) pour la propriété 1 se déduit de :

1. $\vdash \phi \Rightarrow (\psi \Rightarrow (\phi \wedge \psi))$
2. $\vdash \Box(\phi \Rightarrow (\psi \Rightarrow (\phi \wedge \psi)))$
3. $\vdash \Box\phi \Rightarrow \Box(\psi \Rightarrow (\phi \wedge \psi))$
4. $\vdash \Box(\psi \Rightarrow (\phi \wedge \psi)) \Rightarrow \Box\psi \Rightarrow \Box(\phi \wedge \psi)$
5. $\vdash \Box\phi \Rightarrow \Box\psi \Rightarrow \Box(\phi \wedge \psi)$
6. $\vdash \Box\phi \wedge \Box\psi \Rightarrow \Box\phi$
7. $\vdash \Box\phi \wedge \Box\psi \Rightarrow \Box(\phi \wedge \psi)$

Le sens ($\Leftarrow$) pour la propriété 1 se déduit de :

1. $\vdash \phi \wedge \psi \Rightarrow \phi$ et $\vdash \phi \wedge \psi \Rightarrow \psi$
2. $\vdash \Box(\phi \wedge \psi) \Rightarrow \Box\phi$ et $\Box(\phi \wedge \psi) \Rightarrow \Box\psi$ (via K et MP)
3. $\vdash (\Box(\phi \wedge \psi) \Rightarrow \Box\phi) \wedge (\Box(\phi \wedge \psi) \Rightarrow \Box\psi) \Rightarrow (\Box(\phi \wedge \psi) \Rightarrow \Box\phi \wedge \Box\psi)$
4. $\vdash \Box(\phi \wedge \psi) \Rightarrow \Box\phi \wedge \Box\psi$

Enfin la propriété 2 se déduit immédiatement de :

1. $\vdash \phi \Rightarrow \phi \vee \psi$ et $\vdash \psi \Rightarrow \phi \vee \psi$
2. $\vdash \Box\phi \Rightarrow \Box(\phi \vee \psi)$ et $\vdash \Box\psi \Rightarrow \Box(\phi \vee \psi)$
3. $\vdash (\Box\phi \vee \Box\psi) \Rightarrow \Box(\phi \vee \psi)$

□

En ajoutant certaines contraintes sur la relation d'accessibilité du cadre de Kripke comme, par exemple, la réflexivité, la transitivité ou encore la sérialité, il est possible d'obtenir de nouvelles formules valides permettant de considérer d'autres systèmes logiques corrects. Ainsi dans un cadre de Kripke sériel $\mathcal{D}$ (i.e. dont la relation d'accessibilité est sérielle), la formule $\models_{\mathcal{D}} \Box\phi \Rightarrow \neg\Box\neg\phi$, appelée *axiome D*, est toujours valide. Le système logique obtenu est alors nommé le système KD qui est donc la combinaison de l'axiome K et de l'axiome D. L'ensemble de ces correspondances entre axiomatique et leur contrainte associée sur la sémantique, est appelé *l'ensemble des correspondances de Sahlqvist*. La figure 2.1 illustre un ensemble d'axiomes usuels avec leur contrainte correspondante sur la relation d'accessibilité pour préserver la validité de l'axiome dans le système. Pour toute combinaison de ces axiomes, il est prouvé que le système logique résultant est correct, complet mais aussi fortement correct et fortement complet (Blackburn et al., 2002). Parmi les systèmes usuels des logiques modales, nous retrouvons les systèmes $K4$, KT, $KT4$, $KD45$ qui sont décrits par la combinaison des axiomes K, T, D, 4 et 5. Par exemple, le système

	Formule	Nom	Propriété de $\mathcal{R}$
D	$\Box\phi \rightarrow \Diamond\phi$	Sérialité	$\forall w \in \mathcal{W}, \exists w' \in \mathcal{W} : w\mathcal{R}w'$
T	$\Box\phi \rightarrow \phi$	Réflexivité	$\forall w \in \mathcal{W}, w\mathcal{R}w$
4	$\Box\phi \rightarrow \Box\Box\phi$ ou $\Diamond\Diamond\phi \rightarrow \Diamond\phi$	Transitivité	$\forall w, w', w'' \in \mathcal{W}, wRw' \wedge w'\mathcal{R}w'' \Rightarrow w\mathcal{R}w''$
5	$\Diamond\phi \rightarrow \Box\Diamond\phi$	Caractère Euclidien	$\forall w, w', w'' \in \mathcal{W}, w\mathcal{R}w' \wedge w\mathcal{R}w'' \Rightarrow w'\mathcal{R}w''$
B	$\phi \rightarrow \Box\Diamond\phi$	Symétrie	$\forall w, w' \in \mathcal{W}, w\mathcal{R}w' \Rightarrow w'\mathcal{R}w$
CD	$\Diamond\phi \rightarrow \Box\phi$	Caractère fonctionnel	$\forall w, w', w'' \in \mathcal{W}, w\mathcal{R}w' \wedge w\mathcal{R}w'' \Rightarrow w' = w''$
$\Box M$	$\Box(\Box\phi \rightarrow \phi)$	Pseudo réflexivité	$\forall w, w' \in \mathcal{W}, w\mathcal{R}w' \Rightarrow w'\mathcal{R}w'$
C4	$\Box\Box\phi \rightarrow \Box\phi$ ou $\Diamond\phi \rightarrow \Diamond\Diamond\phi$	Densité	$\forall w, w' \in \mathcal{W}, w\mathcal{R}w' \Rightarrow \exists w'' \in \mathcal{W}, w\mathcal{R}w'' \wedge w''\mathcal{R}w'$
G	$\Diamond\Box\phi \rightarrow \Box\Diamond\phi$	Confluence	$\forall w, w', w'' \in \mathcal{W}, w\mathcal{R}w' \wedge w\mathcal{R}w''' \Rightarrow \exists w''' \in \mathcal{W}, w'\mathcal{R}w''' \wedge w''\mathcal{R}w''$

Tableau 2.1 – Correspondances de Sahlqvist

$S5$ considère la combinaison des axiomes K, T, 4 et 5. Dans un tel système, la relation d'accessibilité est donc transitive, réflexive et possède le caractère euclidien. La relation d'accessibilité dans $S5$ est donc une relation d'équivalence (i.e. transitive, réflexive et symétrique).

Plus généralement, ces propriétés standards comme la transitivité, ou la réflexivité peuvent être représentées sous une forme dite *klmn-incestueuse* comme le présente l'exemple 2.5. Une relation binaire $\mathcal{R}$ est dite *klmn-incestueuse* si, et seulement si :

$$\forall x, y, z \in \mathcal{W}, [(x\mathcal{R}^k y \wedge x\mathcal{R}^m z) \Rightarrow \exists t \in \mathcal{W}(y\mathcal{R}^l t \wedge z\mathcal{R}^n t)]$$

Ainsi, il est prouvé que dans tout cadre G *klmn*-incestueux, la formule nommée *Axiome de Geach* $\models_G \Diamond^k\Box^l\phi \Rightarrow \Box^m\Diamond^n\phi$ est toujours valide.

Exemple 2.5 : *Par exemple, les contraintes présentées précédemment peuvent toutes être réécrites sous leur forme klmn – incestueuses :*

— $\mathcal{R}$ *est réflexive ssi* $\models \Box\phi \Rightarrow \phi$ *ssi* $\mathcal{R}$ *est une relation* $0100 - incestueuse$ *;*

— $\mathcal{R}$ *est transitive ssi* $\models \Box\phi \Rightarrow \Box\Box\phi$ *ssi* $\mathcal{R}$ *est une relation* $0120 - incestueuse$ *;*

— $\mathcal{R}$ *est Euclidienne ssi* $\models \neg\Box\phi \Rightarrow \Box\neg\Box\phi$ *ssi* $\mathcal{R}$ *est une relation* $1011 - incestueuse.$

Nous avons présenté la sémantique de Kripke, ainsi que des théorèmes toujours vérifiés dans les cadres de Kripke comme l'axiome K ou les propriétés de l'exemple 2.4. Cependant, cette sémantique de Kripke ne suffit pas toujours pour modéliser certaines modalités comme les capacités d'un agent. Nous avons alors besoin de définir d'autres sémantiques comme les sémantiques de voisinage.

Sémantiques de voisinage

Dans certaines situations, si un agent est capable de venir en vélo au laboratoire et qu'il est aussi capable de venir en voiture au laboratoire, cela ne signifie pas que cet agent est capable de réaliser les deux à la fois. Aussi, nous ne souhaitons pas avoir certains théorèmes standards des logiques modales comme l'axiome K, la nécessitation ou encore les propriétés démontrées dans l'exemple 2.4. De ce fait, il existe une sémantique permettant de s'affranchir de tous ces théorèmes, nommée *sémantique minimale*, ou encore, *sémantique de voisinage* (Pacuit, 2017). Nous parlons alors de *logiques modales non normales*, contrairement aux *logiques modales normales* qui sont fondées sur les cadres de Kripke.

Définition 2.14 - *Modèle de voisinage* : *Nous appelons* $\mathcal{M} = (\mathcal{W}, \mathcal{N}, V)$ *un* modèle de voisinage *sur un langage modal* $\mathcal{L}_{\Box,\Diamond}$ *si, et seulement si :*

— $\mathcal{W}$ *est un ensemble de mondes possibles*

— $\mathcal{N} : \mathcal{W} \to 2^{2^{\mathcal{W}}}$ *est une fonction de voisinage*

— $V : \mathcal{A}(\mathcal{L}_{\Box,\Diamond}) \to 2^{\mathcal{W}}$ *une fonction de valuation*

$$\forall w \in \mathcal{W} : \mathcal{M}, w \models \Box\phi \text{ ssi } \{v \in \mathcal{W} : \mathcal{M}, v \models \phi\} \in \mathcal{N}(w)$$

Le couple $\mathcal{C} = (\mathcal{W}, \mathcal{N})$ *est appelé* cadre de voisinage.

Ce modèle permet de caractériser des notions comme la capacité d'un agent à amener un certain état du monde. Nous illustrons dans l'exemple 2.6, issu de (Pacuit, 2017), une situation dans laquelle la notion de capacité d'un agent à amener un certain état du monde peut être représentée avec une sémantique de voisinage. En particulier, lorsque nous représentons cette notion, nous ne voulons pas avoir certains théorèmes comme $\vdash \Box\phi \wedge \Box\psi \Rightarrow \Box(\phi \wedge \psi)$. En effet, si un agent est capable d'amener ϕ à vrai et qu'il est capable d'amener ψ à vrai, cela ne signifie pas qu'il est capable d'amener à la fois ϕ et ψ à vrai.

Exemple 2.6 : *Imaginons un jeu dans lequel deux agents Anne (A) et Bob (B) doivent jouer à tour de rôle et ont deux choix possibles à chaque tour. Anne commence à jouer et après que Bob ait joué, certaines variables sont vraies et d'autres fausses. La figure* **??** *illustre la situation de ce jeu. Ainsi si Anne joue à gauche et Bob joue lui aussi à gauche, l'état du monde, noté* o_1 *a la variable p à vrai. Si Bob avait plutôt joué à droite, dans* o_2*, les variables p et q auraient été vraies toutes les deux à la place.*

Nous remarquons que lorsqu'Anne joue à gauche, peu importe le choix de Bob, la variable p sera toujours vraie. Si elle avait joué à droite, c'est la variable q qui aurait été toujours vraie dans o_3 *et* o_4*. Ainsi, Anne a la possibilité de rendre vraie soit p, soit q mais pas p et q à la fois car Bob a toujours la possibilité de rendre vraie une seule variable uniquement.*

Un joueur a la capacité de forcer une certaine sortie si un joueur a une stratégie qui garantit que le jeu se terminera bien sur cet état. Ainsi, Anne a donc la capacité d'amener les états $\{o_1, o_2\}$ *ou,* $\{o_3, o_4\}$.

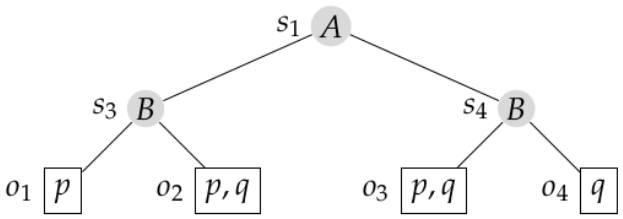

FIGURE 2.3 – Jeu de capacité (Pacuit, 2017)

Pour un modèle de voisinage $\mathcal{M} = (\mathcal{W}, \mathcal{N}, V)$ avec $\mathcal{W} = \mathcal{S} \cup \mathcal{O}$, $\mathcal{S} = \{s_1, s_2, s_3\}$ les états du jeu dans lesquels Anne et Bob peuvent jouer et $\mathcal{O} = \{o_1, o_2, o_3, o_4\}$ les états du jeu dans lesquels le jeu est terminé. Nous définissons $\mathcal{M}, s_1 \models Ab_a p$ si, et seulement si, il existe $X \in \mathcal{N}(s_1)$ tel que $\forall v \in X, \mathcal{M}, v \models p$. Si $\mathcal{N}(s_1) = \{\{o_1, o_2\}, \{o_3, o_4\}\}$, nous avons alors $\mathcal{M}, s_1 \models Ab_a p$ et $\mathcal{M}, s_1 \models Ab_a q$ et $\mathcal{M}, s_1 \models Ab_a p \wedge Ab_a q$. Cependant $Ab_a(p \wedge q)$ est faux dans s_1, i.e. $\mathcal{M}, s_1 \not\models Ab_a(p \wedge q)$. Ainsi, le théorème $Ab_a p \wedge Ab_a q \Rightarrow Ab_a(p \wedge q)$ valide dans tout cadre de Kripke n'est donc pas valide dans les cadres de voisinage.

Par ailleurs, nous notons $|\phi|_{\mathcal{M}} = \{v \in \mathcal{W} : \mathcal{M}, v \models \phi\}$ l'ensemble de tous les mondes de $\mathcal{W}$ dans lesquels ϕ est vraie pour le modèle $\mathcal{M}$. Lorsqu'il n'y a aucune ambiguïté, nous notons simplement l'ensemble $|\phi|$. De plus, la dualité du $\Box$ se définit comme $\mathcal{M}, w \models \Diamond\phi$ si, et seulement si, $\mathcal{W} \setminus |\phi| \notin \mathcal{N}(w)$. Pour démontrer que la dualité, notée (Dual), est préservée par la validité entre le $\Box$ et le $\Diamond$, c'est-à-dire, que la formule $\models \Box\phi \Leftrightarrow \neg\Diamond\neg\phi$ est valide dans tout cadre de voisinage. Nous avons besoin de plusieurs propriétés sur $|\phi|$ (Pacuit, 2017).

Proposition 2.2 : *Soient $\mathcal{M} = (\mathcal{W}, \mathcal{N}, V)$ un modèle de voisinage, pour tout $p \in \mathcal{L}_{\Box,\Diamond}$, et $\phi \in \mathcal{L}_{\Box,\Diamond}$. L'ensemble $|\phi|$ possède les propriétés suivantes :*

1. $|\phi| = V(p)$;
2. $|\neg\phi| = \mathcal{W} \setminus |\phi|$;
3. $|\phi \wedge \psi| = |\phi| \cap |\psi|$;
4. $|\Box\phi| = m_{\mathcal{N}}(|\phi|)$ *avec* $m_{\mathcal{N}} : 2^{\mathcal{W}} \to 2^{\mathcal{W}}$ *et telle que :*

$$\forall X \subseteq \mathcal{W} : m_{\mathcal{N}}(X) = \{v \in \mathcal{W} : X \in \mathcal{N}(v)\}$$

5. $|\Diamond\phi| = \mathcal{W} \setminus m_{\mathcal{N}}(\mathcal{W} \setminus |\phi|)$.

Immédiatement, par application des propriétés issues de 2.2, nous déduisons que la règle (Dual) préserve la validité dans tout cadre de voisinage.

Proposition 2.3 : *Pour tout cadre de voisinage $\mathcal{C}$,*

$$\models_{\mathcal{C}} \Box\phi \Leftrightarrow \neg\Diamond\neg\phi \quad (Dual)$$

	Formule	Propriété sur $\mathcal{N}$
M	$\Box(\phi \wedge \psi) \rightarrow \Box\phi \wedge \Box\psi$	$\forall w \in \mathcal{W} : S \in \mathcal{N}(w) \wedge S \subseteq T \Longrightarrow T \in \mathcal{N}(w)$
C	$\Box\phi \wedge \Box\psi \rightarrow \Box(\phi \wedge \psi)$	$\forall w \in \mathcal{W} : S \in \mathcal{N}(w) \wedge T \in \mathcal{N}(w) \Longrightarrow S \cap T \in \mathcal{N}(w)$
N	$\Box\top$	$\forall w \in \mathcal{W} : \mathcal{W} \in \mathcal{N}(w)$
T	$\Box\phi \Rightarrow \phi$	$\forall w \in \mathcal{W} : S \in \mathcal{N}(w) \Longrightarrow w \in S$

Tableau 2.2 – Propriétés sur la fonction de voisinage

Dans tout cadre de voisinage, une autre règle préserve toujours la validité, notée (RE) et qui consiste à énoncer que si une équivalence $\phi \Leftrightarrow \psi$ est valide, alors il en est de même pour $\Box\phi \Leftrightarrow \Box\psi$ (Pacuit, 2017).

Proposition 2.4 : *Pour tout cadre de voisinage* $\mathcal{C}$*,*

$$Si \ \models_{\mathcal{C}} \phi \Leftrightarrow \psi \ alors \ \models_{\mathcal{C}} \Box\phi \Leftrightarrow \Box\psi \quad (RE)$$

Le plus petit système logique fondé sur une sémantique minimale est nommé **E**. Il est constitué des règles :

— (CP) les tautologies du Calcul Propositionnel ;

— (Dual) la règle de dualité ;

— (MP) la règle du modus ponens ;

— (RE) la règle d'équivalence.

Toutes les règles du système **E** préservent la validité et sont donc vérifiées dans tout cadre de voisinage. De plus, remarquons que la règle de la substitution uniforme (Sub) peut être déduite des règles de **E**, par raisonnement par récurrence.

De la même manière que pour les logiques modales, nous pouvons contraindre la fonction de voisinage à respecter certaines propriétés afin de construire de nouveaux systèmes axiomatiques corrects. De plus, avec la sémantique de voisinage, il est possible de définir de nouveaux systèmes logiques qui ne pouvaient alors être définies avec un cadre de Kripke standard. Certaines contraintes classiques sur les cadres de voisinage sont données dans la figure 2.2. Par exemple, il est possible de définir un système dans lequel la nécessitation préserve la validité en considérant les cadres tels que la propriété $\forall w \in \mathcal{W} : \mathcal{W} \in \mathcal{N}(w)$ est vérifiée. De cette propriété, nous remarquons qu'il est possible, contrairement aux logiques modales normales, de définir une règle dans laquelle $\models \neg\Box\top$ est valide en considérant tout cadre de voisinage tel que $\forall w \in \mathcal{W} : \mathcal{W} \notin \mathcal{N}(w)$ est vérifiée.

Par ailleurs, les cadres de voisinage permettent de redéfinir les systèmes classiques, comme c'est le cas pour le système K. En effet, nous remarquons que la combinaison des axiomes ECMN permet de définir un système logiquement équivalent au système K. Les autres systèmes classiques comme S5 peuvent eux aussi être redéfinis avec une fonction de voisinage. Par exemple, lorsque la

fonction de voisinage est un filtre [2], alors un système axiomatique correct et complet correspond au système S5 (Pacuit, 2017).

Sémantiques multi-modales, temporelles et dynamiques

Il existe encore d'autres sémantiques alternatives aux logiques modales normales et non-normales : les sémantiques de branchement, les sémantiques dynamiques ou encore, les sémantiques multi-modales.

Sémantiques fondées sur les branchements Pour caractériser la temporalité, nous pouvons considérer, par exemple, une relation $\mathcal{R}$ décrivant les états futurs d'un système sous la forme d'un arbre (Blackburn et al., 2002). Ainsi pour caractériser le « toujours vrai dans le futur » à partir d'un monde w, il est nécessaire de considérer tous les états qui succèdent w dans cet arbre. Pour ce faire, si $\mathcal{C} = (\mathcal{W}, \mathcal{R})$ est un arbre, $\mathcal{M} = (\mathcal{W}, \mathcal{R}, V)$ le modèle de Kripke associé au cadre $\mathcal{C}$, $w \in \mathcal{W}$ un monde possible et si le $\Box$ représente cette modalité du « toujours vrai dans le futur », alors cette modalité peut être représenter sémantiquement par la relation :

$$\mathcal{M}, w \models \Box\phi \text{ ssi } \forall v \in \mathcal{W}, w\mathcal{R}^{+}v : \mathcal{M}, v \models \phi, \text{ avec } \mathcal{R}^{+} \text{ la clôture transitive de } \mathcal{R}$$

Si cette sémantique permet de décrire, l'évolution d'un système, par exemple, elle ne considère aucun modèle événementiel pour représenter la conséquence perlocutoire d'un acte de langage sur l'ensemble des croyances des agents dans un système, comme le fait d'affirmer que « demain il va y avoir grève à l'université ».

Sémantiques dynamiques Pour représenter les conséquences des actions sur les états futurs du système, il existe de nombreuses logiques dites *dynamiques*. Nous pouvons par exemple intégrer dans la sémantique un modèle événementiel permettant de considérer les conséquences d'événements dans le système (Van Ditmarsch et al., 2007). Les logiques des annonces publiques sont un exemple de sémantiques dynamiques, leur langage $\mathcal{L}_{PAL}$ consiste en une modalité $\Box$ représentée par la lettre K décrivant le fait qu'un agent sache une proposition ϕ comme étant vraie et une modalité pour l'annonce publique d'une proposition $[\phi!]$. Pour toute lettre propositionnelle p du langage $\mathcal{L}_{PAL}$, ce langage est généré par la grammaire :

$$\phi ::= p|\phi_1 \wedge \phi_2|\neg\phi|\top|\bot|K\phi|[\phi_1!]\phi_2$$

Une formule $[\psi!]\phi$ est lue comme « après l'annonce publique de ψ, ϕ est nécessairement vraie ». Si $\mathcal{M} = (\mathcal{W}, \mathcal{R}, V)$ est un modèle de Kripke, la sémantique de la formule $[\psi!]\phi$ est donnée par la relation d'équivalence suivante :

2. Pour un ensemble de mondes possibles $\mathcal{W}$, une fonction de voisinage $\mathcal{N} : \mathcal{W} \to 2^{2^{\mathcal{W}}}$ est appelée *filtre* sur $\mathcal{W}$ si, et seulement si, les trois propriétés suivantes sont respectées :

1. $\forall w \in \mathcal{W}, \forall S, V \subseteq \mathcal{W}, V \subseteq S, V \in \mathcal{N}(w) \Rightarrow S \in \mathcal{N}(w)$;
2. $\forall w \in \mathcal{W}, \forall S, V \subseteq \mathcal{W}, S, V \in \mathcal{N}(w) \Rightarrow S \cap V \in \mathcal{N}(w)$;
3. $\forall w \in \mathcal{W}, \emptyset \notin \mathcal{N}(w)$.

$$\mathcal{M}, w \models [\psi!]\phi \text{ ssi, si } \mathcal{M}, w \models \psi \text{ alors } \mathcal{M}^{\psi}, w \models \phi$$

avec $\mathcal{M}^{\psi} = (\mathcal{W}^{\psi}, \mathcal{R}^{\psi}, V^{\psi})$ le modèle de Kripke tel que :

— $\mathcal{W}^{\psi} = |\psi|_{\mathcal{M}}$

— $\mathcal{R}^{\psi} = \mathcal{R} \cap (\mathcal{W}^{\psi} \times \mathcal{W}^{\psi})$

— $V^{\psi} = V_{|\mathcal{W}^{\psi}}$

Le modèle $\mathcal{M}^{\psi}$ représente donc le modèle de Kripke après l'annonce de ψ.

Généralisation aux multi-modal et relations n-aires Jusqu'à présent nous n'avons considéré des systèmes de logique modale qui se plaçaient du point d'un unique agent raisonnant. Une extension naturelle des logiques modales consiste à intégrer de nouvelles modalités propres aux autres agents du système (Fagin et al., 2004). Ainsi, un agent peut raisonner sur les connaissances ou croyances des autres agents. Le modèle de Kripke est alors étendu à un ensemble de relations binaires :

$$\mathcal{M} = (\mathcal{W}, \mathcal{R}_1, \ldots, \mathcal{R}_n, V)$$

Pour chaque agent $i \in \mathcal{N}$ du système, numéroté de 1 à n, on lui associe une relation d'accessibilité $\mathcal{R}_i$ et une modalité $\Box_i$ dans le langage modal.

Puisque les opérateurs modaux sont définis généralement comme des prédicats d'arité 1, il est possible de la même façon de considérer des opérateurs modaux d'arité quelconque (Blackburn et al., 2002). Un opérateur $\Box$ d'arité m est alors représenté par le symbole ∇ et est défini pour un modèle $\mathcal{M} = (\mathcal{W}, \mathcal{R}_{\nabla}, V)$ où $\mathcal{R}_{\nabla}$ est une relation $(m+1)$-aire, pour tout $\forall w \in \mathcal{W}$ monde possible, nous avons :

$$\mathcal{M}, w \models \nabla(\phi_1, \ldots, \phi_m) \text{ ssi } \forall v_1, \ldots, v_m \in \mathcal{W}, w\mathcal{R}_{\nabla}v_1, \ldots, v_m : \forall i \in [|1, m|]\mathcal{M}, v_i \models \phi_i$$

Le dual de ∇ est alors noté $\triangle$ et est défini tel que :

$$\mathcal{M}, w \models \triangle(\phi_1, \ldots, \phi_m) \text{ ssi } \exists v_1, \ldots, v_m \in \mathcal{W}, w\mathcal{R}_{\nabla}v_1, \ldots, v_m : \forall i \in [|1, m|]\mathcal{M}, v_i \models \phi_i$$

Toutes les propriétés de correction et de complétude associées aux logiques modales normales sont alors généralisables pour les opérateurs modaux d'arité n, pour tout $n \in \mathbb{N}$. De plus, ces extensions existent aussi pour les logiques modales non-normales (Pacuit, 2017).

2.1.3 Les logiques modales pour agents cognitifs

Dans cette thèse, nous nous intéressons aux logiques modales qui expriment certains états mentaux des agents cognitifs comme les connaissances, les croyances, les intentions et les émotions tels qu'ils ont été présentés au chapitre 1. De plus, puisque certaines stratégies de manipulation peuvent reposer sur les normes d'un système, nous présentons les logiques déontiques qui sont des logiques modales exprimant les notions d'obligations et de permissions.

Logiques doxastiques et épistémiques

Les logiques modales dédiées à la représentation des connaissances et des croyances des agents sont appelées respectivement : les *logiques épistémiques* et les *logiques doxastiques*.

Logiques épistémiques Les logiques épistémiques représentent donc la connaissance qu'un agent possède sur son environnement ou sur ce que savent les autres agents. Le lecteur intéressé peut se référer à (Fagin et al., 2004). La connaissance d'un agent i est représentée par la modalité K_i dans le langage $\mathcal{L}_{EL}$:

$$\phi ::= p|\phi_1 \wedge \phi_2|\neg\phi|\top|\bot|K_i\phi$$

Souvent, la sémantique associée à l'interprétation de cette modalité consiste à considérer la relation d'indiscernabilité $\mathcal{K}_i$ associée à K_i comme une relation d'équivalence (i.e. une relation transitive, réflexive et symétrique). Ainsi, nous obtenons un système axiomatique S5 qui est constitué des axiomes K, T, 4 et 5 avec la règle du *modus ponens*, de la nécessitation, la substitution uniforme et toutes les tautologies du calcul propositionnel. Un modèle définit pour les logiques épistémiques est un N-uplet $\mathcal{M} = (\mathcal{W}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, V\}$ tel que pour tout $i \in \mathcal{N}$, $\mathcal{K}_i$ est une relation d'équivalence.

Les logiques épistémiques permettent aussi de représenter des notions de connaissances générales, de connaissances communes, de connaissances partagées, ou encore, des notions de connaissances distribuées. Par exemple, la connaissance générale se définit comme le prédicat :

$$G\phi := \bigwedge_{i\in\mathcal{N}} K_i\phi$$

La connaissance commune se définit comme :

$$C\phi := \bigwedge_{n\in\mathbb{N}^*} G^n\phi = G\phi \wedge GG\phi \wedge \ldots \wedge G\ldots G\phi$$

La connaissance distribuée dans un groupe d'agents $G \subseteq \mathcal{N}$ se définit quant-à-elle comme le prédicat, noté $D_G\phi$, représentant l'ensemble de toutes les connaissances des agents du groupe G. Toutes ces notions peuvent alors être définies dans un modèle $\mathcal{M}$ tel que, pour tout monde $w \in \mathcal{W}$, pour tout agent $i \in \mathcal{N}$ et pour tout groupe d'agents $G \subseteq \mathcal{N}$, nous avons :

1. $\mathcal{M}, w \models K_i\phi$ si, et seulement si, $\forall v \in \mathcal{K}_i(w), \mathcal{M}, v \models \phi$;
2. $\mathcal{M}, w \models C_G\phi$ si, et seulement si, $\forall v \in (\bigcup_{i\in G} \mathcal{K}_i(w))^+, \mathcal{M}, v \models \phi$,
   avec $(\bigcup_{i\in G} \mathcal{K}_i(w))^+$ la clôture transitive de l'union des $\{\mathcal{K}_i\}_{i\in G}$;
3. $\mathcal{M}, w \models D_G\phi$ si, et seulement si, $\forall v \in \bigcap_{i\in G} \mathcal{K}_i(w), \mathcal{M}, v \models \phi$.

Une extension possible des logiques épistémiques consiste à intégrer un modèle dynamique, nous parlons alors de *logiques épistémiques dynamiques*. De telles logiques sont constituées d'un modèle des événements, ainsi que d'un produit de mise à jour avec lequel les agents vont actualiser leurs connaissances (Van Ditmarsch et al., 2007). Ces logiques possèdent trois composantes :

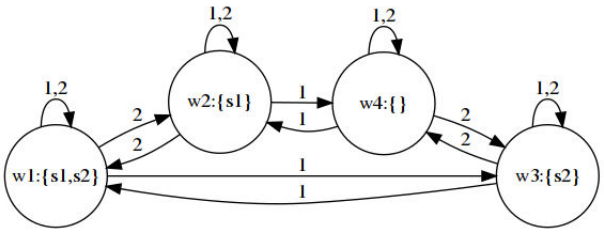

FIGURE 2.4 – Modèle des connaissances des enfants sales

1. Un modèle de Kripke multi-modal pour la logique épistémique $\mathcal{M} = (\mathcal{W}, \mathcal{K}_1, \ldots \mathcal{K}_{|\mathcal{N}|}, V)$ décrivant les connaissances de chacun des agents ;
2. Un modèle événementiel $\mathcal{E} = (\mathcal{W}^\alpha, \mathcal{K}_1^\alpha, \ldots \mathcal{K}_{|\mathcal{N}|}^\alpha, I^\alpha\}$ décrivant la manière dont les événements peuvent influer sur les connaissances des agents où $\mathcal{W}^\alpha$ représente un ensemble non vide d'événements possibles, pour tout $i \in \mathcal{N}$, $\mathcal{K}_i^\alpha$ est une relation d'accessibilité entre les événements possibles, et $I^\alpha : \mathcal{W}^\alpha \to \mathcal{E}_{EL}$ est appelée *la fonction de précondition* et assigne à chaque événement possible une formule de $\mathcal{L}_{EL}$;
3. Un produit de mise à jour des connaissances, tenant compte du modèle de la situation $\mathcal{M}$ et des événements possibles $\mathcal{E}$, et représentant les nouvelles connaissances des agents après l'exécutions des événements possibles. Il se définit par $\mathcal{M} \otimes \mathcal{E} = \{\mathcal{W}^\otimes, \mathcal{K}_1^\otimes, \ldots, \mathcal{K}_{|\mathcal{N}|}^\otimes, I^\otimes\}$ avec :
   - $\mathcal{W}^\otimes = \{(v, f) \in \mathcal{W} \times \mathcal{W}^\alpha : \mathcal{M}, v \models I^\alpha(f)\}$
   - $\forall (v, f) \in \mathcal{W}^\otimes, \forall i \in \mathcal{N}, \mathcal{K}_i^\otimes(v, f) = \{(u, g) \in \mathcal{W}^\otimes : u \in \mathcal{K}_i(v), g \in \mathcal{K}_i^\alpha(f)\}$
   - $\forall (v, f) \in \mathcal{W}^\otimes : V^{\mathcal{W}^\otimes}(v, f) = V(v)$

Exemple 2.7 : *Le problème des enfants sales consiste en une situation dans laquelle, deux enfants face à leur père ne savent pas s'ils sont sales eux-mêmes mais peuvent savoir si l'autre enfant l'est (Van Ditmarsch et al., 2005). Nous considérons s_1 et s_2 deux variables propositionnelles pour décrire respectivement que l'enfant 1 est sale et que l'enfant 2 est sale. Cette situation est alors représentée par le modèle $\mathcal{M}$ décrit par la structure de la figure 2.4. Dans cette situation, nous retrouvons quatre mondes possibles $\mathcal{W} = \{w_1, w_2, w_3, w_4\}$:*

- *le monde dans lequel les deux enfants sont sales $w_1 : s_1, s_2$*
- *le monde dans lequel seul l'enfant 1 est sale $w_2 : s_1$*
- *le monde dans lequel seul l'enfant 2 est sale $w_3 : s_2$*
- *le monde dans lequel aucun des deux enfants n'est sale $w_4 : \emptyset$*

Du point de vue des enfants, chacun ne peut pas distinguer des mondes possibles dans lesquels lui-même est sale du monde dans lequel il ne l'est pas. Nous avons donc les relations $\mathcal{K}_1 = \{(w_1, w_1), (w_1, w_3), (w_3, w_3), (w_3, w_1), (w_4, w_2), (w_2, w_4), (w_2, w_2), (w_4, w_4)\}$, $\mathcal{K}_2 = \{(w_1, w_1), (w_1, w_2), (w_2, w_2), (w_2, w_1), (w_4, w_3), (w_3, w_4), (w_3, w_3), (w_4, w_4)\}$.

À ce modèle, ajoutons un modèle événementielle $\mathcal{E}$ pour décrire le fait que le père annonce uniquement à 1 qu'il est sale sans même que l'enfant 2 ne distingue quoique ce soit. Ce modèle est décrit par la structure donnée par la figure 2.5, dans laquelle, après l'annonce du père, l'enfant 1 sait qu'il est sale mais l'enfant 2 ne remarque rien. Les événements possibles $\mathcal{W}^\alpha = \{e_1, e_2\}$ sont alors les suivants : tous les mondes possibles dans lesquels 1 est sale, l'enfant 1 ne peut alors que discerner e_1 ; tous les autres mondes possibles restent inchangés pour les enfants dans l'événement e_2. L'enfant 1 ne peut pas discerner un autre événement possible que e_1 alors que pour l'enfant 2, puisque celui-ci n'a rien remarqué, il ne peut discerner un autre événement que e_2 , c'est-à-dire l'ensemble des mondes possibles inchangés. Nous avons que dans la figure 2.5 les relations entre événements sont données par $\mathcal{K}_1^\alpha = \{(e_1, e_1), (e_2, e_2)\}$ et $\mathcal{K}_j^\alpha = \{(e_1, e_2), (e_2, e_2)\}$. Pour représenter cet événement la fonction de précondition est alors $I^\alpha = \{(e_1, s_1), (e_2, \top)\}$. $I^\alpha(e_2) = \top$ signifie que e_2 concerne tous les mondes possibles de $\mathcal{W}$.

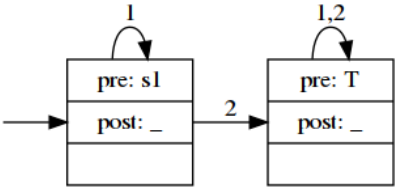

FIGURE 2.5 – Modèle événementiel

Après l'événement produit, la mise à jour décrite par le produit $\mathcal{M}^\otimes = \mathcal{M} \otimes \mathcal{E}$ nous donne alors le nouveau modèle illustré dans la figure 2.6. Dans la configuration où l'événement e_1 s'est produit, l'enfant 1 sait qu'il est sale, l'enfant 2 ne sait pas qu'un tel événement s'est produit. Le modèle nous décrit alors les éléments suivants :

- *l'ensemble de mondes possibles $\mathcal{W}^\otimes = \{(w_1, e_1), (w_2, e_1), (w_1, e_2), (w_2, e_2), (w_3, e_2), (w_4, e_2)\}$;*
- *$\mathcal{K}_1^\otimes = \{((w_1, e_1), (w_1, e_1)); ((w_2, e_1), (w_2, e_1)); ((w_1, e_2), (w_1, e_2)); \ldots\}$;*
- *$\mathcal{K}_2^\otimes = \{((w_1, e_1), (w_1, e_2)); ((w_1, e_1), (w_2, e_2)); ((w_2, e_1), (w_1, e_2)); ((w_2, e_1), (w_2, e_2)); ((w_1, e_2), (w_1, e_2)); \ldots\}$;*
- *La fonction de valuation est donnée par les valuations des mondes w_1, w_2, w_3, et w_4.*

Logiques doxastiques Les logiques doxastiques représentent les croyances via une modalité B_i associée à un agent i (Hintikka, 1965). Contrairement aux logiques épistémiques, les modalités doxastiques ne représentent plus la vérité que connaît un agent mais ce qu'il pense comme possible sans en être certain. Ainsi, l'axiome de vérité (T) est atténué dans les logiques doxastiques. Un système logique très souvent utilisé pour représenter les croyances est le système KD45 où les croyances sont introspectives.

Un modèle doxastique est donc un modèle de Kripke multi-modal $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i \in \mathcal{N}}, V)$ décrivant les croyances de chacun des agents avec pour tout $i \in \mathcal{N}$, $\mathcal{B}_i$ est une relation sérielle, transitive et euclidienne. Un *modèle doxastique multiple pointé* est un couple $(\mathcal{M}, U)$ où $U \subseteq \mathcal{W}$.

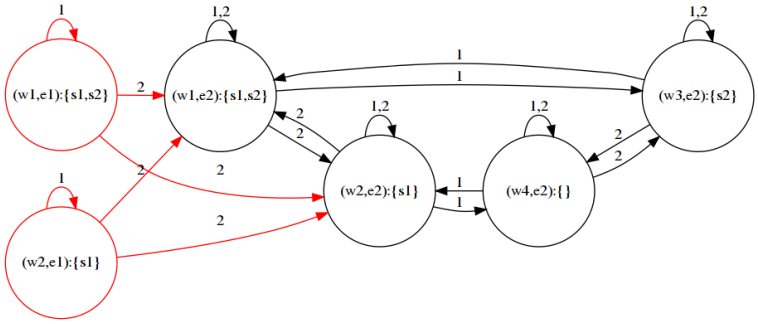

FIGURE 2.6 – Modèle des connaissances des enfants après l'annonce du père

Pour chaque $\mathcal{B}_i$ nous représentons $\xrightarrow{i}$. De plus, nous notons $\mathbb{M}$ la classe des modèles doxastiques multiple pointés.

De façon analogue aux logiques épistémiques, les logiques doxastiques peuvent être étendues aux logiques doxastiques dynamiques. Ces dernières sont associées à un modèle d'action qui est un modèle événementiel $\mathcal{A} = (\mathcal{W}^\alpha, \{\mathcal{B}_i^\alpha\}_{i\in\mathcal{N}}, I^\alpha, P^\alpha)$ et décrit la manière dont les événements peuvent influer sur les croyances des agents. L'ensemble $\mathcal{W}^\alpha$ représente un ensemble non vide d'événements possibles ou d'actions. Pour tout agent $i \in \mathcal{N}$, $\mathcal{B}_i^\alpha$ est une relation d'accessibilité entre les événements qui est transitive, sérielle et euclidienne. La fonction $I^\alpha : \mathcal{W}^\alpha \to \mathcal{L}_B$ est *la fonction de précondition* qui assigne à chaque événement possible une formule de $\mathcal{L}_B$. Enfin, $P^\alpha : \mathcal{W}^\alpha \to Sub_{\mathcal{L}_B}$ est une fonction qui associe à chaque événement possible une substitution qui représente les conséquences des actions sur les variables, cette fonction est appelée *la fonction de postcondition*. $Sub_{\mathcal{L}_B}$ représente l'ensemble des substitutions sur le langage $\mathcal{L}_B$. Si $\mathcal{A}(\mathcal{L}_B) = \{p_1, \ldots, p_n\}$ est l'ensemble des variables propositionnelles de $\mathcal{L}_B$, les éléments de $Sub_{\mathcal{L}_B}$ sont de la forme :

$$\{p_1 \mapsto \sigma_1, \ldots, p_n \mapsto \sigma_n | \sigma_1, \ldots, \sigma_n \in \mathcal{L}_B\}$$

Nous notons $\mathbb{A}$ l'ensemble des modèles d'action multiple pointés et finis et $\mathcal{L}_B$ est le langage qui pour tout agent $i \in \mathcal{N}$, pour toute variable propositionnelle $p \in \mathcal{A}(\mathcal{L}_B)$ et pour tout modèle d'action multiple pointés $(\mathcal{A}, S) \in \mathbb{A}$ avec $S \subseteq \mathcal{W}^\alpha$:

$$\phi ::= p | \phi_1 \wedge \phi_2 | \neg\phi | \top | \bot | [\alpha]\phi | [A, S]\phi$$

$$\alpha ::= i | ?\alpha | \alpha_1; \alpha_2 | \alpha_1 \cup \alpha_2 | \alpha^*$$

Syntaxiquement, $[A, S]\phi$ exprime que la formule ϕ est vraie après application du modèle d'action A aux ensembles d'événements de $S \subseteq \mathcal{W}^\alpha$. Nous représentons usuellement pour chaque agent $i \in \mathcal{N}$, $[i]\phi \triangleq B_i\phi$ et $[(\bigcup_{i\in\mathcal{N}} i)^*]\phi$ l'opérateur de croyances communes. Les formules suivantes sont associées aux relations :

— $\forall i \in \mathcal{N} : \xrightarrow{i} = \mathcal{B}_i$

— $?\phi := \{(w, w) | \mathcal{M}, w \models \phi\}$

— $\alpha_1;\alpha_2 := \{(x,y)|\exists z(x \xrightarrow{\alpha_1} z) \wedge (z \xrightarrow{\alpha_2} y)\}$
— $\xrightarrow{\alpha_1 \cup \alpha_2} = \xrightarrow{\alpha_1} \cup \xrightarrow{\alpha_2}$
— $\xrightarrow{\alpha^*}$ est la clôture transitive réflexive de $\xrightarrow{\alpha}$

Le produit de mise à jour est défini comme l'application $\otimes : \mathbb{M} \times \mathbb{A} \to \mathbb{M}$ telle que pour tout modèle doxastique $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, V)$ pointé sur $U \subseteq \mathcal{W}$ et pour tout modèle d'action $\mathcal{A} = (\mathcal{W}^\alpha, \{\mathcal{B}_i^\alpha\}_{i\in\mathcal{N}}, I^\alpha, P^\alpha)$ pointé sur $S \subseteq \mathcal{W}^\alpha$, nous avons : $(\mathcal{M},U)\otimes(\mathcal{A},S) = ((\mathcal{W}^\otimes, \{\mathcal{B}_i^\otimes\}_{i\in\mathcal{N}}, V^\otimes), U^\otimes)$ où :

— $\mathcal{W}^\otimes = \{(v,f) \in \mathcal{W} \times \mathcal{W}^\alpha : \mathcal{M}, v \models I^\alpha(f)\}$
— $\forall (v,f) \in \mathcal{W}^\otimes, \forall i \in \mathcal{N}$,
$\mathcal{B}_i^\otimes(v,f) = \{(u,g) \in \mathcal{W}^\otimes : u \in \mathcal{B}_i(v), g \in \mathcal{B}_i^\alpha(f)\}$
— $\forall (v,f) \in \mathcal{W}^\otimes : V^\otimes(v,f) = \{p \in \mathcal{A}(\mathcal{L}_B) | \mathcal{M}, w \models P(s)(p)\}$
— $U^\otimes := \{(v,f)|v \in U, f \in S, (v,f) \in \mathcal{W}^\otimes\}$

L'interprétation des formules est donnée par :

1. $\mathcal{M}, w \models [\alpha]\phi$ ssi $\forall v \in \mathcal{W} : w \xrightarrow{\alpha} v, \mathcal{M}, v \models \phi$
2. $\mathcal{M}, w \models [\mathcal{A}, S]\phi$ ssi $(\mathcal{W}^\otimes, \{\mathcal{B}_i^\otimes\}_{i\in\mathcal{N}}, V^\otimes), (w,s) \models \phi$ pour tout $(w,s) \in U^\otimes$ et $((\mathcal{W}^\otimes, \{\mathcal{B}_i^\otimes\}_{i\in\mathcal{N}}, V^\otimes), U^\otimes) = (\mathcal{M}, \{w\}) \otimes (\mathcal{A}, S)$

Exemple 2.8 : *La figure 2.7 illustre la situation dans laquelle un agent a (accusé) sait une certaine proposition p (i.e. a croit p et p est vraie), par exemple p :=« l'accusé a était bien sur la scène de crime » alors que les autres agents b et c ne savent pas si cette proposition p est vraie i.e.* $\neg B_{\{b,c\}}p \wedge \neg B_{\{b,c\}}\neg p$ *avec* $B_{\{b,c\}}\phi \triangleq B_b\phi \vee B_c\phi$. *Le nœud gris de la figure illustre le monde actuel (monde* (0)*) , i.e. l'accusé était bien sur la scène de crime. L'accusé parvient à fait croire*

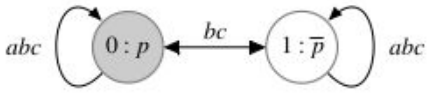

FIGURE 2.7 – Exemple de modèle pointé doxastique $(\mathcal{M}, \{(0)\})$ (Van Ditmarsch et al., 2012)

aux agents b et c qu'il n'était pas sur la scène de crime. Un modèle événementiel associé à cette situation est décrit par la figure 2.8. L'événement actuel en gris $[0]$*, est donc caractérisé par le fait que les agents b et c ne peuvent discerner que des mondes possibles où* $\neg p$ *est vraie.*

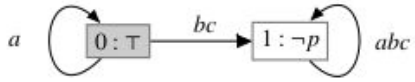

FIGURE 2.8 – Exemple de modèle pointé événementiel $(\mathcal{A}, \{[0]\})$ (Van Ditmarsch et al., 2012)

Le modèle mis à jour des croyances des agents est décrit dans la figure 2.9 : l'agent a est le seul à croire que p est vraie tandis que les autres agents b et c pensent le contraire.

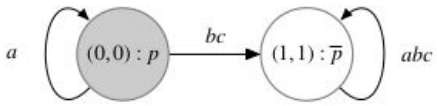

FIGURE 2.9 – Exemple de « product update » sur les modèles pointés $(\mathcal{M}, \{(0)\}) \otimes (\mathcal{A}, \{[0]\}) = (\mathcal{M}^{\otimes}, \{((0), [0])\})$ (Van Ditmarsch et al., 2012)

Combinaisons entre logiques épistémiques et logiques doxastiques Les logiques doxastiques et les logiques épistémiques ne sont pas mutuellement exclusives et peuvent être combinées ensemble. (Stalnaker, 2006) propose un système logique qui inclut une modalité K_i et une modalité B_i pour les croyances des agents. Un modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, V)$ est constitué, pour chaque agent $i \in \mathcal{N}$, d'une relation d'équivalence $\mathcal{K}_i$ pour chaque modalité de connaissance K_i et d'une relation sérielle, transitive et euclidienne $\mathcal{B}_i$ pour chaque modalité de croyance B_i. Il montre que des liens logiques peuvent être construits entre ces deux modalités. Ainsi pour tout agent $i \in \mathcal{N}$ et pour tout $w \in \mathcal{W}$, un modèle $\mathcal{M}$ pour représenter un système de la croyance et de la connaissance doit être tel que les propriétés suivantes doivent être vérifiées :

— $\mathcal{B}_i(w) \subseteq \mathcal{K}_i(w)$, un agent croit ce qu'il sait ;

— $\forall v, u \in \mathcal{W} : w\mathcal{K}_i u \wedge u\mathcal{B}_i v \Rightarrow w\mathcal{K}_i v$, un agent sait qu'il croit ;

— $\forall v, u \in \mathcal{W} : w\mathcal{K}_i u \wedge w\mathcal{B}_i v \Rightarrow u\mathcal{K}_i v$, un agent sait qu'il ne croit pas.

La figure 2.10 présente la liste des axiomes dans le système de (Stalnaker, 2006).

$(\mathbf{KD45}_B)$	$\vdash \phi$, pour tout théorème ϕ de KD45 pour B_i
$(\mathbf{T}_K)$	$\vdash K_i\phi \Rightarrow \phi$
$(\mathbf{4}_{BK})$	$\vdash B_i\phi \Rightarrow K_iB_i\phi$
$(\mathbf{5}_{BK})$	$\vdash \neg B_i\phi \Rightarrow K_i\neg B_i\phi$
$(\mathbf{BK})$	$\vdash K_i\phi \Rightarrow B_i\phi$

FIGURE 2.10 – Système combinant à la fois des croyances et des connaissances (Stalnaker, 2006)

Une contrainte de croyance forte peut aussi être ajoutée à ce système signifiant que si un agent croit alors il croit qu'il sait.

$$\forall u, v \in \mathcal{W} : w\mathcal{B}_i u \wedge u\mathcal{K}_i v \Rightarrow w\mathcal{B}_i v$$

Cette contrainte se traduit comme l'axiome :

$$\vdash B_i\phi \Rightarrow B_iK_i\phi$$

Logiques d'actions

Les logiques d'actions expriment la notion d'intention des agents. Si les logiques épistémiques dynamiques et les logiques doxastiques dynamiques représentent les actions sous la forme d'événements et en expriment les conséquences sur les croyances et connaissances des agents, il existe d'autres formalismes pour exprimer cette notion d'intentions (Segerberg et al., 2009). Par exemple, les logiques dynamiques (Harel et al., 2000) et les logiques temporelles (Prior, 1967) comme ALX (Huang et al., 1996), CTL (Emerson and Halpern, 1986) ou encore ATL (Alur et al., 2002) considèrent chaque action qu'un agent peut effectuer sous la forme d'un programme. (Giordano et al., 2000) considère une modalité pour chaque action qu'un agent peut effectuer. Dans de tels formalismes, une action est représentée comme un programme avec sa sortie et exprime les conséquences associées à cette sortie.

Dans le cas de la représentation de la manipulation, puisque la manipulation est une intention, il n'est donc pas nécessaire de considérer l'action effectuée par un agent de manière explicite, mais uniquement ses conséquences. C'est pourquoi, deux formalismes de logiques d'action nous intéressent plus particulièrement : le formalisme STIT (Belnap and Perloff, 1988; Balbiani et al., 2008; Lorini and Sartor, 2016) et le formalisme BIAT (Santos and Carmo, 1996; Pörn, 2012; Troquard, 2014).

Chacun de ces formalismes représente d'un point de vue abstrait la notion d'assurer que quelque chose est réalisé. Ces approches STIT et BIAT considèrent donc les actions comme le fruit de leurs conséquences. C'est donc un niveau d'abstraction bien adapté dans le cas de la représentation de la manipulation. Les approches BIAT considèrent une modalité E_i qui signifie que l'agent *i amène à ce que quelque chose soit vraie* alors que les approches STIT représentent une modalité $[stit]_i$ qui décrit le fait que l'agent *i veille à ce que quelque chose soit vraie.* Bien que ces deux approches soient souvent confondues, la différence entre ces deux formalismes réside dans le niveau d'abstraction des modalités. Les approches STIT expriment le fait qu'un agent s'assure que quelque chose est réalisé en faisant un choix et introduisent une notion de temporalité. Alors que les approches BIAT sont libérées de cette notion de temporalité et considèrent dans sa forme la plus abstraite le fait d'amener quelque chose à vrai.

Logiques BIAT Dans la littérature, il existe de nombreuses logiques BIAT (Jones and Sergot, 1996; Elgesem, 1997; Carmo and Pacheco, 2000; Pauly, 2002; Troquard, 2014). Par exemple, (Troquard, 2014) considère une axiomatique commune à toutes ces logiques et la généralise aux groupes d'agents. De plus, il intègre la notion de capacité d'un agent à agir. Il considère alors le langage $\mathcal{L}_{BIAT}$ tel que pour tout atome propositionnel p du langage, pour tout groupe d'agents $G \subseteq \mathcal{N}$, le langage $\mathcal{L}_{BIAT}$ est généré par la grammaire BNF suivante :

$$\phi ::= p|\phi_1 \wedge \phi_2|\phi_1 \vee \phi_2|\phi_1 \Rightarrow \phi_2|\phi_1 \Leftrightarrow \phi_2|\neg\phi|\top|\bot|E_G\phi|C_G\phi$$

Pour chaque groupe d'agents $G \subseteq \mathcal{N}$, $E_G\phi$ signifie que le groupe d'agents G amène ϕ à vrai et $C_G\phi$ signifie que le groupe G a la capacité d'amener ϕ à vrai. Pour un agent $i \in \mathcal{N}$, nous notons $E_i\phi \stackrel{\triangle}{=} E_{\{i\}}$ et $C_i\phi \stackrel{\triangle}{=} C_{\{i\}}$ pour décrire respectivement l'intention d'un agent, seul, d'amener quelque chose et la capacité de cet agent d'amener ϕ à vrai.

Les modèles pour interpréter ce langage sont fondés sur une sémantique de voisinage telle que présentée à la section 2.1.2. Un modèle pour BIAT est donc un modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{E}_G\}_{G\subseteq\mathcal{N}}, \{\mathcal{C}_G\}_{G\subseteq\mathcal{N}}, V)$ tel que pour tout groupe d'agents $G \subseteq \mathcal{N}$, :

— $\mathcal{W}$ est un ensemble non vide de mondes possibles ;

— $\mathcal{E}_G : \mathcal{W} \to 2^{2^{\mathcal{W}}}$ est une fonction de voisinage associée à l'opérateur E_G ;

— $\mathcal{C}_G : \mathcal{W} \to 2^{2^{\mathcal{W}}}$ est une fonction de voisinage associée à l'opérateur C_G ;

— $V : \mathcal{A}(\mathcal{L}_{BIAT}) \to 2^{\mathcal{W}}$ est une fonction de valuation.

De plus, pour tout monde possible $w \in \mathcal{W}$ et tout groupe d'agents $G \subseteq \mathcal{N}$, $\mathcal{M}$, nous retrouvons une sémantique classique telle que :

1. $\mathcal{M}, w \models \top$
2. $\mathcal{M}, w \not\models \bot$
3. $\mathcal{M}, w \models p$ ssi $w \in V(p)$
4. $\mathcal{M}, w \models \neg\phi$ ssi $\mathcal{M}, w \not\models \phi$
5. $\mathcal{M}, w \models \phi \wedge \psi$ ssi $\mathcal{M}, w \models \phi$ et $\mathcal{M}, w \models \psi$
6. $\mathcal{M}, w \models E_G\phi$ ssi $|\phi| \in \mathcal{E}_G(w)$
7. $\mathcal{M}, w \models C_G\phi$ ssi $|\phi| \in \mathcal{C}_G(w)$

Pour décrire la notion d'intention collective (Troquard, 2014) ajoute des contraintes sur ce modèle notée de (C1) à (C6) :

(C1) $\mathcal{W} \notin \mathcal{C}_G(w)$ signifie qu'aucun groupe d'agents n'a la capacité de rendre vrai les tautologies.

(C2) $\forall X \in \mathcal{E}_G(w) : w \in X$ signifie que si un groupe d'agent veille à faire quelque chose dans un monde, alors ce quelque chose est vrai dans ce monde possible. Une conséquence immédiate de cette contrainte est que $\emptyset \notin \mathcal{E}_G(w)$ et donc qu'aucun groupe ne peut veiller à rendre vrai l'impossible.

(C3) $\emptyset \notin \mathcal{C}_G(w)$ un groupe d'agent est donc incapable de rendre vrai l'impossible.

(C4) $\forall X_1 \in \mathcal{E}_G(w), \forall X_2 \in \mathcal{E}_G(w) : (X_1 \cap X_2) \in \mathcal{E}_G(w)$ signifie que si X_1 est un choix possible pour le groupe G et si X_2 est un autre choix, alors l'intersection de ces deux choix de mondes possibles est un troisième choix possible pour le groupe G.

(C5) $\mathcal{E}_G(w) \subseteq \mathcal{C}_G(w)$ signifie que tout groupe G qui amène quelque chose à vrai en est alors capable. Une conséquence directe de cette contrainte combinée avec la contrainte $(C1)$ est que puisqu'aucun groupe d'agent n'est capable de rendre vrai une tautologie alors nécessairement, par contraposition, aucun groupe d'agent ne peut veiller à rendre vrai une tautologie i.e. $\mathcal{W} \notin \mathcal{E}_G(w)$.

(C6) $\mathcal{C}_\emptyset(w) = \emptyset$ un groupe d'agents vide ne peut rien faire. Une conséquence de (C5) est donc qu'on a aussi $\mathcal{E}_\emptyset(w) = \emptyset$.

Le système axiomatique correct et complet est résumé en figure 2.11.

Si certains principes de la théorie des jeux de coalitions sont représentés par cette logique BIAT, tous les principes ne peuvent pas être considérés comme le montre l'exemple 2.9. En effet,

(**Ax0**)	$\vdash \phi$, pour tout théorème ϕ du calcul propositionnel
(**Ax1**)	$\vdash E_G\phi \wedge E_G\psi \Rightarrow E_G(\phi \wedge \psi)$
(**Ax2**)	$\vdash E_G\phi \Rightarrow E_G\phi$
(**Ax3**)	$\vdash E_G\phi \Rightarrow C_G\phi$
(**Ax4**)	$\vdash \neg C_G\bot$
(**Ax5**)	$\vdash \neg C_G\top$
(**Ax6**)	$\vdash \neg C_\emptyset\phi$
(**ERE**)	Si $\vdash \phi \Leftrightarrow \psi$ alors $\vdash E_G\phi \Leftrightarrow E_G\psi$
(**ERC**)	Si $\vdash \phi \Leftrightarrow \psi$ alors $\vdash C_G\phi \Leftrightarrow C_G\psi$

FIGURE 2.11 – Un système axiomatique pour BIAT (Troquard, 2014)

le *principe de superadditivité*[3], souvent accepté dans les jeux de coalitions (Aumann and Dreze, 1974), ne peut fonctionner dans un tel système logique.

Exemple 2.9 : *Supposons une situation dans laquelle deux groupes d'agents G_1 et G_2 s'opposent sur une destination de vacances à prendre. Si p représente le fait que « l'ensemble des agents du groupe va à New York City » et q désigne « l'ensemble des agents du groupe va à Barcelone ». De manière indépendante, les deux groupes G_1 et G_2 veillent de manière respective à ce que $E_{G_1}p$ et $E_{G_2}q$. Cependant, nous remarquons que $E_{G_1\cup G_2}(p \wedge q)$ est physiquement impossible. De plus, par* (**Ax3**) *nous déduisons que les deux groupes d'agents G_1 et G_2 sont aussi capables d'amener, pour G_1, p à vrai et, pour G_2, q à vrai. Cependant, l'union des groupes $G_1 \cup G_2$ n'est pas capable d'amener $p \wedge q$ à vrai.*

Logiques STIT Le formalisme STIT (Chellas, 1992; Broersen, 2008; Xu, 2010) représente la notion de *veiller à ce que* quelque chose se produise. Ils considèrent plusieurs modalités comme $[stit]_i$ qui représente le fait qu'un agent i veille à ce que quelque chose soit vraie, $[stit]_\mathcal{N}$ qui représente la prise de décision collective, une modalité de nécessité $\Box$, de futur X, et de passé Y. Il existe beaucoup de variantes à ce STIT standard comme celle de (Lorini and Sartor, 2016) qui considère en plus une modalité de prise de décision rationnelle, notée $[rstit]_i$. Le langage $\mathcal{L}_{rstit}$ de (Lorini and Sartor, 2016) est alors généré par la grammaire :

$$\phi ::= p|\neg\phi|\phi \wedge \phi|X\phi|Y\phi|\Box\phi|[stit]_i\phi|[stit]_\mathcal{N}\phi|[rstit]_i\phi$$

Contrairement aux logiques BIAT qui considèrent une sémantique de voisinage pour les actions, les logiques STIT considèrent un modèle de Kripke $\mathcal{M} = (\mathcal{W}, \equiv, \rightarrow, \leftarrow, \{\sim_i\}_{i\in\mathcal{N}}, \sim_\mathcal{N}, \{\mathcal{R}_i\}_{i\in\mathcal{N}}, V)$ où :

3. Le *principe de superadditivité* est la propriété qui, pour deux groupes distincts G_1 et G_2, font des choix X_1 et X_2, l'union des deux groupes correspond à l'intersection des choix réalisés par ces deux groupes, c'est-à-dire :

$$\forall G_1, G_2 \subseteq \mathcal{N}, G_1 \cap G_2 = \emptyset : X_1 \in \mathcal{E}_{G_1}(w), X_2 \in \mathcal{E}_{G_2}(w) \Rightarrow (X_1 \cap X_2) \in \mathcal{E}_{G_1\cup G_2}(w)$$

— $\mathcal{W}$ est un ensemble non vide de mondes possibles

— $\equiv$ est une relation d'équivalence

— $\rightarrow$ une relation sérielle et déterministe

— $\leftarrow$ la relation inverse de $\rightarrow$, i.e. $\leftarrow = \{(v,w)|(w,v) \in \mathcal{W}^2, w \rightarrow v\}$

— $\{\sim_i\}_{i\in\mathcal{N}} \cup \{\sim_\mathcal{N}\}$ un ensemble de relations d'équivalences sur $\mathcal{W}$

— pour tout agent $i \in \mathcal{N}$, $\forall w \in \mathcal{W} : \mathcal{R}_i(w) \subseteq 2^{\{v\in\mathcal{W}:w\sim_i v\}}$

— $V : \mathcal{A}(\mathcal{L}_{rstit}) \rightarrow 2^\mathcal{W}$ est une fonction de valuation.

Pour tout monde possible $w \in \mathcal{W}$, $\mathcal{M}$ est tel que :

1. $\mathcal{M}, w \models X\phi$ ssi $\forall v \in \rightarrow (w) : \mathcal{M}, v \models \phi$
2. $\mathcal{M}, w \models Y\phi$ ssi $\forall v \in \leftarrow (w) : \mathcal{M}, v \models \phi$
3. $\mathcal{M}, w \models \Box\phi$ ssi $\forall v \in \mathcal{W}, w \equiv v : \mathcal{M}, v \models \phi$
4. $\mathcal{M}, w \models [stit]_i\phi$ ssi $\forall v \in \sim_i (w) : \mathcal{M}, v \models \phi$
5. $\mathcal{M}, w \models [stit]_\mathcal{N}\phi$ ssi $\forall v \in \sim_\mathcal{N} (w) : \mathcal{M}, v \models \phi$
6. $\mathcal{M}, w \models [rstit]_i\phi$ ssi si $\sim_i (w) \in \mathcal{R}_i$ alors $\forall v \in \sim_i (w) : \mathcal{M}, v \models \phi$

Chaque classe d'équivalence induit par la relation $\equiv$ est appelée *moment*. Par exemple, $m_w = \{v \in \mathcal{W} : w \equiv v\}$ est la classe d'équivalence du monde possible $w \in \mathcal{W}$. Cette classe d'équivalence représente intuitivement l'ensemble des mondes possibles à un instant donné dans le système. Cette relation $\equiv$ est associée à la sémantique de l'opérateur de nécessité $\Box$. Ensuite, pour représenter la notion de futur, STIT reprend la vision de la temporalité de (Prior, 1967; Zanardo, 1996) pour qui, à chaque monde possible w est associé un unique successeur par la relation $\rightarrow (w)$. Cette relation $\rightarrow$ est considérée comme déterministe (i.e. un unique successeur) et sérielle (i.e. il en existe au moins un). Quant-à la relation inverse $\leftarrow$, qui représente la relation passée, celle-ci est déterministe mais non sérielle car dans leur vision du temps, le passé n'est pas sans fin. Ainsi une liste de moments $(m_1, m_2, \ldots)$ possède toujours un point de départ, cette liste forme ce qu'ils définissent comme *l'historique*. Pour un moment donné, il est donc possible qu'un agent i veille à ce que certaines choses soient réalisées $[stit]_i\phi$. Cet agent i fait alors des choix sur les états des mondes et pour définir sémantiquement le choix d'un agent, un moment est alors partitionné en fonction des choix possibles qu'un agent peut réaliser à cet instant et est représenté par la relation $\sim_i$.

Exemple 2.10 : *La figure 2.12 illustre une situation, issue de (Lorini and Sartor, 2016), dans laquelle un agent* 1 *peut décider d'inciter un autre agent* 2 *(noté par la variable in), à acheter une voiture électrique en lui offrant une remise de* 3000$ *sur celle-ci ou bien, ne pas proposer à l'agent* 2 *cette remise (noté par la variable no). Nous notons ce moment* $m_1 = \{w_1, w_2, w_3, w_4\}$ *où* $\{w_1, w_2\}$ *représente les mondes possibles dans lesquels l'agent* 1 *incite l'agent* 2*, tandis que l'ensemble* $\{w_3, w_4\}$ *représente les mondes dans lesquels l'agent* 1 *n'incite pas l'agent* 2*. Ces deux sous-ensembles partitionnent les choix de l'agent en un ensemble* $\sim_1 (m_1) =$

$\{\{w_1, w_2\}, \{w_3, w_4\}\}$. *De plus, dans cette situation, nous supposons que l'agent* 2 *possède un budget de* 7000$ *pour le choix d'une voiture et que la voiture électrique vaut* 10000$ *(sans remise) et que la voiture diesel est déjà à* 7000$. *Nous pouvons donc considérer qu'en présence de l'offre de la remise, la décision rationnelle pour un tel agent* 2 *est de choisir la voiture électrique (el) plutôt que la voiture diesel (ds).*

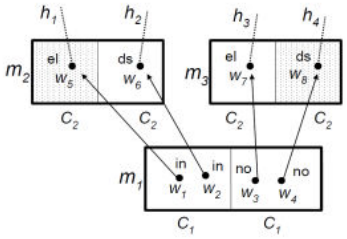

FIGURE 2.12 – Exemple de modèle RSTIT (Lorini and Sartor, 2016)

Pour caractériser cette notion de choix rationnel, (Lorini and Sartor, 2016) ajoutent au formalisme STIT standard, une modalité de décision rationnelle ([rstit]) qui permet de considérer certains choix possibles comme rationnels. Il s'agit de la relation $\mathcal{R}_i$ *qui est associée pour chaque agent* i *du système. Ainsi, le moment* $m_2 = \{w_5, w_6\}$, *représente la situation dans laquelle l'agent* 1 *vient de proposer la remise de* 3000$ *à l'agent* 2. *Dans* m_2, *même si le choix rationnel de l'agent* 2 *est d'acheter la voiture électrique, il reste possible pour lui d'acheter la voiture diesel. Sinon lorsque l'agent* 1 *ne propose pas la remise, la décision rationnelle pour l'agent* 2 *est alors de prendre la voiture diesel dans* m_3. *Par conséquent, l'ensemble des choix rationnels pour l'agent* 2 *sont donc décrits par l'ensemble* $\mathcal{R}_2 = \{\{w_5\}, \{w_8\}\}$.

Pour décrire la notion d'intentions des agents dans le temps, telle que donnée dans l'exemple 2.10, les modèles STIT ont besoin de contraintes. Ces contraintes sont notées de (C1) à (C6).

(C1) $\forall w, v \in \mathcal{W} : w \rightarrow^+ v \Rightarrow w \not\equiv v$ [4]

(C2) $\forall i \in \mathcal{N} : \forall w \in \mathcal{W}, \sim_i (w) \subseteq \equiv (w)$

(C3) $\forall (u_i)_{i \in \mathcal{N}} \in \mathcal{W}^{|\mathcal{N}|} : (\forall i, j \in \mathcal{N}, u_i \equiv u_j) \Longrightarrow \bigcap_{i \in \mathcal{N}} \sim_i (u_i) \neq \emptyset$

(C4) $\forall i \in \mathcal{N}, \forall w \in \mathcal{W}, \exists v \in \mathcal{W} : w \equiv v, \sim_i (v) \in \mathcal{R}_i$

(C5) $\forall w \in \mathcal{W} : \sim_{\mathcal{N}} (w) = \bigcap_{i \in \mathcal{N}} \sim_i (w)$

(C6) $\rightarrow \circ \equiv \subseteq \sim_{\mathcal{N}} \circ \rightarrow$ avec $\circ$ l'opération de composition standard [5].

La contrainte $(C1)$ traduit le fait que les mondes futurs ne peuvent être contenus dans le même moment. La contrainte $(C2)$ traduit que tous les choix possibles à un moment donné sont contenus dans ce moment. La contrainte $(C3)$ traduit qu'il existe au moins un monde possible

4. $\rightarrow^+$ est la cloture transitive de $\rightarrow$

5. Si F et R sont deux relations binaires, alors l'opération de composition, notée $F \circ R$, est définie comme $F \circ R := \{(x, z) | \exists y : (x, y) \in R \wedge (y, z) \in F\}$.

(**CP**)	$\vdash \phi$, pour tout théorème ϕ du calcul propositionnel
(**S5**($\Box$))	$\vdash \phi$, pour tout théorème ϕ de S5 associé à $\Box$
(**KD**(X))	$\vdash \phi$, pour tout théorème ϕ de KD associé à X
(**K**(Y))	$\vdash \phi$, pour tout théorème ϕ de K associé à Y
(**S5**($[stit]_i$))	$\vdash \phi$, pour tout théorème ϕ de S5 associé à $[stit]_i$
(**S5**($[stit]_{\mathcal{N}}$))	$\vdash \phi$, pour tout théorème ϕ de S5 associé à $[stit]_{\mathcal{N}}$
(**Alt**$_X$)	$\vdash \neg X\phi \Rightarrow X\neg\phi$
(**Alt**$_Y$)	$\vdash \neg Y\phi \Rightarrow Y\neg\phi$
(**Conv**$_{X,Y}$)	$\vdash \phi \Rightarrow XY\phi$
(**Conv**$_{Y,X}$)	$\vdash \phi \Rightarrow YX\phi$
(**Rel**$_{\Box_i,[stit]_i}$)	$\vdash \Box\phi \Rightarrow [stit]_i\phi$
(**AIA**)	$\vdash \Diamond[stit]_j\phi_1 \wedge \ldots \wedge \Diamond[stit]_n\phi_n \Rightarrow \Diamond([stit]_1\phi_1 \wedge \ldots \wedge [stit]_n\phi_n)$
(**Rel**$_{[rstit]_i,[stit]_i}$)	$\vdash \langle rstit\rangle_i\top \Rightarrow ([rstit]_i\phi \Leftrightarrow [stit]_i\phi)$
(**RatCh**)	$\vdash \langle rstit\rangle_i\top \Rightarrow ([stit]_i\langle rstit\rangle_i\top)$
(**OneRat**)	$\vdash \Diamond\langle rstit\rangle_i\top$
(**Rel**$_{[stit]_i,[stit]_N}$)	$\vdash \bigwedge_{n\in N}[stit]_n\phi_n \Rightarrow [stit]_N \bigwedge_{n\in N}\phi_n$
(**NCUH**)	$\vdash X\neg B_i\neg\phi \Rightarrow \langle stit\rangle_N X\phi$
(**MP**)	$\dfrac{\vdash \phi \qquad \vdash \phi \Rightarrow \psi}{\vdash \psi}$

FIGURE 2.13 – Axiome du système RSTIT (Lorini and Sartor, 2016)

pour chaque moment qui considère l'ensemble des choix réalisés par tous les agents de $\mathcal{N}$. La contrainte $(C4)$ traduit qu'à chaque moment, tous les agents ont un choix rationnel possible. La contrainte $(C5)$ traduit que l'action réalisée par l'ensemble des agents est définie comme l'intersection de tous les choix réalisés par les agents de $\mathcal{N}$. Enfin, la contrainte $(C6)$ représente sémantiquement que tous les futurs possibles sont contenus dans les choix possibles de futurs pour l'ensemble des agents $\mathcal{N}$[6]. La figure 2.13 résume l'ensemble des axiomes du système. Ce système axiomatique est correct et complet (Lorini and Sartor, 2016).

Logiques déontiques

Les logiques déontiques (Von Wright, 1951) sont des logiques modales qui permettent d'exprimer les notions d'obligation et de permission dans un système normatif. Ces logiques considèrent alors deux opérateurs modaux : $O\phi$ qui représente l'obligation que la proposition ϕ soit à vrai, et $P\phi$ qui représente la permission que ϕ soit à vrai. Ces deux opérateurs sont souvent interprétés comme dual et la relation $P\phi \equiv \neg O\neg\phi$ est vérifiée. Le langage des logiques déontiques est donné par la grammaire suivante :

$$\phi ::= p|\phi_1 \wedge \phi_2|\phi_1 \vee \phi_2|\phi_1 \Rightarrow \phi_2|\phi_1 \Leftrightarrow \phi_2|\neg\phi|\top|\bot|O\phi|P\phi$$

6. En effet, $\rightarrow \circ \equiv := \{(x,z)|\exists y : x \equiv y \wedge y \rightarrow z\}$ traduit la relation de tous les futurs possibles à partir de l'ensemble des mondes possibles pour un moment m donné par $\equiv$ et $\sim_{\mathcal{N}} \circ \rightarrow := \{(x,z)|\exists y : x \rightarrow y \wedge y \sim_{\mathcal{N}} z\}$ traduit la relation vers l'ensemble des choix de futurs possibles pour les agents de $\mathcal{N}$.

Pour interpréter le langage des logiques déontiques, nous considérons un modèle de Kripke $\mathcal{M} = (\mathcal{W}, \mathcal{R}, V)$ tel que $\mathcal{R}$ est une relation sérielle et les opérateurs O et P sont définis tels que pour tout monde $w \in \mathcal{W}$:

$$\mathcal{M}, w \models O\phi \text{ ssi } \forall v \in \mathcal{W}, w\mathcal{R}v : \mathcal{M}, v \models \phi$$

$$\mathcal{M}, w \models P\phi \text{ ssi } \exists v \in \mathcal{W}, w\mathcal{R}v : \mathcal{M}, v \models \phi$$

Le système logique obtenu est un système KD. Il existe des variantes aux logiques déontiques qui intègrent d'autres types d'obligations. (Meyer et al., 1988) considèrent le « ce qui doit être » [7] qui imposent aux agents les états du système dans lequel celui-ci doit être et le « ce qui doit se faire » [8] qui contraint les agents à effectuer certaines actions obligatoirement. Pour traduire sémantiquement le « ce qui doit être » et le « ce qui doit se faire », (Carmo and Pacheco, 2000) intègre un opérateur BIAT et ajoute des contraintes sémantiques à son système de telle sorte que, par exemple, si il est obligatoire que p, alors il est obligatoire que tout agent i amène à ce que p soit à vrai et réciproquement. Les opérateurs STIT peuvent aussi être utilisés, à la place des opérateurs BIAT, pour décrire d'autres logiques déontiques comme dans (Bentzen, 2010).

Cependant, les logiques déontiques ont de nombreux paradoxes (Hansen, 2006) comme le paradoxe de Ross donné dans l'exemple 2.11. Pour une revue détaillée des logiques déontiques, le lecteur intéressé peut se référer à (McNamara, 2006; Hilpinen, 2012)

Exemple 2.11 : *Supposons une situations dans laquelle un agent i doit surveiller une base et considérons la variable propositionnelle pour le fait que « l'agent i informe l'équipe de sécurité sur la situation ». L'agent doit donc nécessairement informer l'équipe de sécurité en cas d'intrusion, i.e. Op. De plus, considérons la variable propositionnelle q pour « l'agent i ne fait rien ». Or $\vdash p \Rightarrow (p \vee q)$ est une tautologie. Donc par nécessitation $\vdash O(p \Rightarrow (p \vee q))$ est aussi un théorème et par application du modus ponens sur l'axiome K, nous déduisons $\vdash Op \Rightarrow O(p \vee q)$ est un théorème. Par conséquent, dans cette situation, nous déduisons alors que $O(p \vee q)$, c'est-à-dire, qu'il est obligatoire que « l'agent i informe l'équipe de sécurité sur la situation » ou « l'agent i ne fait rien ». Il y a donc un paradoxe entre le fait qu'il est nécessaire que l'agent i informe l'équipe de sécurité sur la situation et le fait qu'il est nécessaire que l'agent i ne fasse rien. Ce paradoxe porte un nom, le* paradoxe de Ross.

Logiques des désirs et des émotions

Les logiques modales permettent aussi d'exprimer la notion de désirs et d'émotions. Usuellement, un désir peut être vu comme une relation de préférences (un pré-ordre) qu'un agent essaie de maximiser (Liu, 2010; Bienvenu et al., 2010). Les logiques *ceteris paribus* (Van Benthem et al., 2009; Grossi et al., 2015) sont des exemples de logiques qui représentent les préférences. Du point de vue des systèmes multi-agents, les désirs peuvent aussi être vus comme un but d'un agent à accomplir. Si de nombreuses logiques des préférences et des désirs existent (Lang et al., 2002;

7. Traduit « ought-to-be » en anglais.
8. Traduit « ought-to-do » en anglais.

Lang et al., 2003; Doyle et al., 1991), de manière générale, les désirs et les émotions peuvent être vus comme des relations d'accessibilité et peuvent donc être exprimées avec des logiques modales.

Par exemple, (Adam et al., 2006) modélisent avec une logique modale des états émotionnels comme la joie, la peur, l'espoir, la fierté, ou encore, la honte. Pour définir ces états émotionnels, ils définissent des prédicats combinant des modalités d'actions, de croyances, de désirs, d'idéaux et de temps. Ils considèrent tout d'abord le langage, noté $\mathcal{L}_{OCC}$, décrit par la grammaire sous forme de Backus-Naur, où $Act = \{\alpha, \beta, \gamma, \ldots\}$ représente un ensemble d'actions, $\alpha \in Act$ une action et $i \in \mathcal{N}$ un agent :

$$\phi ::= p|\phi_1 \wedge \phi_2|\neg\phi|\top|\bot|After_{i:\alpha}\phi|Before_{i:\alpha}\phi|Bel_i\phi|Prob_i\phi|G\phi|H\phi|Des_i\phi|Undes_i\phi|Idl_i\phi$$

Toutes ces modalités signifient :

- $After_{i:\alpha}\phi$ exprime que :
  « la conséquence ϕ est vraie après toute exécution de l'action α par l'agent i » ;
- $Before_{i:\alpha}\phi$ exprime que :
  « ϕ doit être vraie avant toute exécution de l'action α par l'agent i » ;
- $Bel_i\phi$ exprime que « l'agent i croit que ϕ est vraie », $Prob_i\phi$ exprime que « l'agent i croit que ϕ est plus probable que $\neg\phi$ » ;
- $G\phi$ exprime que « ϕ sera toujours vraie » ;
- $H\phi$ exprime que « ϕ a toujours été vraie dans le passé » ;
- $Des_i\phi$ exprime que « l'agent i désire ϕ » ;
- $Undes_i\phi$ exprime que « l'agent i ne désire pas ϕ » ;
- $Idl_i\phi$ exprime que « idéalement ϕ est vraie pour i » pour représenter les notions de lois ou d'obligations morales.

Les formules du langage sont interprétées dans le modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{A}_\alpha\}_{\alpha \in Act}, \{\mathcal{B}_i\}_{i \in \mathcal{N}}, \{\mathcal{P}_i\}_{i \in \mathcal{N}}, \mathcal{G}, \{\mathcal{L}_i\}_{i \in \mathcal{N}}, \{\mathcal{D}_i\}_{i \in \mathcal{N}}, \{\mathcal{I}_i\}_{i \in \mathcal{N}}, V)$ où :

1. $\mathcal{W}$ est un ensemble non vide de mondes possibles ;
2. $\{\mathcal{A}_{i:\alpha}\}_{i:\alpha \in \mathcal{N} \times Act}$ est un ensemble de relations binaires associées à agent $i \in \mathcal{N}$ et à chaque action $\alpha \in Act$;
3. $\{\mathcal{B}_i\}_{i \in \mathcal{N}}$ est un ensemble de relations binaires transitives, sérielles et euclidiennes ;
4. $\forall i \in \mathcal{N}, \mathcal{P}_i : \mathcal{W} \to 2^{2^{\mathcal{W}}}$ est une fonction de voisinage tq $\forall w \in \mathcal{W}, \forall U \in \mathcal{P}_i(w), U \subseteq \mathcal{B}_i(w)$;
5. $\mathcal{G}$ est une relation connexe, transitive et antisymétrique tq $\forall \alpha \in Act, \forall i \in \mathcal{N}, \mathcal{A}_{i:\alpha} \subseteq \mathcal{G}$;
6. $\{\mathcal{L}_i\}_{i \in \mathcal{N}}$ est un ensemble de relations binaires ;
7. $\{\mathcal{D}_i\}_{i \in \mathcal{N}}$ est un ensemble de relations binaires ;
8. $\{\mathcal{I}_i\}_{i \in \mathcal{N}}$ est un ensemble de relations binaires sérielles ;
9. $V : \mathcal{A}(\mathcal{L}_{OCC}) \to 2^{\mathcal{W}}$ est une fonction de valuation.

$(\mathbf{KD45}_{B_i})$	$\vdash \phi$, pour tout théorème ϕ de KD45 pour B_i
$(\mathbf{KD}_{Idl_i})$	$\vdash \phi$, pour tout théorème ϕ de KD pour Idl_i
$(\mathbf{K}_{H,G})$	$\vdash \phi$, pour tout théorème ϕ de K pour G et H
$(\mathbf{Conv}_{GP})$	$\vdash \phi \Rightarrow GP\phi$
$(\mathbf{Conv}_{HF})$	$\vdash \phi \Rightarrow HF\phi$
$(\mathbf{BP})$	$\vdash Bel_i\phi \Rightarrow Prob_i\phi$
$(\mathbf{GAfter})$	$\vdash G\phi \Rightarrow After_{i:\alpha}\phi$
$(\mathbf{4}_{Des_i})$	$\vdash Des_i\phi \Rightarrow B_iDes_i\phi$
$(\mathbf{5}_{Des_i})$	$\vdash \neg Des_i\phi \Rightarrow B_i\neg Des_i\phi$
$(\mathbf{4}_{Undes_i})$	$\vdash Undes_i\phi \Rightarrow B_iUndes_i\phi$
$(\mathbf{5}_{Undes_i})$	$\vdash \neg Undes_i\phi \Rightarrow B_i\neg Undes_i\phi$
$(\mathbf{RE}_{Des_i})$	Si $\vdash \phi \leftrightarrow \psi$ alors $\vdash Des_i\phi \leftrightarrow Des_i\psi$
$(\mathbf{D}_{Des_i})$	$\vdash Des_i\phi \Rightarrow \neg Des_i\neg\phi$
$(\neg\top_{Des_i})$	$\vdash \neg Des_i\top$
$(\neg\bot_{Des_i})$	$\vdash \neg Des_i\bot$
$(\mathbf{Pers}_{Des_i})$	$\vdash Des_i\phi \rightarrow GDes_i\phi$
$(\mathbf{Pers}_{\neg Des_i})$	$\vdash \neg Des_i\phi \rightarrow G\neg Des_i\phi$
$(\mathbf{RDU}_{Des_i})$	$\vdash Des_i\phi \rightarrow \neg Undes_i\phi$

FIGURE 2.14 – Axiomatique du système logique fondé sur OCC (Adam et al., 2006)

Un agent ne peut pas « désirer » et « ne pas aimer » à la fois une même proposition. Ainsi, le modèle est contraint de telle sorte que les propriétés suivantes soient vérifiées sur le cadre, pour tout monde $w \in \mathcal{W}$:

$$\mathcal{M}, w \models Des_i\phi \text{ ssi } \forall v \in \mathcal{L}_i(w), \mathcal{M}, v \models \phi \text{ et } \exists u \in \mathcal{D}_i(w), \mathcal{M}, u \not\models \phi$$

$$\mathcal{M}, w \models Undes_i\phi \text{ ssi } \forall v \in \mathcal{D}_i(w), \mathcal{M}, v \models \phi \text{ et } \exists u \in \mathcal{L}_i(w), \mathcal{M}, u \not\models \phi$$

De plus, nous retrouvons l'introspection sur les désirs et sur ce qui est probable :

$$\text{Si } w \in \mathcal{B}_i(w') \text{ alors } \mathcal{B}_i(w) = \mathcal{B}_i(w') \text{ et } \mathcal{D}_i(w) = \mathcal{D}_i(w') \text{ et } \mathcal{L}_i(w) = \mathcal{L}_i(w') \text{ et } \mathcal{P}(w) = \mathcal{P}(w')$$

L'interprétation de ces modalités est l'interprétation standard des logiques modales normales et non normales, à ceci près que la modalité de temporalité associée au passé est la relation inverse de $\mathcal{G}$. Nous avons alors pour tout monde possible $w \in \mathcal{W}$:

$$\mathcal{M}, w \models H\phi \text{ ssi } \forall v \in \mathcal{W} : \text{ si } w \in \mathcal{G}(v) \text{ est définie, alors } \mathcal{M}, v \models \phi.$$

De plus, d'autres opérateurs de temporalité sont introduits dans le langage comme $F\phi \stackrel{\triangle}{=} \neg G\neg\phi$ qui décrit « ϕ est vraie ou sera vraie dans le futur » et $P\phi \stackrel{\triangle}{=} \neg H\neg\phi$ pour « ϕ est vraie ou était vraie dans le passé ».

Ce cadre logique donne une axiomatique du désir décrite par la figure 2.14. En particulier, les axiomes $(\mathbf{Pers}_{Des_i})$, $(\mathbf{Pers}_{\neg Des_i})$ décrivent l'idée que les désirs sont atemporels. Ce modèle

permet d'exprimer un grand nombre d'états émotionnels d'agents comme ceux donnés dans l'exemple 2.1.3. (Adam et al., 2006) identifient et caractérisent différents états émotionnels comme la joie, l'espoir, la satisfaction, la jubilation, etc. Par exemple :

— la joie est caractérisée comme le fait de croire qu'un de nos désir est réalisé, c'est-à-dire $Joy_i\phi \triangleq Bel_i\phi \wedge Des_i\phi$;

— les attentes sont définies comme le fait de penser probable une proposition ϕ mais ne pas y croire, c'est-à-dire $Expect_i\phi \triangleq Prob_i\phi \wedge \neg B_i\phi$;

Cette notion d'attente permet à (Adam et al., 2006) de définir d'autres notions comme l'espoir, la peur, ou encore la honte :

— le prédicat $Hope_i\phi \triangleq Expect_i\phi \wedge Des_i\phi$ signifiant qu'un agent a de l'espoir sur une certaine proposition ϕ si, et seulement si, cet agent s'attend que ϕ soit vraie tout en désirant cette proposition ;

— la peur, quant-à-elle, se définit comme le prédicat dans lequel un agent s'attend à une proposition ϕ et ne désire pas cette proposition, i.e. $Fear_i\phi \triangleq Expect_i\phi \wedge Undes_i\phi$;

— la honte est $Shame_{i:\alpha}\phi \triangleq Bel_i Done_{i:\alpha}(Idl_i \neg Happens_{i:\alpha}\phi \wedge Prob_i After_{i:\alpha}\neg\phi) \wedge Bel_i\phi$ où $Happens_{i:\alpha}\phi \triangleq \neg After_{i:\alpha}\neg\phi$ et $Done_{i:\alpha}\phi \triangleq \neg Before_{i:\alpha}\neg\phi$.

Ainsi, nous pouvons imaginer qu'un agent manipulateur puisse utiliser de tels états émotionnels pour construire des stratégies de manipulation. En effet, un agent manipulateur pourrait chercher à créer une attente, jouer sur les désirs, s'appuyer sur la peur, ou encore, instaurer un sentiment de honte. Par exemple, si les intentions du manipulateur sont dissimulées, la coercition peut faire partie d'une stratégie de manipulation fondée sur l'état émotionnel de la peur instaurée dans l'esprit de la victime.

Si nous avons présenté de nombreux états mentaux pouvant être exprimés par les logiques modales, nous n'avons pas présenté un état mental au cœur de toute interaction entre agents et pouvant avoir un rôle important dans la manipulation : *la confiance*. Or, les logiques modales permettent aussi d'exprimer la confiance avec une modalité ou un prédicat. Ainsi, dans la suite de cette thèse, nous prenons le choix de modéliser la notion de confiance, et par conséquent, nous présentons en détails l'ensemble des approches logiques liées à la modélisation de la confiance.

2.2 Des logiques qui caractérisent la confiance

Dans les systèmes multi-agents, la représentation de la confiance a été très étudiée comme nous l'avons présenté en section 1.1.3 avec les systèmes de confiance et les systèmes de réputation. Elle permet d'intégrer au système un mécanisme de régulations afin d'exclure toute interaction avec des agents reconnus comme non fiables, malhonnêtes ou manipulateurs. Dans cette section, nous présentons les approches logiques qui ont modélisé la notion de confiance. Dans un premier temps, en section 2.2.1, nous présentons la confiance comme une déclinaison en différents aspects. Dans un second temps, en section 2.2.2, nous présentons des mécanismes logiques pour inférer de la confiance. Enfin, en section 2.2.3, nous présentons la confiance définie comme un degré de confiance.

2.2.1 Différents aspects de la confiance

La confiance prend un sens différent en fonction du contexte (Castelfranchi and Falcone, 2010). Par exemple, lorsqu'Alice dit à Bob « j'ai confiance en toi », il est possible qu'Alice ait confiance en Bob pour une tâche qu'il doit réaliser, pour la fiabilité d'une information qui lui a donnée, ou encore pour sa sincérité. Nous distinguons dans cette section, les travaux qui considèrent la confiance sur les compétences d'un agent à bien réaliser une certaine tâche, la confiance sur la disposition d'un agent à agir, la confiance en les institutions et la confiance dans les communications.

Confiance sur les compétences d'un agent

La confiance sur les compétences d'un agent est exprimée comme une croyance de l'agent qui fait confiance envers l'agent qui reçoit la confiance, pour une action qui lui a été déléguée. (Herzig et al., 2010) considèrent la confiance comme un prédicat signifiant qu'un agent i fait confiance à un autre agent j par rapport à une action α aboutissant à une proposition ϕ, si et seulement si, toutes les propriétés suivantes sont vraies :

1. i a le but que ϕ soit réalisé,
2. i croit que :
   (a) j est capable de réaliser l'action α,
   (b) j, après avoir réalisé α assurera ϕ,
   (c) j a l'intention de faire α.

Ceci permet de définir un prédicat de *confiance immédiate*[9]. Cette notion traduit un aspect de la confiance dans le présent, à savoir qu'un agent j s'apprête bien à réaliser l'action pour laquelle i lui fait confiance. Une seconde notion de confiance, la *confiance dispositionnelle*[10], est la confiance accordée par un agent i à un agent j que cet agent j réalisera le but ϕ de i dans un contexte spécifique.

(Smith et al., 2011) considèrent aussi une confiance immédiate signifiant qu'un agent i fait confiance à un autre agent j pour ϕ si, et seulement si, toutes les propriétés suivantes sont vraies :

1. i a le but ϕ,
2. i croit que j réalise ϕ,
3. i a l'intention que :
   (a) j réalise ϕ,
   (b) il n'est pas le cas que i fasse ϕ.
4. i a le but que j ait l'intention de faire ϕ,
5. i croit que j a l'intention de ϕ.

9. Occurrent trust.
10. Dispositional trust.

Enfin, (Drawel et al., 2017) abordent le problème de la modélisation de la confiance sur les compétences d'un agent à bien réaliser une tâche en utilisant la logique temporelle CTL. Ils considèrent alors le langage $\mathcal{L}_{TCTL}$:

$$\phi ::= p|\neg\phi|\phi \wedge \phi|AG\phi|EX\phi|EG\phi|E(\phi \cup \phi)|T_{i,j}\phi$$

La modalité $AG\phi$ signifie que dans tous les chemins, la formule ϕ est vérifiée, $EX\phi$ est une modalité qui signifie qu'il existe un chemin dans lequel ϕ est vrai dans l'état suivant, $EG\phi$ signifie qu'il existe un chemin dans lequel ϕ est toujours vraie et $E(\phi \cup \psi)$ signifie qu'il existe un chemin dans lequel il existe un futur moment pour ψ et ϕ est vérifiée jusqu'à ce futur moment.

Un modèle $\mathcal{M} = (\mathcal{W}, \mathcal{R}, I, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$ est un modèle où $\mathcal{W}$ est un ensemble non vide de mondes possibles appelés *états du système*, $\mathcal{R} : \mathcal{W} \to \mathcal{W}$ une fonction de transition du système, $I \subseteq \mathcal{W}$ un ensemble d'états initiaux globaux pour le système, $\mathcal{T}_{i,j}$ une relation d'accessibilité définie comme une relation d'équivalence pour tout couple d'agents $(i,j) \in \mathcal{N}^2$ et $V : \mathcal{W} \to 2^{\mathcal{A}(\mathcal{L}_{TCTL})}$ une fonction de valuation. De plus, (Drawel et al., 2017) définissent la relation d'accessibilité pour la confiance à partir d'une matrice de valeurs entières $(a_{i,j})_{i,j\in\mathcal{N}}$ pour chaque état $s \in \mathcal{W}$ du système. Ainsi, pour tout agent $i, j \in \mathcal{N}$ et deux états $s, s' \in \mathcal{W}$, $s\mathcal{T}_{i,j}s'$ si, et seulement si, $a_{i,j}(s) = a_{i,j}(s')$. Les propriétés de relation d'équivalence sont alors immédiates. Ce système logique aboutit alors sur un ensemble de théorèmes comme :

- $\models \phi \Rightarrow \neg T_{i,j}\phi$, i.e. la confiance a été achevée ;
- $\models T_{i,j}(\phi \wedge \psi) \wedge \neg\phi \Rightarrow T_{i,j}\phi$, i.e. l'agent i fait confiance à l'agent j pour deux tâches, si l'une n'est pas déjà réalisée alors l'agent i fait confiance à j pour la tache non encore réalisée ;
- $\models T_{i,j}\phi \Rightarrow \neg T_{i,j}\neg\phi$, i.e. la confiance évite les conflits, l'axiome D est vérifié, sous une autre forme la formule $\models \neg T_{i,j}\bot$ est équivalente ;
- $\models T_{i,j}\phi \wedge T_{i,j}\psi \Rightarrow T_{i,j}(\phi \wedge \psi)$, i.e. la confiance combine les tâches ;
- $\models AG\neg\phi \Rightarrow \neg T_{i,j}\phi$, i.e. si une tâche est irréalisable, un agent ne peut pas faire confiance en un autre agent pour cette tâche à réaliser.

Confiance sur la disposition d'un agent à agir

Si (Herzig et al., 2010) présentent une notion de confiance dispositionnelle comme un prédicat, d'autres approches comme (Singh, 2011) ou (Liau, 2003; Dastani et al., 2004; Dundua and Uridia, 2010) considèrent plutôt des modalités pour exprimer la disposition d'un agent à agir ou à croire. (Singh, 2011), exprime une notion de confiance dispositionnelle par l'intermédiaire d'une modalité d'arité 2, $T_{i,j}(\phi, \psi)$ signifiant que, dans un contexte ϕ, un agent i fait confiance à un autre agent j pour réaliser ψ et considère une modalité d'engagement, notée $C_{i,j}(\phi, \psi)$, d'un agent i envers un autre agent j pour réaliser la proposition ψ lorsqu'un contexte ϕ est vérifié. Il voit la confiance comme étant équivalente à un engagement d'un agent envers un autre agent. Par exemple, si ϕ est vraie, la confiance de l'agent i envers j est active, l'agent j s'engage alors envers i pour mener ψ à vrai. De plus, (Singh, 2011) exprime aussi une forme de confiance immédiate, exprimée par la formule $T_{i,j}(\top, \psi)$ signifiant qu'à tout instant (et donc dans l'instant présent) i fait confiance

à j pour réaliser ψ. D'autres travaux ont intégré une modalité de confiance. Par exemple, les travaux de (Dundua and Uridia, 2010) introduisent une modalité de confiance $T_{i,j}$ représentant la confiance comme une forme de croyance qu'un agent entretient sur un autre agent. Pour le langage généré par la grammaire :

$$\phi ::= p|\neg\phi|\phi_1 \wedge \phi_2|\Box\phi|T_{i,j}\phi$$

Ils considèrent un modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{R}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$ tel que :

— $\mathcal{W}$ est un ensemble non vide de mondes possibles ;

— $\{\mathcal{R}_i\}_{i\in\mathcal{N}}$ un ensemble de relations binaires symétriques et faiblement transitives, i.e. :

$$\forall i, j \in \mathcal{N}, \forall w, v, u \in \mathcal{W}, w\mathcal{R}_i v \wedge v\mathcal{R}_i u \wedge w \neq u \Rightarrow w\mathcal{R}_i u$$

— Pour tout $i, j \in \mathcal{N}$, $\mathcal{T}_{i,j} : \mathcal{W} \to 2^{2^{\mathcal{W}}}$ est une fonction de voisinage ;

— V une fonction de valuation.

De plus, pour tout monde $w \in \mathcal{W}$ et pour tout agents $i, j \in \mathcal{N}$, (Dundua and Uridia, 2010) considèrent que $\mathcal{T}_{i,j}$ est telle que :

$$\mathcal{T}_{i,j}(w) = \bigcap_{u\in\mathcal{R}_i(w)} \mathcal{T}_{i,j}(u)$$

Ces contraintes sémantiques permettent alors de construire le système axiomatique correct et complet donné par la figure 2.15. En particulier, ce système logique permet de déduire le théorème $\vdash T_{i,j}\phi \Leftrightarrow B_i T_{i,j}\phi$, signifiant qu'un agent i a confiance en un autre agent j si, et seulement si l'agent i croit qu'il a confiance en l'agent j.

$(\mathbf{CP})$	$\vdash \phi$, pour toute tautologie du calcul propositionnel
$(\mathbf{K}_{\Box_i})$	$\vdash \phi$, pour tout théorème du système $\mathbf{K}$ pour $\Box_i$
$(\mathbf{B}_{\Box_i})$	$\vdash \phi \Rightarrow \Box_i\Diamond_i\phi$
$(\mathbf{4S}_{\Box_i})$	$\vdash \Box_i\phi \wedge \phi \Rightarrow \Box_i\Box_i\phi$
$(\mathbf{C4}_{T\Box_i})$	$\vdash T_{i,j}\phi \Rightarrow \Box_i T_{i,j}\phi$
$(\mathbf{C5}_{T\Box_i})$	$\vdash \neg T_{i,j}\phi \Rightarrow \Box_i\neg T_{i,j}\phi$

FIGURE 2.15 – Système axiomatique de la confiance de (Dundua and Uridia, 2010)

Confiance sur les communications

Un autre aspect important de la confiance est celui de la *confiance dans le discours d'un agent.* En effet, lorsque deux agents communiquent ensemble, il peut arriver que la communication ne soit pas fiable entre ces deux agents ou bien qu'un agent ne soit pas sincère ou honnête dans les informations qu'il donne. Des travaux comme (Primiero and Taddeo, 2012; Christianson and Harbison, 1997; Demolombe, 2004) ont proposé de représenter la confiance qu'entretient

un agent dans la communication qu'il a avec les autres agents du système. Ils définissent ainsi des propriétés qui doivent être nécessairement satisfaites dans les communications pour pouvoir considérer qu'un agent a confiance dans la communication. Pour (Demolombe, 2004), dans une communication, un agent peut faire confiance à un autre agent pour différentes propriétés. La confiance en un discours peut être associée à la sincérité de l'autre, à sa fiabilité, à sa crédibilité, ou encore la coopération de l'autre agent à informer l'agent qui fait confiance. Demolombe définit un un cadre logique dans lequel il considère différentes modalités comme la connaissance d'un agent i (K_i), la croyance d'un agent i (B_i), l'information transmise d'un agent i vers un agent j ($I_{i,j}$), une modalité BIAT pour représenter l'intention d'un agent d'amener à ce que quelque chose soit vrai (E_i) ainsi que des modalités d'obligations (O) et de permission (P). Il définit alors différentes propriétés en lien avec la communication comme :

Sincérité : un agent j est *sincère* au regard d'un agent i si, et seulement si, lorsque l'agent j informe i d'une proposition ϕ alors l'agent j croit cette proposition,

$$sin_{i,j}(\phi) \stackrel{\triangle}{=} I_{j,i}\phi \Rightarrow B_j\phi$$

Coopération : un agent j est *coopératif* au regard d'un agent i si, et seulement si, lorsque l'agent j croit une proposition ϕ, celui-ci en informe l'agent i

$$coo_{i,j}(\phi) \stackrel{\triangle}{=} B_j\phi \Rightarrow I_{j,i}\phi$$

Honnêteté : un agent i est *honnête* si, et seulement si, lorsque l'agent i veille à ce que ϕ soit vraie, alors il est permis que l'agent i veille à ce que ϕ soit vraie,

$$hon_i(\phi) \stackrel{\triangle}{=} E_j\phi \Rightarrow PE_i\phi$$

De manière générale, pour (Demolombe, 2004), la confiance est nécessairement associée à une des propriétés mentionnée précédemment, notée *prop*. Ainsi, la confiance peut alors être définie comme :

$$Tprop_{i,j}(\phi) \stackrel{\triangle}{=} K_i prop(\phi)$$

Par exemple, la confiance en l'honnêteté est $Thon_{i,j}(\phi) \stackrel{\triangle}{=} K_i(E_j\phi \Rightarrow PE_j\phi)$, elle se caractérise comme le fait qu'un agent sache que lorsqu'un agent fait une action, cette action est nécessairement permise, ou encore la confiance en la sincérité se caractérisant par le prédicat $Tsinc_{i,j}(\phi) \stackrel{\triangle}{=} K_i(Com_{j,i}\phi \Rightarrow B_j\phi)$. Enfin, Demolombe ajoute une notion de confiance conditionnelle décrite par le fait que dans une certaine condition, si les prémisses sont vérifiées alors l'agent pour qui on a confiance, respectera la propriété *prop* :

$$Tprop_{i,j}(\phi|\psi) \stackrel{\triangle}{=} K_i(\psi \Rightarrow prop(\phi))$$

Les définitions de Demolombe considèrent que lorsqu'un agent a confiance en un autre agent pour quelque chose (par exemple, pour sa sincérité), il sait alors que cette propriété est vérifiée. Or pour Castelfranchi et Falcone, puisque la confiance représente un risque, elle ne peut être une

certitude. Ainsi, la confiance serait davantage plus proche de la croyance que de la connaissance. Dans des articles plus récents Demolombe (Demolombe, 2009) remplace dans ses propriétés la modalité de connaissance (K_i) par celle de la croyance (B_i).

Dans la littérature qui s'intéresse à représenter la confiance avec des logiques modales, nous avons donc vu qu'il existait principalement deux types d'approches pour représenter la confiance : les logiques modales qui définissent la confiance comme une combinaison de modalités comme (Demolombe, 2004; Herzig et al., 2010; Smith et al., 2011) qui combinent des modalités pour l'intention, la communication, la croyance, ou encore, la connaissance ; et les logiques modales qui considèrent explicitement une modalité de confiance comme (Dundua and Uridia, 2010; Singh, 2011; Drawel et al., 2017). Cependant, d'autres approches logiques existent et représentent comment des agents peuvent inférer de la confiance et quelles sont les implications logiques associées au fait de faire confiance.

2.2.2 Inférer avec la confiance

Les agents peuvent utiliser la confiance qu'ils accordent aux autres agents pour inférer de nouveaux états mentaux comme, par exemple, de nouvelles croyances. Cependant, qu'en est-il pour savoir si une confiance est bien fondée ? Les travaux de (Liau, 2003; Dastani et al., 2004) se sont intéressés, d'une part, à l'influence de la confiance dans l'assimilation de nouvelles informations au travers d'un cadre logique appelé le formalisme BIT. D'autre part, ce formalisme est utilisé pour exprimer la notion de thème sur lequel un agent peut accorder sa confiance en un autre agent. De plus, ce cadre est aussi utilisé pour permettre à un agent de vérifier, par le moyen d'un questionnement, s'il peut faire confiance ou non à un autre agent.

Déduire des croyances par la confiance

(Liau, 2003) définit le formalisme BIT pour exprimer le lien existant entre la confiance en la fiabilité d'une source, l'acquisition d'information provenant de cette source, et l'inférence de nouvelles croyances. Liau représente alors l'aspect de la confiance qui est la disposition d'un agent à croire des faits. Il définit un langage composé d'une modalité d'acquisition d'information, notée $I_{i,j}$, signifiant qu'un agent i a acquis une information de j. Cette modalité est associée à une sémantique normale, et plus particulièrement à un système KD. De plus, cette modalité possède l'introspection positive et l'introspection négative par rapport aux croyances B_i ; cette modalité B_i représente les croyances d'un agent i et est associée aussi à une sémantique normale décrivant un système KD45. En revanche, la confiance qu'un agent i émet pour une source j, notée $T_{i,j}$, est une modalité non normale et est associée à une sémantique de voisinage telle que présentée à la section 2.1.2. Un modèle standard pour le formalisme BIT est donc un N-uplet $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{I}_{i,j}\}_{i,j\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$ tel que :

- $\mathcal{W}$ est un ensemble non vide de mondes possibles ;
- $\{\mathcal{B}_i\}_{i\in\mathcal{N}}$ un ensemble de relations binaires sérielles, transitives et euclidiennes ;
- $\{\mathcal{I}_{i,j}\}_{i,j\in\mathcal{N}}$ est un ensemble de relations binaires sérielles ;
- Pour tout agent $i, j \in \mathcal{N}$, $\mathcal{T}_{i,j} : \mathcal{W} \to 2^{2^{\mathcal{W}}}$ est une fonction de voisinage.

Ce cadre permet alors de construire le système axiomatique standard pour BIT. Ce système axiomatique est résumé sur la figure 2.16.

(**CP**)	$\vdash \phi$, pour tout théorème ϕ du calcul propositionnel
(**KD45**(B_i))	$\vdash \phi$, pour tout théorème ϕ de KD45 associé à B_i
(**KD**$(I_{i,j})$)	$\vdash \phi$, pour tout théorème ϕ de KD associé à $I_{i,j}$
(**MP**)	Si $\vdash \phi \Rightarrow \psi$ et $\vdash \phi$ alors $\vdash \psi$
(**Nec**)	Si $\vdash \phi$ alors $\vdash B_i\phi$ et $\vdash I_{i,j}\phi$
(**Dual**)	$\vdash T_{i,j}\phi \Leftrightarrow \neg T_{i,j}\neg\phi$, $\vdash B_i\phi \Leftrightarrow \neg B_i\neg\phi$, et $\vdash I_{i,j}\phi \Leftrightarrow \neg I_{i,j}\neg\phi$
(**RE**)	Si $\vdash \phi \Leftrightarrow \psi$ alors $\vdash T_{i,j}\phi \Leftrightarrow T_{i,j}\psi$, $\vdash B_i\phi \Leftrightarrow B_i\psi$ et $\vdash I_{i,j}\phi \Leftrightarrow I_{i,j}\psi$

FIGURE 2.16 – Un système axiomatique standard pour BIT (Liau, 2003)

Liau considère un ensemble de nouvelles contraintes pouvant être ajoutées à son cadre pour permettre, par exemple, de déduire qu'un agent infère de nouvelles croyances ou permettre une introspection de la confiance par rapport aux croyances de l'agent, c'est-à-dire pour tout monde $w \in \mathcal{W}$ et pour tout agent $i, j \in \mathcal{N}$:

— (**m1**) $\forall S \in \mathcal{T}_{i,j} : \mathcal{B}_i \circ \mathcal{I}_{i,j}(w) \subseteq S \Rightarrow \mathcal{B}_i(w) \subseteq S$ signifiant que si un agent i a confiance en un agent i et croit que cet agent j lui a transmis une certaine information ϕ alors l'agent i croit cette information. Cette contrainte est alors traduite par le théorème suivant dans le système :

$$\vdash B_i I_{i,j}\phi \wedge T_{i,j}\phi \Rightarrow B_i\phi$$

— (**m2**) $\mathcal{T}_{i,j}(w) = \bigcap_{u \in \mathcal{B}_i(w)} \mathcal{T}_{i,j}(w)$ traduit que un agent i a confiance en un agent j si, et seulement si i croit qu'il a confiance en j, c'est-à-dire :

$$\vdash T_{i,j}\phi \Leftrightarrow B_i T_{i,j}\phi$$

(Liau, 2003) nomme alors ce système, le système **BA**. Il propose ensuite la possibilité d'étendre BA en permettant par exemple de considérer le fait que si un agent i fait confiance en un autre agent j pour ϕ alors cet agent i fait aussi confiance à j pour le contraire $\neg\phi$. Cette contrainte s'appelle la symétrie et elle permet de décrire que si un agent peut potentiellement croire un agent pour ϕ alors il peut aussi croire pour son contraire. La figure 2.3 représente l'ensemble des systèmes étendant le formalisme BIT et sont complets et corrects (Liau, 2003).

Remarquons toutefois que (Liau, 2003) ne considère pas dans le système BIT standard les contraintes relationnelles entre $\{\mathcal{B}_i\}_{i\in\mathcal{N}}$ et $\{\mathcal{I}_{i,j}\}_{i,j\in\mathcal{N}}$ pour décrire l'introspection positive et négative vis-à-vis des informations transmises et des croyances [11].

11. Les contraintes d'introspection positive et négative sont données par :

— $\forall w, v, u \in \mathcal{W} : w\mathcal{B}_i u \wedge u\mathcal{I}_{i,j} v \Rightarrow w\mathcal{I}_{i,j} v$ pour décrire le fait qu'un agent lorsqu'il reçoit une information transmise par un agent j, croit qu'il a reçu l'information transmise de j. Cette contrainte est traduite dans le système par le théorème suivant $\vdash I_{i,j}\phi \Rightarrow B_i I_{i,j}\phi$;

— $\forall v, u \in \mathcal{W} : w\mathcal{B}_i u \wedge w\mathcal{I}_{i,j} v \Rightarrow u\mathcal{I}_{i,j} v$ et signifiant que s'il n'est pas le cas qu'un agent i a reçu une information de j, alors l'agent i croit qu'il n'est pas le cas que l'agent i a reçu une information de j. Cette contrainte

Nom	Contraintes sémantiques	Axiomes syntaxiques correspondants
$*$	Contraintes standards de BIT	$\vdash \phi$, pour tout théorème ϕ de BIT
(**BA**)	(m1) $\forall S \in \mathcal{T}_{i,j} : \mathcal{B}_i \circ \mathcal{I}_{i,j}(w) \subseteq S \Rightarrow \mathcal{B}_i(w) \subseteq S$	$\vdash B_i I_{i,j}\phi \wedge T_{i,j}\phi \Rightarrow B_i\phi$
	(m2) $\mathcal{T}_{i,j}(w) = \bigcap_{u \in \mathcal{B}_i(w)} \mathcal{T}_{i,j}(w)$	$\vdash T_{i,j}\phi \Leftrightarrow B_i T_{i,j}\phi$
(**SY**)	(m3) $\forall S \subseteq \mathcal{W} : S \in \mathcal{T}_{i,j}(w) \Rightarrow S \in \mathcal{T}_{i,j}(w)$	$\vdash T_{i,j}\phi \Rightarrow T_{i,j}\neg\phi$
(**TR**)	(m4) $\bigcap_{u \in \mathcal{B}_i(w)} \mathcal{T}_{j,k}(u) \subseteq \mathcal{T}_{i,k}(w)$	$\vdash B_i T_{j,k}\phi \Rightarrow T_{i,k}\phi$
(**CA**)	$\mathcal{T}^c_{i,j} \subseteq \mathcal{W} \times 2^{\mathcal{W}} : \mathcal{T}^c_{i,j}(w,S)$ ssi 1 et 2 1)$\forall u \in \mathcal{B}_i(w), \mathcal{I}_{i,j}(u) \subseteq S \Rightarrow \mathcal{B}_j(u) \subseteq S$ 2)$\forall u \in \mathcal{B}_i(w) : \mathcal{B}_j(u) \subseteq S \Rightarrow u \in S$	$\vdash T^c_{i,j}\phi \Leftrightarrow B_i((I_{i,j} \Rightarrow B_j\phi) \wedge (B_j\phi \Rightarrow \phi))$
	(m5) $\mathcal{T}_{i,j}(w) \subseteq \mathcal{T}^c_{i,j}(w)$	$\vdash T_{i,j}\phi \Rightarrow T^c_{i,j}\phi$
(**IC**)	(m6) $\mathcal{B}_i \circ \mathcal{I}_{i,j} \equiv \mathcal{I}_{i,j}$	$\vdash I_{i,j}\phi \Leftrightarrow B_i I_{i,j}\phi$

Tableau 2.3 – Ensemble des systèmes axiomatiques fondés sur BIT (Liau, 2003)

En effet, il ne considère pas ces contraintes sémantiques car, selon lui, il est possible que dans certaines situations, comme les attaques de l'homme du milieu [12], une information semble provenir d'une certaine source j mais en réalité provient d'une autre source que j. Ainsi, un agent vigilant peut ne pas croire que cette information a été réellement transmise par cette source j. Cependant, Liau considère un autre système, nommé le système **IC** permettant d'exprimer explicitement cette relation d'équivalence $I_{i,j}\phi \equiv B_i I_{i,j}\phi$. Liau intègre, dans son système nommé **CA** la notion de *confiance prudente*, notée $T^c_{i,j}$, et définit par le prédicat :

$$T^c_{i,j}\phi \stackrel{\triangle}{=} B_i((I_{i,j} \Rightarrow B_j\phi) \wedge (B_j\phi \Rightarrow \phi))$$

Cette notion de confiance prudente, se traduit comme le fait que l'agent i croit que lorsqu'il reçoit une information de j, l'agent j croit cette information et lorsque l'agent j croit une information, cette information est nécessairement juste. Un agent est prudent si, et seulement si, cet agent, lorsqu'il accorde sa confiance, le fait nécessairement avec une confiance prudente, c'est-à-dire $\mathcal{T}_{i,j}(w) \subseteq \mathcal{T}^c_{i,j}(w)$ est vérifiée. Enfin, Liau propose d'étendre le formalisme BIT pour représenter le fait qu'un agent peut faire confiance sur un ensemble de propositions, appelé *thème* [13].

Confiance d'un agent par rapport à des thèmes

Un thème est un ensemble de propositions ayant un point en commun. Par exemple, les propositions « changer la courroie de distribution », « installer un nouveau système de freinage », « changer la carrosserie » appartiennent au thème de la réparation automobile. Il existe des logiques modales exprimant cette notion de thème comme celle de (Demolombe and Jones, 1999) qui introduisent une modalité $A(t,'\phi')$ signifiant que la proposition ϕ est associée au thème t.

est traduite dans le système par le théorème suivant $\vdash \neg I_{i,j}\phi \Rightarrow B_i\neg I_{i,j}\phi$.

12. *Man In The Middle* en anglais.
13. *Topics* en anglais.

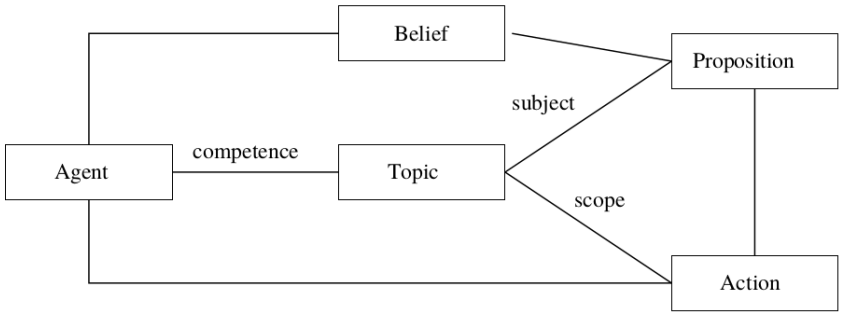

FIGURE 2.17 – Modèle de la confiance (Herzig and Longin, 2000)

Cette notion de thème permet à (Herzig and Longin, 2000) de considérer un mécanisme de révision des croyances en fonction de l'acquisition de nouvelles informations en lien avec un thème donné. Ce mécanisme doit nécessairement contenir les composantes décrites par la figure 2.17, c'est-à-dire :

— Une fonction de compétence qui définit les liens entre les sujets pour lesquels un agent est compétent ;

— Une fonction sujet qui définit le lien entre les propositions et leurs sujets de thème ;

— Une fonction qui fait le lien entre les thèmes et les actions affectées par ce thème. Par exemple, le fait d'informer d'autres agents, le fait de rédiger un manuscrit, etc.

(Dastani et al., 2004) reprend ces travaux et les intègre au formalisme BIT de Liau. Cette notion de thème leur permet alors de considérer la confiance qu'un agent accorde sur un thème donné. Ainsi, un agent peut, par exemple, faire confiance à un autre agent sur le thème de la réparation automobile mais pas sur le thème de l'informatique. Si un agent est compétent sur un thème et qu'une proposition appartient à ce thème, alors l'agent informé de cette proposition va croire cette nouvelle information.

Vérifier la confiance par le questionnement

(Dastani et al., 2004) propose en plus de vérifier la confiance par un mécanisme de questionnements. Il considère le formalisme BIT présenté précédemment et dans lequel il intègre une modalité non normale $Q_{i,j}\phi$ signifiant qu'un agent i questionne un agent j sur la vérité d'une proposition ϕ. Il contraint son cadre de telle sorte à pouvoir caractériser le fait que si un agent i qui questionne un autre agent j sur une proposition ϕ que i croit vraie et que i croit que j lui a transmis l'information ϕ, alors l'agent i a confiance en j sur ϕ, c'est-à-dire il rend valide dans son système, le théorème $\vdash Q_{i,j}\phi \wedge B_i\phi \wedge B_i I_{i,j}\phi \Rightarrow T_{i,j}\phi$. De plus, il considère aussi que si un agent i questionne un autre agent j sur une proposition ϕ que i croit fausse et que i croit que j lui a transmis l'information ϕ, alors l'agent i n'a pas confiance en j sur ϕ, c'est-à-dire $\vdash Q_{i,j}\phi \wedge B_i\neg\phi \wedge B_i I_{i,j}\phi \Rightarrow \neg T_{i,j}\phi$. Dastani *et al.* combinent alors le cadre logique exprimant la

confiance et les thèmes mentionnés précédemment et intègre à celui-ci le questionnement pour permettre aux agents du système de vérifier s'ils peuvent faire confiance ou non à un agent en fonction des réponses données par l'agent.

Nous avons donc présenté jusqu'à présent des approches logiques qui caractérisaient la confiance et ses liens existant avec d'autres états mentaux comme les croyances. Cependant, nous avons présenté des travaux sur la confiance qui ne considéraient la confiance uniquement comme binaire. En effet, dans ces travaux, nous avons supposé qu'un agent ne pouvait uniquement accorder ou ne pas accorder sa confiance. Ainsi, dans la suite de cet état de l'art, nous présentons les approches logiques qui modélisent la confiance, non plus comme une notion binaire, mais comme un degré de confiance qu'un agent accorde envers un autre agent.

2.2.3 Confiance graduée et multi-valuée

(Falcone et al., 2002) proposent de considérer la confiance comme une logique floue où la confiance est une valeur comprise entre 0 et 1. La valeur 0 signifie qu'un agent n'est pas de confiance et 1 pour un agent de confiance. Pour évaluer cette confiance, un agent évalue la confiance qu'il a envers un autre agent en fonction de différents critères subjectifs comme :

1. sa capacité à bien agir ou donner une réponse fiable ;
2. sa volonté d'agir ou d'être fiable dans ses réponses ;
3. sa malhonnêteté ;
4. l'opportunité représentée par le fait d'accorder sa confiance ;
5. le danger représenté par le fait d'accorder sa confiance.

Une fois l'évaluation de chacun des critères faite, l'agent agrège ces évaluations dans une fonction et évalue quel degré de confiance il accorde à l'autre agent.

Plutôt que de représenter la confiance à partir d'une logique floue comme (Falcone et al., 2002), d'autres travaux représentent une notion de confiance graduée dans une logique modale (Demolombe, 2004; Demolombe, 2009). La confiance est alors interprétée comme une forme de croyance graduée en différents niveaux ordonnés de 1 à n. Demolombe définit un modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{B}_i^g\}_{i\in\mathcal{N}}^{g\in[|1,n|]}, \{R^g\}^{g\in[|1,n|]}, V)$ où chaque élément de l'ensemble $\{\mathcal{B}_i\}_{i\in\mathcal{N}}$ est une relation binaire associée à une modalité de croyance B_i pour chaque agent $i \in \mathcal{N}$. À chaque niveau de croyance $g \in [|1,n|]$ et pour chaque agent i est associé une modalité B_i^g qui représente le niveau de croyance de l'agent i au niveau g. Cette modalité est associée à une fonction de voisinage $\mathcal{B}_i^{g\in[|1,n|]} : \mathcal{W} \to 2^{2^{\mathcal{W}}}$. Contrairement à la modalité B_i dans laquelle l'agent ne statut pas le niveau de croyance, la modalité B_i^g statut la force de la croyance de l'agent i. Ainsi, plus le niveau considéré est élevé et plus la croyance est forte pour l'agent. Dans ce modèle, il définit un opérateur d'implication $\phi \Rightarrow^h \psi$ sémantiquement par rapport à un contexte $X \subseteq \mathcal{W}$ de sous-ensembles de mondes possibles :

$$\forall w \in X : \mathcal{M}, X, w \models \phi \Rightarrow^h \psi \text{ ssi } |\psi|_X \in R^h(w, |\phi|_X), \text{ où } |\phi|_X = \{v \in X : \mathcal{M}, X, v \models \phi\}$$

Cet opérateur représente le fait que dans le contexte X au niveau h, si ϕ est vraie dans ce contexte, nécessairement ψ est aussi vraie. Cette notation de contexte de mondes possibles revient à écrire de manière équivalente, lorsque $X = \mathcal{W}, \mathcal{M}, \mathcal{W}, w \models \phi$ si, et seulement si, $\mathcal{M}, w \models \phi$. Le contexte permet alors de préciser dans quel sous-ensemble de mondes possibles nous nous plaçons :

$$\forall w \in X : \mathcal{M}, X, w \models B_i\phi \text{ ssi } \forall v \in \mathcal{B}_i(w) : \mathcal{M}, \mathcal{B}_i(w), v \models \phi$$

De la même façon, la modalité de croyance graduée est définie sémantiquement par la relation :

$$\forall w \in X : \mathcal{M}, X, w \models B_i^g\phi \text{ ssi } |\phi|_{\mathcal{B}_i(w)} \in \mathcal{B}_i^g(w)$$

Ainsi pour Demolombe, la confiance au niveau g se définit alors comme le degré de croyance de niveau g que lorsque la formule ϕ_j est vraie, nécessairement ψ_j est aussi vraie, pour un certain degré d'implication h, c'est-à-dire :

$$Trust_{i,j}^g(\phi_j, \psi_j) \triangleq B_i^g(\phi_j \Rightarrow^h \psi_j)$$

Sur la base du modèle précédent, (Lorini and Demolombe, 2008) ont caractérisé une notion de confiance graduée en la délégation d'une tache α par un agent i à un autre agent j. Cette notion est définie comme le fait qu'un agent i a pour but de rendre vrai une proposition ϕ et croit que, pour un certain degré g, que l'agent j en est capable et en a bien l'intention de réaliser la tache α qui amènera à ϕ. Ainsi, cette confiance est alors exprimée par :

$$Trust_{i,j}^g(\alpha, \phi) \triangleq AGoal_i\phi \wedge B_i^g(After_{j:\alpha}\phi \wedge Does_{j:\alpha}\top)$$

Ici, $AGoal_i\phi$ est un prédicat pour décrire que l'agent i a bien le but de rendre vrai ϕ, $After_{j:\alpha}\phi$ une modalité qui décrit que l'agent j après avoir exécuté l'action α amène ϕ à vrai et $Does_{j:\alpha}\top$ la modalité qui décrit l'intention de j de réaliser l'action α.

En conclusion, nous avons vu que différents aspects de la confiance pouvaient être exprimés avec des logiques modales comme la confiance en les compétences d'un agent, la confiance en la disposition d'un agent à agir ou encore la confiance dans la communication. Les logiques modales sont aussi utilisées pour exprimer des mécanismes d'inférences impliqués dans le fait d'accorder sa confiance. De manière générale, si nous venons de présenter différents états mentaux pouvant être impliqués dans une manipulation comme la confiance, les croyances, ou les intentions, nous n'avons pas présenté de travaux portant sur la représentation de la manipulation ou de notions s'en rapprochant comme le mensonge, la malhonnêteté, l'influence, ou la prise de conscience. Dans la suite, nous présentons des travaux en lien avec ces notions.

2.3 Des logiques en lien avec la manipulation

Peu de travaux se sont intéressés à exprimer la manipulation telle que nous la définissons à la section 1.2.2. Quelques travaux se sont intéressés à des notions proches comme le mensonge, la tromperie, la malhonnêteté, l'influence, ou la prise de conscience mais pas directement à la manipulation. Dans un premier temps, nous présentons les travaux portant sur la modélisation du mensonge, puis présentons une théorie de la malhonnêteté. Dans un second temps, puisque la manipulation est une intention délibérée d'instrumentaliser, nous portons un regard sur les logiques qui se sont intéressées à représenter l'influence et l'intention délibérée. Enfin, nous présentons les travaux portant sur la modélisation de la prise de conscience.

2.3.1 Représenter le mensonge et la malhonnêteté

Mentir est l'intention d'un agent i d'informer un agent j sur quelque chose afin que j le croit alors que l'agent i croit en son contraire (Van Ditmarsch et al., 2012; Sakama et al., 2015). (Van Ditmarsch et al., 2012) utilise des logiques doxastiques dynamiques avec une modalité pour décrire l'action des annonces privées qui est utilisée pour décrire le mensonge. (Sakama et al., 2015) utilisent une logique modale et introduisent une modalité de communication entre deux agents, une modalité de croyance ainsi qu'une modalité d'intention pour décrire des notions comme le mensonge, le baratinage, la dissimulation ou encore la tromperie.

Une logique du mensonge de Hans Van Ditmarsch

(Van Ditmarsch et al., 2012) représente le mensonge comme le fait d'annoncer publiquement quelque chose de faux. Il utilise les modèles de l'annonce publique adapté dans un modèle de logique doxastique dynamique tel que présenté à la section 2.1.3. La figure 2.18 représente son modèle d'action où p est, par exemple, l'annonce publique : « l'accusé était sur la scène du crime ». Dans ce modèle d'action, Hans van Ditmarsch fait l'hypothèse que ses agents sont crédules et vont donc nécessairement croire n'importe quelle annonce : qu'elle soit vraie ou fausse. Ainsi, lorsque l'annonce p est faite, tous les agents du système croit cette annonce.

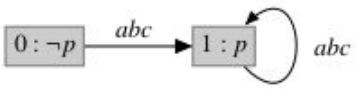

FIGURE 2.18 – Modèle événementiel de la logique du mensonge (Van Ditmarsch et al., 2012)

Il considère un langage $\mathcal{L}_{ll}$, pour tout agent $i \in \mathcal{N}$:

$$\phi ::= p|\phi_1 \wedge \phi_2|\neg\phi|\top|\bot|B_i\phi|[\ddagger\phi_1]\phi_2|[!\phi_1]\phi_2|[¡\phi_1]\phi_2$$

L'opérateur $[!\phi]\psi$ représente le fait que ψ est vraie après l'annonce publique ϕ et $[¡\phi]\psi$ signifie que ψ est vraie après l'annonce du mensonge ϕ et $[\ddagger\phi]\psi$ représente le fait que ψ est vraie après l'annonce publique de ϕ mais aussi vraie après le mensonge ϕ. Si $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, V)$ est

un modèle doxastique et $\mathcal{A}_p = (\mathcal{W}^\alpha, \{\mathcal{B}_i^\alpha\}_{i\in\mathcal{N}}, I^\alpha, P^\alpha)$ son modèle d'action tel que décrit par la figure 2.18 où p peut-être substitué à n'importe quelle formule ϕ. Nous avons $\mathcal{W}^\alpha = \{0, 1\}$ où l'événement 0 représente un mensonge possible alors que l'événement possible 1 représente l'annonce publique standard. Ainsi, la sémantique des opérateurs du langage $\mathcal{L}_{ll}$ est donnée par :

1. $\mathcal{M}, w \models B_i\phi$ ssi $\forall v \in \mathcal{W} : w\mathcal{B}_i v, \mathcal{M}, v \models \phi$
2. $\mathcal{M}, w \models [!\phi]\psi$ ssi $(\mathcal{M}, \{w\}) \otimes (\mathcal{A}_\phi, \{1\}), (w, 1) \models \psi$
3. $\mathcal{M}, w \models [\phi]\psi$ ssi $(\mathcal{M}, \{w\}) \otimes (\mathcal{A}_\phi, \{0\}), (w, 0) \models \psi$
4. $\mathcal{M}, w \models [\ddagger\phi]\psi$ ssi $(\mathcal{M}, \{w\})\otimes(\mathcal{A}_\phi, \{0, 1\}), (w, 0) \models \psi$ et $(\mathcal{M}, \{w\})\otimes(\mathcal{A}_\phi, \{0, 1\}), (w, 1) \models \psi$

Cette sémantique donne alors le système axiomatique résumé sur la figure 2.19, l'annonce manipulatrice est décrite alors par $[\ddagger\phi]\psi$ et par **Ax0**, $[\ddagger\phi]\psi \Leftrightarrow [!\phi]\psi \wedge [¡\phi]\psi$. La règle **A1** décrit que si une annonce publique ϕ amène la conséquence p, alors si ϕ est vraie, nécessairement p aussi et réciproquement. La formule $[!\phi]p \Leftrightarrow \phi \Rightarrow p$ est alors une tautologie du système.

(**CP**)	$\vdash \phi$, pour tout théorème ϕ du calcul propositionnel
(**Ax0**)	$\vdash [\ddagger\phi]\psi \Leftrightarrow [!\phi]\psi \wedge [¡\phi]\psi$
(**A1**)	$\vdash [!\phi]p \Leftrightarrow \phi \Rightarrow p$
(**A2**)	$\vdash [!\phi]\neg\psi \Leftrightarrow \phi \Rightarrow \neg[!\phi]\psi$
(**A3**)	$\vdash [!\phi](\psi_1 \wedge \psi_2) \Leftrightarrow [!\phi]\psi_1 \wedge [!\phi]\psi_2$
(**A4**)	$\vdash [!\phi]B_i\psi \Leftrightarrow \phi \Rightarrow B_i[!\phi]\psi$
(**L1**)	$\vdash [¡\phi]p \Leftrightarrow \neg\phi \Rightarrow p$
(**L2**)	$\vdash [¡\phi]\neg\psi \Leftrightarrow \neg\phi \Rightarrow [¡\phi]\psi$
(**L3**)	$\vdash [¡\phi](\psi_1 \wedge \psi_2) \Leftrightarrow [¡\phi]\psi_1 \wedge [¡\phi]\psi_2$
(**L4**)	$\vdash [¡\phi]B_i\psi \Leftrightarrow \neg\phi \Rightarrow B_i[¡\phi]\psi$

FIGURE 2.19 – Système axiomatique simplifié de (Van Ditmarsch et al., 2012)

Si (Van Ditmarsch et al., 2012) utilise les logiques dynamiques doxastiques pour décrire le mensonge, une autre approche se défait de cet aspect dynamique et consiste à considérer le mensonge comme une combinaison de modalités normales telles que la croyance, les intentions et la communication entre agents.

La théorie de la malhonnêteté de Sakama *et al.*

(Sakama et al., 2015) définissent une théorie logique de la malhonnêteté dans laquelle ils caractérisent des notions de mensonge, de baratinage, de dissimulation de l'information, de tromperie, ou encore de demi-vérité. Pour décrire formellement ces notions, ils considèrent le formalisme BIC, dans lequel le langage considère, pour tout agent $i, j \in \mathcal{N}$, des modalités de croyances B_i, d'intentions I_i et de communications, notée $C_{i,j}$, d'un agent i vers un agent j. Un modèle est un N-uplet $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{I}_i\}_{i\in\mathcal{N}}, \{\mathcal{C}_{i,j}\}_{i,j\in\mathcal{N}}, V)$ avec $\mathcal{W}$ un ensemble non-vide de mondes possibles, $\{\mathcal{B}_i\}_{i\in\mathcal{N}}$ un ensemble de relations sérielles, transitives et euclidiennes, $\{\mathcal{I}_i\}_{i\in\mathcal{N}}$ un ensemble de relations sérielles et $\{\mathcal{C}_{i,j}\}_{i,j\in\mathcal{N}}$, un ensemble de relation sérielle pour décrire

respectivement la sémantiques des opérateurs B_i, I_i et $C_{i,j}$. De plus, (Sakama et al., 2015) considèrent l'introspection entre les modalités d'intentions, de communications et de croyances. Ainsi, ces relations sont telles que :

— $\forall w, v, u \in \mathcal{W} : w\mathcal{B}_i u \wedge u\mathcal{C}_{i,j} v \Rightarrow w\mathcal{C}_{i,j} v$

— $\forall w, v, u \in \mathcal{W} : w\mathcal{B}_i u \wedge w\mathcal{C}_{i,j} v \Rightarrow u\mathcal{C}_{i,j} v$

— $\forall w, v, u \in \mathcal{W} : w\mathcal{B}_i u \wedge u\mathcal{I}_i v \Rightarrow w\mathcal{I}_i v$

— $\forall w, v, u \in \mathcal{W} : w\mathcal{B}_i u \wedge w\mathcal{I}_i v \Rightarrow u\mathcal{I}_i v$

— $\forall w, v, u \in \mathcal{W} : w\mathcal{I}_i u \wedge u\mathcal{C}_{i,j} v \Rightarrow w\mathcal{I}_i v$

— $\forall w, v, u \in \mathcal{W} : w\mathcal{I}_i u \wedge w\mathcal{C}_{i,j} v \Rightarrow u\mathcal{C}_{i,j} v$

(Mahon, 2008) définit tout d'abord le mensonge comme le fait de faire croire une fausse déclaration à une autre personne avec l'intention que cette déclaration soit crue comme vraie pour l'autre personne. (Sakama et al., 2015) reprennent cette définition et définissent la notion de mensonge simple par le prédicat :

$$SimpleLie_{i,j}(\phi) \stackrel{\triangle}{=} C_{i,j}\phi \wedge B_i\neg\phi \wedge I_i B_j \phi$$

Cependant, (Mahon, 2008) considère aussi que mentir à une autre personne peut être défini comme l'intention de faire croire à l'interlocuteur que le locuteur croit ce qu'il affirme, ou bien croire que cette déclaration est crue comme vraie par le locuteur. Ainsi, (Sakama et al., 2015) décrivent aussi le mensonge comme :

$$Lie^*_{i,j}(\phi) \stackrel{\triangle}{=} SimpleLie_{i,j}(\phi) \vee (C_{i,j}\phi \wedge B_i\neg\phi \wedge I_i B_j B_i \phi)$$

À cette notion de mensonge, ils ajoutent la notion de mensonge par objectif lorsque l'agent menteur i a l'intention de faire croire une proposition ϕ par déduction à son interlocuteur. Ici, l'agent menteur ment alors sur une déclaration qu'il ne croit pas σ, mais croit que cette déclaration va amener son interlocuteur à croire ϕ. Cette notion est alors décrite par le prédicat :

$$O - Lie^*_{i,j}(\sigma, \phi) \stackrel{\triangle}{=} I_i B_j \phi \wedge \neg B_i B_j \phi \wedge B_i B_j(\sigma \Rightarrow \phi) \wedge B_i\neg\sigma \wedge C_{i,j}\phi$$

D'autres notions sont caractérisées comme le baratinage qui décrit l'action de communiquer une information ϕ dont il n'est pas le cas que l'agent communiquant i croit ϕ ou croit en son contraire $\neg\phi$, c'est-à-dire le prédicat :

$$BS_{i,j}(\phi) \stackrel{\triangle}{=} C_{i,j}\phi \wedge \neg B_i\phi \wedge \neg B_i\neg\phi$$

Sakama *et al.* caractérisent la dissimulation d'une information comme le fait qu'un agent i ait l'intention de ne pas révéler une information ϕ dont i croit comme vraie, c'est-à-dire :

$$WI_{i,j}(\phi) \stackrel{\triangle}{=} \neg C_{i,j}\phi \wedge B_i\phi \wedge I_i\neg B_j\phi$$

Enfin, la notion de demi-vérité consiste à donner une information que nous croyons comme

vraie afin d'induire l'autre en erreur en jouant sur une erreur de raisonnement chez l'interlocuteur. Elle est formellement décrite par :

$$HT_{i,j}(\phi,\psi) \triangleq (C_{i,j}\phi \wedge B_i\phi \wedge I_iB_j\phi) \wedge \neg B_iB_j\psi \wedge B_iB_j(\phi \Rightarrow \psi) \wedge B_i\neg\psi \wedge \neg C_{i,j}\neg\psi \wedge I_iB_j\psi$$

La partie de ce prédicat $(C_{i,j}\phi \wedge B_i\phi \wedge I_iB_j\phi)$ est appelée la *sincérité intentionnelle*. Ainsi, dans leur système, ils montrent, par exemple, que mentir sur une proposition ϕ tout en ayant l'intention d'être sincère est impossible et déduisent alors le théorème $\models (C_{i,j}\phi \wedge B_i\phi \wedge I_iB_j\phi) \wedge Lie^B_{i,j}(\phi) \Rightarrow \bot$.

De plus, Sakama *et al.* caractérisent des maximes pour des agents malintentionnés. En se fondant sur les maximes introduites par Paul Grice pour la conversation (Grice, 1991), ils considèrent qu'un agent malhonnête et rationnel doit suivre les règles suivantes. Un agent malhonnête, pour atteindre ses objectifs doit respecter le fait que :

1. Plus petit le mensonge est, mieux c'est ;
2. Plus petit le baratinage est, mieux c'est ;
3. Plus petit la dissimulation est, mieux c'est.

À ces trois maximes, ils ajoutent des règles de priorité sur l'usage du mensonge, par rapport au baratinage et à la dissimulation :

1. Ne mens jamais si tu peux parvenir à tes objectifs par le baratinage ;
2. Ne mens jamais ou ne baratine jamais si tu peux parvenir à tes objectifs en dissimulant des informations ;
3. Ne mens jamais, ou ne baratine jamais, ou encore, ne dissimule aucune information si tu peux parvenir à tes objectifs par une demi-vérité.

Ces travaux caractérisent donc le mensonge, la tromperie et plus généralement la malhonnêteté. Or, la manipulation est une intention délibérée d'instrumentaliser tout en dissimulant cette intention. Ainsi, dans le section suivante, nous présentons les travaux portant sur l'influence et ceux sur la représentation de l'intention délibérée.

2.3.2 Représenter l'influence et l'intention délibérée

L'influence est usuellement définie comme l'intention d'un agent d'amener un autre agent à s'assurer que quelque chose soit vrai. Toutefois, l'influence peut aussi porter sur l'intention de changer d'autres états mentaux comme les croyances, les désirs, ou la confiance de l'agent influencé. Cependant, à notre connaissance, seul le premier aspect est considéré dans la littérature.(Lorini and Sartor, 2016) utilisent la logique RSTIT, présentée à la section 2.1.3, pour définir l'influence comme le fait qu'un agent i s'assure que l'agent j réalise dans le futur l'action attendue par l'agent i, c'est-à-dire :

$$[infl]_{i,j}\phi \triangleq [stit]_iX[rstit]_j\phi$$

(Bottazzi and Troquard, 2015) utilisent une logique BIAT pour définir l'influence comme l'intention d'un agent i d'amener un autre agent j à faire quelque chose, c'est-à-dire :

$$[infl]_{i,j}\phi \triangleq E_i(E_j\phi \wedge \psi)$$

Ici, les E_i (resp. E_j) sont les modalités non normales du formalisme BIAT pour désigner l'intention d'*amener quelque chose à vrai* par l'agent i (resp. l'agent j). La formule ψ, représente n'importe quelle formule du langage qui ne contredit pas $E_j\phi$. Ce ψ est explicitement intégré à la définition du prédicat de l'influence, en raison de l'axiomatique du système qui ne permet pas d'avoir la propriété des logiques normales $\Box(\phi \wedge \psi) \equiv \Box\phi \wedge \psi$ (cf. exemple 2.4). En effet, si l'influence était définie seulement par $E_iE_j\phi$ et si un agent i amène un autre agent j à faire le travail (ϕ), tout en veillant à ce que le travail soit bien fait (ψ), nous ne pourrions alors pas déduire de l'influence en raison de cette propriété que nous n'avons pas car seule la formule $E_i(E_j\phi \wedge \psi)$ est vérifiée.

Si la manipulation est une intention d'influencer, elle est aussi une intention délibérée. (Lorini and Sartor, 2016) définissent une notion d'intention délibérée et considèrent que quelque chose est fait délibérément par un agent i si, et seulement si, i veille à ce que quelque chose soit vraie alors que ce n'est pas nécessairement le cas, c'est-à-dire :

$$[dstit]_i\phi \triangleq [stit]_i\phi \wedge \neg\Box\phi$$

Cependant, cette définition n'est pas sans problème. Imaginons une situation dans laquelle un agent i a causé un accident de voiture délibérément pour profiter de l'assurance automobile. Juste après cet accident de voiture, une personne inconsciente était allongée sur la route, morte. En suivant la définition formelle du STIT délibéré de (Lorini and Sartor, 2016), nous déduirions que l'agent i aurait d'une part, délibérément veillé à ce que « la voiture s'écrase » mais aussi, d'autres part, toutes conséquences indirectes comme « une personne est morte ». Or, nous affirmons que le contraire est également possible. Même si l'agent i a délibérément causé cet accident de voiture, il n'a pas eu nécessairement l'intention délibérée de tuer la victime. Le STIT standard ne peut donc pas exprimer cette situation. Par conséquent, l'opérateur STIT délibéré ne correspond pas à ce dont nous avons besoin dans le cadre de cette thèse. De plus, une intention délibérée doit être connue de l'agent, ce qui n'est pas le cas pour l'opérateur $[dstit]$. En effet, quand nous délibérons pour faire ou ne faire quelque chose, alors nous savons que nous avons eu l'intention délibérée de le faire ou non. Puisque Lorini et Sartor utilisent un cadre logique STIT, ils ne considèrent pas d'introspection positive et négative des intentions sur les connaissances des agents. Cependant, à notre connaissance, il n'existe aucun travaux dans les approches BIAT définissant explicitement une modalité d'intention délibérée.

La manipulation est une intention délibérée d'instrumentaliser un agent victime, mais la manipulation est aussi une intention de dissimuler ses intentions. Cette notion même de dissimulation peut être exprimée du point de vue des connaissances d'un agent mais aussi de son niveau de conscience. Dans la suite, nous nous intéressons aux approches logiques qui ont représenté cette notion de « prise de conscience ».

2.3.3 Représenter la prise de conscience

(Modica and Rustichini, 1994; Hill, 2010; Schipper, 2014; Van Ditmarsch et al., 2018) se sont intéressés à représenter formellement la sémantique du fait de « prendre conscience ». Les premiers à définir formellement cette notion de « prise de conscience » en logique sont (Fagin and Halpern, 1987) qui considèrent la logique de Levesque sur les connaissances implicites et explicites (Levesque, 1984) pour modéliser une logique de la non-omniscience. (Modica and Rustichini, 1994) considèrent qu'un agent $i \in \mathcal{N}$ a conscience de quelque chose, noté $A_i\phi$, si, et seulement si, cet agent sait ϕ ou il ne sait pas ϕ et il sait qu'il ne sait pas ϕ, c'est-à-dire, vérifie le prédicat $A_i\phi := K_i\phi \vee (\neg K_i\phi \wedge K_i\neg K_i\phi)$ où K_i est une modalité de connaissance. Si K_i est associée à un système modal S5, alors la définition précédente de la conscience devient équivalente à $A_i\phi := K_i\phi \vee K_i\neg K_i\phi$.

Par la suite, (Schipper, 2014) proposent alors un langage modal $\mathcal{L}$ dans lequel nous retrouvons une modalité L_i pour désigner la connaissance implicite d'un agent i, ainsi qu'une modalité $A_i\phi$ pour exprimer la « prise de conscience » d'un agent i sur une proposition ϕ. La sémantique de cette modalité A_i est définie à partir d'une fonction $\mathcal{A}_i : \mathcal{W} \to 2^{\mathcal{L}}$ dite de *correspondance avec la conscience*[14]. Ainsi, un agent i a conscience d'une proposition ϕ si, et seulement si, cette formule ϕ appartient à un ensemble de formules donné par cette fonction de correspondance. $\mathcal{A}_i : \mathcal{W} \to 2^{\mathcal{L}}$ associe pour un monde donné, l'ensemble des formules dont un agent i a conscient. La connaissance est alors définit par le prédicat $K_i\phi \stackrel{\triangle}{=} L_i\phi \wedge A_i\phi$.

Dans ce formalisme, un modèle est un N-uplet, $\mathcal{M} = (\mathcal{W}, \{\mathcal{R}_i\}_{i\in\mathcal{N}}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, V)$ tel que $(\mathcal{W}, \{\mathcal{R}_i\}_{i\in\mathcal{N}}, V)$ est un cadre de Kripke où $\{\mathcal{R}_i\}_{i\in\mathcal{N}}$ est un ensemble de relations d'équivalences associées à la sémantique de l'opérateur L_i pour chaque agent $i \in \mathcal{N}$. Pour définir la fonction de correspondance $\mathcal{A}_i$, il est alors nécessaire de recourir à une fonction qui renvoie l'ensemble des propositions atomiques d'une formule, c'est-à-dire, une fonction At telle que :

- $At(\top) = \emptyset$
- $At(p) = \{p\}$ si p est un atome du langage $\mathcal{L}$
- $At(\neg\phi) = At(\phi)$
- $At(\phi \wedge \psi) = At(\phi) \cup At(\psi)$
- $At(K_i\phi) = At(\phi)$
- $At(A_i\phi) = At(\phi)$

Ainsi, pour tout agent $i \in \mathcal{N}$ et pour tout monde possible $w \in \mathcal{W}$, la fonction de correspondance est telle que :

1. $\phi \in \mathcal{A}_i(w)$ si, et seulement si, $At(\phi) \subseteq \mathcal{A}_i(w)$
2. $\forall v \in \mathcal{W} : w\mathcal{R}_i v \Rightarrow \mathcal{A}_i(w) = \mathcal{A}_i(v)$

La propriété 1 signifie simplement que, pour qu'un agent i ait conscience d'une formule ϕ, il est nécessaire que cet agent soit conscient de tous les atomes propositionnels contenus dans une formule. La propriété 2 correspond au fait qu'un agent sait ce dont il a conscience. Enfin,

14. *Awareness correspondence* en anglais.

un modèle est défini de façon standard. Nous avons, pour tout agent $i \in \mathcal{N}$ et pour tout monde $w \in \mathcal{W}$, la sémantique suivante :

— $\mathcal{M}, w \models L_i\phi$ ssi $\forall v \in \mathcal{W}, w\mathcal{R}_i v : \mathcal{M}, v \models \phi$

— $\mathcal{M}, w \models A_i\phi$ ssi $\phi \in \mathcal{A}_i(w)$

— $\mathcal{M}, w \models K_i\phi$ ssi $\mathcal{M}, w \models L_i\phi$ et $\mathcal{M}, w \models A_i\phi$

Cette sémantique nous donne alors le système axiomatique de la figure 2.20. Un théorème immédiat de la conscience est que si un agent i sait que quelque chose est vrai alors nécessairement cet agent i a conscience de ce quelque chose. En effet :

— $\vdash K_i\phi \Leftrightarrow L_i\phi \wedge A_i\phi$

— $\vdash L_i\phi \wedge A_i\phi \Rightarrow A_i\phi$ (tautologie R_8 du calcul propositionnel)

— $\vdash K_i\phi \Rightarrow A_i\phi$

Ce système axiomatique présenté dans la figure 2.20 peut bien sûr être défini sans la modalité de connaissance implicite tout en préservant les propriétés de correction et de complétude du système (Schipper, 2014).

(**CP**)	$\vdash \phi$, pour tout théorème ϕ du calcul propositionnel
(**KL**)	$\vdash K_i\phi \Leftrightarrow L_i\phi \wedge A_i\phi$
(**S5**(L_i))	$\vdash \phi$, pour tout théorème ϕ de S5 associé à L_i
(**AS**)	$\vdash A_i\phi \Leftrightarrow A_i\neg\phi$
(**AC**)	$\vdash A_i(\phi \wedge \psi) \Leftrightarrow A_i\phi \wedge A_i\psi$
(**AKR**)	$\vdash A_i\phi \Leftrightarrow A_iK_i\phi$
(**AR**)	$\vdash A_i\phi \Leftrightarrow A_iA_i\phi$
(**ALR**)	$\vdash A_i\phi \Leftrightarrow A_iL_i\phi$
(**UL**)	$\vdash \neg A_i\phi \Leftrightarrow L_i\neg A_i\phi$
(**MP**)	Si $\vdash \phi \Rightarrow \psi$ et $\vdash \phi$ alors $\vdash \psi$
(**Nec**)	Si $\vdash \phi$ alors $\vdash L_i\phi$

FIGURE 2.20 – Système axiomatique de la logique de la prise de conscience de (Schipper, 2014)

Enfin, tous ces travaux ont été étendus dans (Van Ditmarsch et al., 2018) où ils définissent à partir de la logique de la connaissance implicite et explicite de (Fagin and Halpern, 1987) et la logique de la prise de conscience de (Schipper, 2014) une logique de la connaissance spéculative permettant de raisonner sur la notion de non conscience. Cette notion de connaissance spéculative est distincte de celle de la connaissance implicite et de la connaissance explicite. Elle traduit qu'un agent a une connaissance spéculative d'une proposition ϕ dans un modèle $\mathcal{M}$ et un monde w si cette formule est vérifiée dans tous les mondes accessibles dans tous les modèles pointés du cadre qui sont $\mathcal{A}_i(w)$-bissimilaires par rapport au monde w.

2.4 Positionnement du manuscrit

En conclusion, nous avons donc vu que dans la littérature, il existait de nombreux travaux qui utilisaient les logiques modales pour représenter les états mentaux des agents comme les connaissances, les croyances, la confiance, les intentions ou encore la prise de conscience. D'autres travaux ont représenté les notions d'influence, de malhonnêteté, de mensonge ou encore de tromperie. Cependant, nous constatons qu'il n'existe aucune approche qui ne s'est intéressée à la représentation de la manipulation telle que nous l'avons définie dans ce manuscrit au chapitre 1.

Dans la suite de cette thèse, nous considérons des systèmes multi-agents dans lesquels les agents peuvent être humains ou artificiels. Ces agents sont cognitifs et possèdent donc des états mentaux comme des croyances, des connaissances, des désirs, des intentions et sont capables d'accorder leur confiance à d'autres agents du système. De plus, ces agents sont supposés rationnels et fiables dans leurs raisonnements. Ainsi, un agent n'accordera jamais sa confiance à un autre agent si celui-ci ne possède pas les caractéristiques nécessaires pour être considéré comme un agent de confiance. Enfin, certains agent peuvent être malhonnêtes, c'est-à-dire, qu'ils peuvent ne pas respecter délibérément les normes du système. De tels agents peuvent aussi être manipulateurs, c'est-à-dire qu'ils peuvent veiller de façon délibérée à instrumentaliser d'autres agents sans que ces agents ne sachent qu'ils ont été instrumentalisés par cet agent.

Dans le chapitre suivant, nous proposons un système, nommé le *système KBE*, pour exprimer la notion de manipulation telle qu'elle a été définie au chapitre 1 et en s'appuyant sur les travaux présentés dans ce chapitre.

Deuxième partie

KBE et TB - Deux systèmes logiques pour caractériser la manipulation et la confiance en la sincérité

Chapitre 3

Le système KBE - raisonner sur la manipulation

Dans le chapitre 1, nous avons défini la manipulation comme l'intention délibérée d'instrumentaliser une victime tout en veillant à lui dissimuler cette intention. Au chapitre suivant, nous avons présenté différentes logiques modales pour exprimer des états mentaux comme les connaissances, les croyances mais aussi les intentions des agents. Nous constatons que pour exprimer une telle définition, il est nécessaire de considérer une logique modale combinant plusieurs modalités : une modalité pour représenter les effets des actions qu'ils soient délibérés ou non, les effets délibérés des actions et des modalités de croyances et de connaissances pour exprimer la dissimulation. Ainsi, nous définissons dans ce chapitre un nouveau système logique nommé KBE, qui intègre ces modalités en section 3.1. La particularité de ce système est de définir une nouvelle modalité d'intention délibérée pour exprimer la manipulation. Nous prouvons ensuite en section 3.2 que ce système est correct, complet et nous montrons des théorèmes logiques pouvant être déduits dans ce système. En section 3.3, nous donnons une définition logique de la manipulation et étudions des propriétés logiques vérifiées dans ce système comme le principe *qui facit per alium facit per se*, c'est-à-dire, « celui qui agit à travers un autre fait acte lui-même ». Enfin, en section 3.4, nous proposons une instanciation du système logique sur un exemple.

3.1 Le système KBE, une logique non normale

Dans un premier temps, nous définissons le langage KBE, puis donnons une sémantique aux formules exprimées de ce langage. Dans un second temps, nous construisons un système axiomatique associé à ce cadre logique.

3.1.1 Le langage $\mathcal{L}_{KBE}$

Soient un ensemble de lettres propositionnelles $\mathcal{P} = \{a, b, c, ...\}$, un ensemble d'agents $\mathcal{N}$ avec $i, j \in \mathcal{N}$ deux agents, et $p \in \mathcal{P}$ une variable propositionnelle. Le langage $\mathcal{L}_{KBE}$ est généré par la

grammaire sous forme de Backus-Naur suivante :

$$\phi ::= p \mid \neg\phi \mid \phi \Rightarrow \phi \mid K_i\phi \mid B_i\phi \mid E_i\phi \mid E_i^d\phi$$

Nous considérons des modalités E_i, E_i^d, K_i et B_i pour chaque agent i. La formule $E_i\phi$ signifie que l'agent i effectue une ou plusieurs actions menant à une conséquence ϕ[1]. Cette modalité représente les effets, i.e. les conséquences des actions effectuées par l'agent i, qu'elles soient délibérées ou non. La formule $E_i^d\phi$ signifie qu'un agent i effectue de manière délibérée une ou plusieurs actions menant à une conséquence ϕ. Cette modalité capture les effets délibérés, i.e. les stratégies mises en œuvre par l'agent. Enfin, les formules $K_i\phi$ et $B_i\phi$ signifient que l'agent i sait que ϕ est vraie, et que l'agent i croit que ϕ est vraie.

3.1.2 Une sémantique de l'intention délibérée

Dans le chapitre 2, nous avons présenté les logiques épistémiques et doxastiques caractérisant la notion de connaissance et de croyance des agents. Ces logiques vont nous permettre en section 3.3 de représenter la notion de dissimulation en lien avec la manipulation. Concernant la représentation des intentions des agents, nous avons présenté deux formalismes logiques : le formalisme BIAT et le formalisme STIT. Cependant, même si ces formalismes représentent bien la notion d'intention, ils ne représentent pas précisément celle d'intention délibérée. En particulier, le formalisme STIT (Lorini and Sartor, 2016) omet une propriété fondamentale qui est l'introspection. Un agent sait toujours ce qu'il a commis de façon délibérée et sait aussi ce qu'il n'a pas commis de façon délibérée. C'est pourquoi, il convient d'introduire un nouvel opérateur modal pour représenter explicitement l'intention délibérée. Nous donnons à ce nouvel opérateur une sémantique formelle, reposant sur une logique non-normale que nous justifions par la suite.

Le cadre KBE

Pour interpréter les formules du langage $\mathcal{L}_{KBE}$, nous considérons le cadre $\mathcal{C} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^d\}_{i\in\mathcal{N}})$ tel que :

1. $\mathcal{W}$ un ensemble de mondes possibles non vide ;
2. $\{\mathcal{B}_i\}_{i\in\mathcal{N}}$ un ensemble de relations binaires telles que :

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{B}_i(w) := \{v \in \mathcal{W} | w\mathcal{B}_i v\}$$

3. $\{\mathcal{K}_i\}_{i\in\mathcal{N}}$ un ensemble de relations binaires telles que :

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{K}_i(w) := \{v \in \mathcal{W} | w\mathcal{K}_i v\}$$

4. $\{\mathcal{E}_i\}_{i\in\mathcal{N}}$ un ensemble de relations binaires telles que :

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{E}_i(w) := \{v \in \mathcal{W} | w\mathcal{E}_i v\}$$

1. Par soucis de lisibilité, nous pourrons utiliser les expressions « faire en sorte que ». Dans tous les cas, l'interprétation sémantique est « effectuer une ou plusieurs actions qui mènent à ».

5. $\{\mathcal{E}_i^d\}_{i\in\mathcal{N}}$ un ensemble de fonctions de voisinage :

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{E}_i^d(w) \in 2^{2^{\mathcal{W}}}$$

Nous définissons un *modèle de KBE* comme $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^d\}_{i\in\mathcal{N}}, V)$ avec $V : \mathcal{P} \to 2^{\mathcal{W}}$ une fonction de valuation. Pour tout monde $w \in \mathcal{W}$, pour toute formule $\phi, \psi \in \mathcal{L}_{KBE}$ et pour tout atome propositionnel $p \in \mathcal{P}$, nous considérons la sémantique :

1. $\mathcal{M}, w \models \top$
2. $\mathcal{M}, w \not\models \bot$
3. $\mathcal{M}, w \models p$ ssi $w \in V(p)$
4. $\mathcal{M}, w \models \neg\phi$ ssi $\mathcal{M}, w \not\models \phi$
5. $\mathcal{M}, w \models \phi \vee \psi$ ssi $\mathcal{M}, w \models \phi$ ou $\mathcal{M}, w \models \psi$
6. $\mathcal{M}, w \models \phi \wedge \psi$ ssi $\mathcal{M}, w \models \phi$ et $\mathcal{M}, w \models \psi$
7. $\mathcal{M}, w \models \phi \Rightarrow \psi$ ssi $\mathcal{M}, w \models \neg\phi$ ou $\mathcal{M}, w \models \psi$
8. $\mathcal{M}, w \models B_i\phi$ ssi $\forall v \in \mathcal{W} : w\mathcal{B}_i v, \mathcal{M}, v \models \phi$
9. $\mathcal{M}, w \models K_i\phi$ ssi $\forall v \in \mathcal{W} : w\mathcal{K}_i v, \mathcal{M}, v \models \phi$
10. $\mathcal{M}, w \models E_i\phi$ ssi $\forall v \in \mathcal{W} : w\mathcal{E}_i v, \mathcal{M}, v \models \phi$
11. $\mathcal{M}, w \models E_i^d\phi$ ssi $|\phi| \in \mathcal{E}_i^d(w)$, avec $|\phi| := \{v \in \mathcal{W} : \mathcal{M}, v \models \phi\}$

Nous rappelons les notations duales pour les modalités normales i.e. pour pour tout monde $w \in \mathcal{W}$, pour tout $(\Diamond_i, \mathcal{R}_i) \in \{(\langle B_i\rangle, \mathcal{B}_i), (\langle K_i\rangle, \mathcal{K}_i), (\langle E_i\rangle, \mathcal{E}_i)\}$, nous avons $\mathcal{M}, w \models \Diamond_i\phi$ si, et seulement si, $\exists v \in \mathcal{W} : w\mathcal{R}_i v, \mathcal{M}, v \models \phi$. Puis, $\mathcal{M}, w \models \langle E_i^d\rangle\phi$ si, et seulement si, $\mathcal{W}\backslash|\phi| \notin \mathcal{E}_i^d(w)$.

Enfin, ϕ est valide dans un modèle $\mathcal{M}$ (noté $\mathcal{M} \models \phi$) si, et seulement si, pour tout monde $w \in \mathcal{W}$, ϕ est satisfiable dans w i.e. $\mathcal{M}, w \models \phi$ est vraie. Une formule ϕ est valide dans un cadre $\mathcal{C}$ (noté $\models_{\mathcal{C}} \phi$ ou $\mathcal{C} \models \phi$) si, et seulement si, pour tout modèle $\mathcal{M}$ fondé sur $\mathcal{C}$, $\mathcal{M} \models \phi$. Nous disons que ϕ est la *conséquence sémantique* d'un ensemble de formules Γ dans $\mathcal{C}$ et notons $\Gamma \models_{\mathcal{C}} \phi$ si, et seulement si, pour tout modèle de KBE $\mathcal{M} \models \Gamma$, implique $\mathcal{M} \models \phi$.

Représentation sémantique des croyances et des connaissances

Les modalités de connaissance et de croyance sont contraintes de telle sorte que, pour tout agent $i \in \mathcal{N}$, $\mathcal{K}_i$ est une relation d'équivalence (transitive, réflexive et symétrique) et $\mathcal{B}_i$ est une relation sérielle, transitive et euclidienne. De plus, les contraintes entre ces deux modalités ont déjà été étudiées dans (Stalnaker, 2006) et il est usuel de considérer qu'un agent i croit ce qu'il sait, c'est-à-dire :

$$\forall w \in \mathcal{W} : \mathcal{B}_i(w) \subseteq \mathcal{K}_i(w) \quad (KB1)$$

Ensuite, si un agent i croit quelque chose, alors il sait qu'il le croit :

$$\forall w, u, v \in \mathcal{W} : w\mathcal{K}_i u \wedge u\mathcal{B}_i v \Rightarrow w\mathcal{B}_i v \quad (KB2)$$

De la même façon, un agent sait ce qu'il ne croit pas, c'est-à-dire :

$$\forall w, u, v \in \mathcal{W} : w\mathcal{K}_i u \wedge w\mathcal{B}_i v \Rightarrow u\mathcal{B}_i v \quad (KB3)$$

Représentation sémantique des effets des actions

Il est usuellement accepté par les approches modernes comme les approches STIT de considérer les intentions des agents de « veiller à ce que quelque chose soit vrai » comme des relations d'équivalences. Nous avons donc pour tout agent $i \in \mathcal{N}$, $\mathcal{E}_i$ est une relation réflexive, transitive et euclidienne [2].

La réflexivité notée E_T, exprime le fait qu'une fois les actions menant à une certaine conséquence ϕ ont été effectuées par un agent i, alors cette conséquence ϕ est nécessairement vraie dans le monde courant.

$$\forall w \in \mathcal{W} : w\mathcal{E}_i w \quad (E_T)$$

De cette contrainte, nous déduisons immédiatement que la relation $\mathcal{E}_i$ est aussi *sérielle*. Cela traduit que si un agent i veille à ce qu'une ou plusieurs actions mènent à une certaine conséquence ϕ, il n'est pas le cas que cette action ou série d'actions puissent mener au contraire dans le monde courant.

La relation $\mathcal{E}_i$ est aussi *transitive* puisque lorsqu'un agent i effectue une ou plusieurs actions qui ont pour effets de rendre vrai une proposition ϕ, ce dernier effectue une ou plusieurs actions qui ont pour effets que son ou ses actions soient bien effectuées et mènent à ce que ϕ soit vraie, c'est-à-dire :

$$\forall w, u, v \in \mathcal{W} : w\mathcal{E}_i u \wedge u\mathcal{E}_i v \Rightarrow w\mathcal{E}_i v \quad (E_4)$$

Enfin, s'il n'est pas le cas qu'un agent i effectue une ou plusieurs actions qui mènent à une certaine conséquence, alors l'agent i effectue une ou plusieurs actions qui mènent à ce qu'il ne réalise pas la ou les actions qui mènent à rendre vrai cette conséquence. Ainsi, la relation $\mathcal{E}_i$ est *euclidienne*, c'est-à-dire :

$$\forall w, u, v \in \mathcal{W} : w\mathcal{E}_i u \wedge w\mathcal{E}_i v \Rightarrow u\mathcal{E}_i v \quad (E_5)$$

Représentation sémantique des intentions délibérées

Si représenter les effets des actions, au sens large est une relation d'équivalence, il n'en est pas de même pour la notion d'intention délibérée. En effet, contrairement à la représentation des effets des actions qui est représentée par une relation de Kripke, la notion d'effets délibérés est représentée par une fonction de voisinage. Cette différence sémantique est justifiée par le fait que lorsqu'un agent a une intention délibérée, ce dernier envisage un ensemble de mondes possibles dans lesquels son intention a été effectuée. De plus, les relations de Kripke ne permettent pas d'exprimer, par exemple, le fait qu'un agent ne puisse pas amener de façon délibérée quelque chose qu'il sait toujours vrai comme les tautologies, et ce, à cause du principe de nécessitation qui est

2. Une relation d'équivalence est par définition une relation réflexive, transitive et symétrique, mais de manière équivalente, cela revient à considérer toute relation réflexive, transitive et euclidienne.

valide dans tout cadre de Kripke. Ainsi, la première différence sémantique avec la modalité E_i qui représente les effets des actions, est qu'un agent i ne peut pas effectuer une ou plusieurs actions de façon délibérée qui mènent à rendre vrai une tautologie puisqu'il sait qu'elle est toujours vraie. Cette contrainte, notée $\overline{E^d_{Nec}}$, est traduite sémantiquement par le fait que l'ensemble de tous les mondes possibles ne peut appartenir à aucun voisinage, c'est-à-dire :

$$\forall w \in \mathcal{W} : W \notin \mathcal{E}^d_i(w) \quad (\overline{E^d_{Nec}})$$

D'autre part, il existe un lien logique entre la modalité des effets délibérés et celle des effets des actions qui peuvent être délibérés ou non. En effet, si un agent i effectue de façon délibérée une ou plusieurs actions, cet agent effectue bien ces actions. Ce lien est représenté par la contrainte E^dE qui est traduite sémantiquement par le fait que l'ensemble des mondes possibles atteignables par la relation $\mathcal{E}_i$ sont toujours inclus dans tout voisinage de mondes possibles de $\mathcal{E}^d_i$, c'est-à-dire :

$$\forall w \in \mathcal{W} : S \in \mathcal{E}^d_i(w) \Rightarrow \mathcal{E}_i(w) \subseteq S \quad (E^dE)$$

Par ailleurs, lorsqu'un agent i effectue de manière délibérée une ou plusieurs actions menant à rendre vrai une proposition ϕ tout en effectuant de manière délibérée une ou plusieurs actions menant à vrai une autre proposition ψ, alors cet agent i effectue une ou plusieurs actions de manière délibérée menant à rendre vrai la conjonction $\phi \wedge \psi$, c'est-à-dire :

$$\forall w \in \mathcal{W} : S \in \mathcal{E}^d_i(w) \wedge T \in \mathcal{E}^d_i(w) \Longrightarrow S \cap T \in \mathcal{E}^d_i(w) \quad (E_{\Rightarrow,\wedge})$$

Cependant, nous ne pouvons pas considérer la réciproque de cette propriété car l'intention délibérée concerne un tout et n'est pas équivalente à la somme de ses parties. Par exemple, quand nous décidons de manger un gâteau aux noisettes, nous ne décidons pas de manger délibéremment la pâte du gâteau et de manger les noisettes de façon complètement indépendantes. De plus, pour des raisons purement techniques, la réciproque, qui est donc la propriété $\forall w \in \mathcal{W} : S \cap T \in \mathcal{E}^d_i(w) \Longrightarrow S \in \mathcal{E}^d_i(w) \wedge T \in \mathcal{E}^d_i(w)$ et qui est associée au théorème $E^d_i(\phi \wedge \psi) \Rightarrow E^d_i\phi \wedge E^d_i\psi$, ne peut pas être considérée. En effet, un résultat immédiat et issue de la topologie (Pacuit, 2017) montre que la réciproque est aussi équivalente à $\forall w \in \mathcal{W} : S \in \mathcal{E}^d_i(w) \wedge S \subseteq T \Longrightarrow T \in \mathcal{E}^d_i(w)$. Par conséquent, si la réciproque était considérée, nous obtiendrions un système logique inconsistant en raison de la contrainte $\forall w \in \mathcal{W} : W \notin \mathcal{E}^d_i(w)(\overline{E^d_{Nec}})$.

Enfin, une caractéristique très importante de l'intention délibérée est qu'elle est introspective par rapport aux connaissances des agents. Un agent sait toujours ce qu'il effectue et ce qu'il n'effectue pas de manière délibérée. Ainsi, la modalité d'action délibérée dispose de l'introspection positive (E^d_{KP}) et de l'introspection négative (E^d_{KN}) par rapport à la connaissance des agents. Ces contraintes liées à l'introspection, sont décrites de façon indépendantes par les relations suivantes :

$$\forall w \in \mathcal{W} : \mathcal{E}^d_i(w) \subseteq \bigcap_{v \in \mathcal{W} : w\mathcal{K}_i v} \mathcal{E}^d_i(v) \quad (E^d_{KP})$$

$$\forall w, v \in \mathcal{W}, \forall S \in 2^{\mathcal{W}} : S \notin \mathcal{E}^d_i(w) \Longrightarrow (w\mathcal{K}_i v \Rightarrow S \notin \mathcal{E}^d_i(v)) \quad (E^d_{KN})$$

Ces contraintes signifient littéralement que lorsqu'un agent i effectue une ou plusieurs actions de façon délibérée menant à une certaine conséquence, il sait ce qu'il est en train de faire de façon délibérée. Il en est de même lorsqu'il n'est pas le cas qu'un agent i effectue une ou plusieurs actions de façon délibérée menant à une certaine conséquence, alors cet agent i sait qu'il ne les a pas effectué de façon délibérée. Ces deux contraintes reviennent alors à considérer une seule et unique contrainte qui est représentée par la relation ensembliste suivante :

$$\forall w \in \mathcal{W} : \mathcal{E}_i^d(w) = \bigcap_{v \in \mathcal{W} : w \mathcal{K}_i v} \mathcal{E}_i^d(v) \quad (E_{KP}^d + E_{KN}^d)$$

Illustration par un exemple des relations sémantiques

L'exemple 3.1 illustre les différences fondamentales entre les intentions délibérées et celles non délibérées. L'intention délibérée se caractérise avant tout comme le choix calculé d'un agent pour atteindre un but. Cet agent a pleinement connaissance de ces conséquences. L'intention non délibérée représente l'ensemble des conséquences des actions effectuées par un agent, qu'elles aient été délibérées ou non. L'agent n'a pas nécessairement connaissance de toutes ces conséquences.

Exemple 3.1 :

Supposons une situation dans laquelle il existe un agent meurtrier i veillant de façon délibérée à assassiner une victime j. Pour représenter cette situation nous considérons trois variables propositionnelles p pour désigner que « l'agent i tue l'agent j en la poignardant », q pour désigner que « l'agent i se fait arrêter par la police » et r pour désigner que « il n'y a aucun témoin ». Pour cette situation, plusieurs mondes possibles sont envisageables comme par exemple, l'agent i tue la victime et ne se fait pas arrêter par la police, la victime était déjà morte avant que l'agent ne la poignarde, l'agent i est pris de remords et ne tue pas la victime, ou bien la victime j se réveille et fait fuir l'agent i, etc. Faisons l'hypothèse qu'un modèle $\mathcal{M}$ décrit tous ces mondes possibles[3]*. Nous supposons $\mathcal{M}$ tel que :*

- $\mathcal{W} = \{w, u, v, x, y, z, a\}$;
- $V(p) = \{w, u, v\}, V(r) = \{w, x, y, z, a\}, V(q) = \{v, y, z\}$.

Les mondes possibles considérés ici représentent donc les situations :

- *w : « il n'y a aucun témoin et l'agent i tue la victime j et ne se fait pas arrêter » ;*
- *u : « il y a un témoin et l'agent i tue la victime j mais ne se fait par arrêter » ;*
- *v : « il y a un témoin et l'agent i tue la victime j et l'agent i se fait arrêter » ;*
- *x : « la victime était déjà morte, aucun témoin, et l'agent i ne se fait pas arrêter » ;*
- *y : « la victime était déjà morte, aucun témoin, et l'agent i se fait arrêter » ;*
- *z : « le tueur est pris de remords, aucun témoin, et ne tue pas la victime mais l'agent i se fait quand même arrêter pour tentative d'assassinat » ;*
- *a : « il ne se passe rien ».*

3. Nous pourrions considérer un grand nombre de mondes possibles pour décrire cet exemple, mais pour des raisons de clarté, nous ne considérons qu'un nombre restreint de mondes possibles.

Dans le monde possible w, l'agent i a donc l'intention délibérée de tuer la victime et parvient à le réaliser. De plus, il a aussi l'intention qu'il n'y ait aucun témoin. Ainsi la fonction de voisinage est telle que $\mathcal{E}_i^d(w) = \{\{w, u, v\}, \{w, x, y, z, a\}\}$. Elle représente chaque stratégie indépendante que l'agent i a eu pour intention de mettre en place dans w.

- *$\{w, u, v\}$ représente l'intention délibérée de l'agent i de tuer la victime ;*
- *$\{w, x, y, z, a\}$ représente l'intention délibérée de l'agent i de veiller à ce qu'il n'y ait aucun témoin de la scène de crime.*

Dans le monde w, l'agent i parvient avec succès à tuer la victime sans se faire arrêter, ainsi $\mathcal{E}_i(w) = \{w\}$. De plus, dans w puisque l'agent i a veillé de façon délibérée à ce qu'il n'y ait aucun témoin, nous déduisons que l'agent sait que p et r sont vraies. Nous avons donc $\mathcal{K}_i(w) = \{w\}$, unique monde possible où p et r sont vraies simultanément. L'agent i ne peut donc discerner dans w un autre montre que celui-ci. Par ailleurs, dans w, l'agent sait que la police ne l'arrête pas.

Nous remarquons que dans ce modèle, les propriétés du cadre comme $\forall w \in \mathcal{W} : S \in \mathcal{E}_i^d(w) \Rightarrow \mathcal{E}_i(w) \subseteq S$ sont naturellement respectées et illustrent bien le fait que puisque l'agent i veille de façon délibérée dans w à tuer la victime, alors cet agent i veille bien à ce qu'elle soit tuée. De plus, puisque $|p| \in \mathcal{E}_i(w)$ nous avons $\mathcal{M}, w \models E_i^d p$, c'est-à-dire l'agent i veille de façon délibérée à tuer la victime. Puisque le choix stratégique de l'agent était de parvenir à tuer la victime, il a donc mis en place une stratégie permettant de rendre vrai p.

3.1.3 Système axiomatique

Nous notons $\vdash \phi$ pour dire que ϕ est un théorème qui peut être déduit à partir d'un système de preuves à la Hilbert. Ainsi, la figure 3.1 représente l'ensemble des axiomes dans un tel système. Chaque axiome provient directement des contraintes sémantiques imposées sur le cadre KBE. De plus, nous avons comme règles de déduction : le modus ponens et la substitution qui s'appliquent sur toutes les modalités. Cependant, la règle de nécessitation ne peut s'appliquer uniquement sur les modalités normales, i.e. B_i, K_i et E_i pour tout agent $i \in \mathcal{N}$. Enfin, la définition de KBE-déductibilité est fondée sur la notion de $\mathcal{S}$-déductibilité présentée en section 2.1.1 où $\mathcal{S} = KBE$.

(CP)	Tous les théorèmes du CP
(RE)	$\forall \Box_i \in \{B_i, K_i, E_i, E_i^d\}$ Si $\vdash \phi \Leftrightarrow \psi$ alors $\vdash \Box_i\phi \Leftrightarrow \Box_i\psi$
$(DUAL)$	$\forall(\Box_i, \Diamond_i) \in \{(B_i, \langle B_i\rangle), (K_i, \langle K_i\rangle), (E_i, \langle E_i\rangle), (E_i^d, \langle E_i^d\rangle)\} \vdash \Box_i\phi \Leftrightarrow \neg\Diamond_i\neg\phi$
$(S5_{K_i})$	Tous les théorèmes de S5 sont vérifiés pour K_i
$(S5_{E_i})$	Tous les théorèmes de S5 sont vérifiés pour E_i
$(KD45_{B_i})$	Tous les théorèmes de KD45 sont vérifiés pour B_i
$(E_i^d E_i)$	$\vdash E_i^d\phi \Rightarrow E_i\phi$
$(C_{E_i^d})$	$\vdash E_i^d\phi \wedge E_i^d\psi \Rightarrow E_i^d(\phi \wedge \psi)$
$(\neg N_{E_i^d})$	$\vdash \neg E_i^d\top$
$(4_{K_iE_i})$	$\vdash E_i^d\phi \Rightarrow K_iE_i^d\phi$
$(5_{K_iE_i})$	$\vdash \neg E_i^d\phi \Rightarrow K_i\neg E_i^d\phi$
(KB)	$\vdash K_i\phi \Rightarrow B_i\phi$
(4_{KB})	$\vdash B_i\phi \Rightarrow K_iB_i\phi$
(5_{KB})	$\vdash \neg B_i\phi \Rightarrow K_i\neg B_i\phi$

FIGURE 3.1 – Système axiomatique du cadre KBE

3.2 Correction et complétude du système KBE

Dans cette section, dans un premier temps, nous démontrons que le système axiomatique présenté par la figure 3.1 est correct. Dans un second temps, nous démontrons que ce système axiomatique est complet. Dans un troisième temps, nous montrons que ce système vérifie les théorèmes de déduction, est fortement correct et fortement complet. Enfin, nous démontrons quelques théorèmes pouvant être déduits dans le système KBE.

3.2.1 Correction

Pour démontrer la correction de notre système KBE, nous considérons dans cette sous-section un cadre quelconque $\mathcal{C} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^d\}_{i\in\mathcal{N}})$.

Modalités normales

Pour tout agent $i \in \mathcal{N}$, puisque les modalités K_i, B_i et E_i sont des modalités normales, il est standard de prouver la correction d'un système S5 pour des relations d'équivalences telles que $\mathcal{E}_i$ et $\mathcal{K}_i$ (Blackburn et al., 2002). Ainsi, pour des raisons de clarté et de concision, nous ne les donnons pas dans ce chapitre et nous renvoyons le lecteur intéressé à ces preuves. De plus, puisque B_i est aussi une modalité normale, il est de même standard de prouver qu'un système KD45 est correct vis à vis des relations $\mathcal{B}_i$ qui sont sérielles, transitives et euclidiennes.

Modalités non normales

Cependant, pour ce qui est des modalités non normales pour représenter les intentions délibérées ainsi que les liens relationnels existants entre les modalités E_i et E_i^d du système. Nous

prouvons tout d'abord que la contrainte d'introspection positive des connaissances sur les intentions délibérées est valide dans ce cadre, i.e. $\models E_i^d p \Rightarrow K_i E_i^d p$.

Proposition 3.1 :

$$\mathcal{C} \models E_i^d p \Rightarrow K_i E_i^d p$$

si, et seulement si,

$$\forall w \in \mathcal{W} : \mathcal{E}_i^d(w) \subseteq \bigcap_{v \in \mathcal{W} : w \mathcal{K}_i v} \mathcal{E}_i^d(v)$$

Démonstration. ($\Rightarrow$) Par contraposition, supposons un cadre $\mathcal{C}$ tel que :

$$\exists w \in \mathcal{W} : \mathcal{E}_i^d(w) \not\subseteq \bigcap_{v \in \mathcal{W} : w \mathcal{K}_i v} \mathcal{E}_i^d(v)$$

Définissons un modèle $\mathcal{M}$ tel qu'il existe deux mondes distincts $w, v \in \mathcal{W} : v \neq w$, $\mathcal{E}_i^d(w) = \{\{w\}\}$, $\mathcal{E}_i^d(v) = \{\{\}\}$, $w \mathcal{K}_i v$ et $V(p) = \{w\}$. Comme $|p| \in \mathcal{E}_i^d(w)$, nous avons que $\mathcal{M}, w \models E_i^d p$. De plus, puisque $|p| \notin \mathcal{E}_i^d(v)$, nous avons $\mathcal{M}, v \models \neg E_i^d p$. Or, comme $w \mathcal{K}_i v$, nous avons donc que $\mathcal{M}, w \models \neg K_i E_i^d p$. Nous avons donc prouvé qu'il existe un modèle tel que $\mathcal{M} \not\models E_i^d p \Rightarrow K_i E_i^d p$ et donc $\mathcal{C} \not\models E_i^d p \Rightarrow K_i E_i^d p$.

($\Leftarrow$) Supposons $\mathcal{C} \not\models E_i^d p \Rightarrow K_i E_i^d p$, c'est-à-dire, il existe un modèle $\mathcal{M}$ et un monde $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models E_i^d p \wedge \neg K_i E_i^d p$. Nous avons donc que $|p| \in \mathcal{E}_i^d(w)$. De plus, il existe $v \in \mathcal{W}$ tel que : $w \mathcal{K}_i v$ et $\mathcal{M}, v \models \neg E_i^d p$, i.e $|p| \notin \mathcal{E}_i^d(v)$. Par conséquent, nous déduisons que :

$$|p| \notin \bigcap_{v \in \mathcal{W} : w \mathcal{K}_i v} \mathcal{E}_i^d(v)$$

Nous venons donc de prouver que :

$$\exists w \in \mathcal{W} : \mathcal{E}_i^d(w) \not\subseteq \bigcap_{v \in \mathcal{W} : w \mathcal{K}_i v} \mathcal{E}_i^d(v)$$

□

Ensuite, nous prouvons que l'introspection négative des intentions délibérées par rapport aux connaissances est valide dans notre cadre.

Proposition 3.2 :

$$\mathcal{C} \models \neg E_i^d p \Rightarrow K_i \neg E_i^d p$$

si, et seulement si,

$$\forall w, v \in \mathcal{W}, \forall S \in 2^{\mathcal{W}} : S \notin \mathcal{E}_i^d(w) \Longrightarrow (w \mathcal{K}_i v \Rightarrow S \notin \mathcal{E}_i^d(v))$$

Démonstration. ($\Rightarrow$) Par contraposition, supposons un cadre $\mathcal{C}$ tel qu'il existe $w, v \in \mathcal{W}$, et il existe $S \in 2^{\mathcal{W}}$:

$$S \notin \mathcal{E}_i^d(v) \wedge w\mathcal{K}_i v \wedge S \in \mathcal{E}_i^d(v)$$

Définissons un modèle $\mathcal{M}$ tel que $\mathcal{E}_i^d(w) = \{\{\}\}$, $\mathcal{E}_i^d(v) = \{\{w, v\}\}$, $w\mathcal{K}_i v$ et $V(p) = \{w, v\}$. Puisque $|p| \notin \mathcal{E}_i^d(w)$, nous avons alors que $\mathcal{M}, w \models \neg E_i^d p$. De plus, comme $|p| \in \mathcal{E}_i^d(v)$, nous avons que $\mathcal{M}, v \models E_i^d p$. Or puisque $w\mathcal{K}_i v$, nous avons que $\mathcal{M}, w \models \neg K_i \neg E_i^d p$. Ainsi $\mathcal{M}, w \models \neg E_i^d p \wedge \neg K_i \neg E_i^d p$. Nous avons donc prouvé qu'il existe un modèle tel que $\mathcal{M} \not\models \neg E_i^d p \Rightarrow K_i \neg E_i^d p$ et donc $\mathcal{C} \not\models \neg E_i^d p \Rightarrow K_i \neg E_i^d p$.

($\Leftarrow$) Supposons par contraposition que $\mathcal{C} \not\models \neg E_i^d p \Rightarrow K_i \neg E_i^d p$, c'est-à-dire il existe un modèle $\mathcal{M}$ et un monde $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models \neg E_i^d p \wedge \neg K_i \neg E_i^d p$. Nous avons donc que $|p| \notin \mathcal{E}_i^d(w)$. De plus, il existe $v \in \mathcal{W}$ tel que : $w\mathcal{K}_i v$ et $\mathcal{M}, v \models E_i^d p$, i.e $|p| \in \mathcal{E}_i^d(v)$. Par conséquent, nous venons donc de prouver que :

$$\exists w, v \in \mathcal{W}, \exists S \in 2^{\mathcal{W}} : S \notin \mathcal{E}_i^d(w) \wedge w\mathcal{K}_i v \wedge S \in \mathcal{E}_i^d(v)$$

□

L'axiome (C) est aussi valide dans le système KBE.

Proposition 3.3 :

$$\mathcal{C} \models E_i^d p \wedge E_i^d q \Rightarrow E_i^d(p \wedge q)$$

si, et seulement si,

$$\forall w \in \mathcal{W} : S \in \mathcal{E}_i(w) \wedge T \in \mathcal{E}_i(w) \Longrightarrow S \cap T \in \mathcal{E}_i(w)$$

Démonstration. ($\Rightarrow$) Par contraposition, considérons un modèle $\mathcal{M}$ tel que : $i(p) = \{u, v\}$, $i(q) = \{v, w\}$ et $\mathcal{E}_i^d(w) = \{\{u, v\}, \{v, w\}\}$. Nous avons donc $|p|, |q| \in \mathcal{E}_i^d(w)$, et $|p| \cap |q| \notin \mathcal{E}_i^d(w)$. Or $|p| \cap |q| = |p \wedge q|$. Donc $|p \wedge q| \notin \mathcal{E}_i^d(w)$ et $\mathcal{M}, w \models \neg E_i^d(p \wedge q)$. De plus, comme $|p| \in \mathcal{E}_i(w) \wedge |q| \in \mathcal{E}_i(w)$, nous avons $\mathcal{M}, w \models E_i^d p \wedge E_i^d q$. Par conséquent, $\mathcal{M}, w \models E_i^d p \wedge E_i^d q \wedge \neg E_i^d(p \wedge q)$. Ainsi, nous venons de prouver que $\mathcal{C} \not\models E_i^d p \wedge E_i^d q \Rightarrow E_i^d(p \wedge q)$.

($\Leftarrow$) Supposons que $\mathcal{C} \not\models E_i^d p \wedge E_i^d q \Rightarrow E_i^d(p \wedge q)$, c'est-à-dire il existe un modèle $\mathcal{M}$ et un monde $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models E_i^d p \wedge E_i^d q \wedge \neg E_i^d(p \wedge q)$. Donc $|p| \in E_i(w)$ et $|q| \in E_i(w)$. De plus, comme $\mathcal{M}, w \models \neg E_i^d(p \wedge q)$, alors $|p \wedge q| \notin \mathcal{E}_i^d(w)$ i.e $|p| \cap |q| \notin \mathcal{E}_i^d(w)$. Ainsi, nous venons de prouver qu'il existe un monde $w \in \mathcal{W}$, $S \in \mathcal{E}_i(w) \wedge T \in \mathcal{E}_i(w) \wedge S \cap T \notin \mathcal{E}_i(w)$.

□

Proposition 3.4 :

$$\mathcal{C} \models \neg E_i^d \top$$

si, et seulement si,

$$\forall w \in \mathcal{W} : \mathcal{W} \notin \mathcal{E}_i(w)$$

Démonstration. ($\Rightarrow$) Par contraposition, considérons un modèle $\mathcal{M}$ tel que pour $w \in \mathcal{W}$, $\{\mathcal{W}\} = \mathcal{E}_i^d(w)$ donc $\mathcal{W} \in \mathcal{E}_i^d(w)$. Or $|\top| = \mathcal{W}$. Donc $\mathcal{M}, w \models E_i^d\top$. Nous venons donc de prouver qu'il existe un modèle tel que si $\exists w \in \mathcal{W} : \mathcal{W} \in \mathcal{E}_i^d(w)$ alors $\mathcal{C} \not\models \neg E_i^d\top$.

($\Leftarrow$) Par contraposition, supposons qu'il existe un modèle $\mathcal{M}$ tel qu'il existe $w \in \mathcal{W}$, $\mathcal{M}, w \models E_i^d\top$. Nous avons donc immédiatement que $|\top| \in \mathcal{E}_i(w)$ i.e $\mathcal{W} \in \mathcal{E}_i^d(w)$.

□

Ensuite, nous prouvons que les effets délibérées impliquent les mêmes effets sur les actions. L'axiome ($E_i^d E_i$) préserve la validité dans le cadre KBE.

Proposition 3.5 :

$$\mathcal{C} \models E_i^d p \Rightarrow E_i p$$

si, et seulement si,

$$\forall w \in \mathcal{W}, \forall S \in 2^{\mathcal{W}} : S \in \mathcal{E}_i^d(w) \Longrightarrow \mathcal{E}_i(w) \subseteq S$$

Démonstration. ($\Rightarrow$) Par contraposition, supposons un cadre $\mathcal{C}$ tel que :

$$\exists (w, S) \in \mathcal{W} \times 2^{\mathcal{W}} : S \in \mathcal{E}_i^d(w) \wedge \mathcal{E}_i(w) \not\subseteq S$$

Soit $w \in \mathcal{W}$. Considérons un modèle $\mathcal{M}$ tel que $\mathcal{E}_i^d(w) = \{\{\}\}$, $V(p) = \emptyset$ avec $p \in \mathcal{P}$, $w\mathcal{E}_i w$ et posons $S = |p|$. Nous avons donc que $|p| = \emptyset$, $|p| \in \mathcal{E}_i^d(w)$ et donc que $\mathcal{M}, w \models E_i^d p$. De plus, comme $\mathcal{M}, w \models \neg p$ et $w\mathcal{E}_i w$. Nous avons donc qu'il existe $v = w$ tel que $w\mathcal{E}_i v, \mathcal{M}, v \models \neg p$. Donc $v \in \mathcal{E}_i(w)$ et $v \notin |p|$, i.e. $\mathcal{M}, w \models \neg E_i p$. Donc $\mathcal{M}, w \models E_i^d p \wedge \neg E_i p$. Nous avons donc montré qu'il existe un modèle $\mathcal{M}$ satisfaisant la contrainte du cadre $\mathcal{C}$ et un monde $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models E_i^d p \Rightarrow E_i p$.

($\Leftarrow$) Supposons $\mathcal{C} \not\models E_i^d p \Rightarrow K_i E_i^d p$, c'est-à-dire il existe un modèle $\mathcal{M}$ et un monde $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models E_i^d p \wedge \neg E_i p$. Nous avons donc que $|p| \in \mathcal{E}_i^d(w)$ et il existe $v \in \mathcal{W}$ tel que $w\mathcal{E}_i v$ et $\mathcal{M}, v \models \neg p$. Donc par définition $v \notin |p|$. Par conséquent, nous venons de prouver que :

$$\exists (w, S) \in \mathcal{W} \times 2^{\mathcal{W}} : S \in \mathcal{E}_i^d(w) \wedge \mathcal{E}_i(w) \not\subseteq S$$

□

Enfin, les règles (RE) i.e. si $\vdash \phi \Leftrightarrow \psi$ alors $\vdash E_i^d\phi \Leftrightarrow E_i^d\psi$ et ($DUAL$) i.e. $\models E_i^d\phi \Leftrightarrow \neg\langle E_i^d\rangle\neg\phi$ préservent aussi la validité et elles sont habituelles dans les sémantiques de voisinage (Pacuit, 2017). Nous pouvons désormais prouver que le système KBE est correct.

Proposition 3.6 : *Le système KBE est correct.*

Démonstration. (1) Prouvons que la substitution préserve la validité. Soient $\phi \in \mathcal{L}_{KBE}$ une formule et $p_{a_1}, ..., p_{a_n} \in \mathcal{P}$ les atomes propositionnels contenus dans la formule ϕ avec $n \in \mathbb{N}$. Considérons $\theta = \phi(\psi_1/p_{a_1}, ..., \psi_n/p_{a_n})$ la formule obtenue par substitution uniforme sur ϕ et $\psi_1, \ldots, \psi_n \in \mathcal{L}_{KBE}$ des formules. Nous voulons prouver que si ϕ est valide dans un cadre de KBE $\mathcal{C}$ alors θ est valide dans $\mathcal{C}$. Par contraposition, supposons $\mathcal{C} \not\models \theta$. Donc, il existe un modèle $\mathcal{M} = (\mathcal{C}, V)$ et un monde $w \in \mathcal{W}$ tel que : $\mathcal{M}, w \not\models \theta$. Construisons un modèle pour ϕ, $\mathcal{M}' = (\mathcal{C}, V')$ tel que :

— $\forall j \in \mathbb{N} : 1 \geq j \geq n, \mathcal{M}, w \models \psi_j \Longrightarrow w \in V'(p_{a_j})$
— $\forall j \in \mathbb{N} : 1 \geq j \geq n, \mathcal{M}, w \not\models \psi_j \Longrightarrow w \notin V'(p_{a_j})$
— $\forall p \in \mathcal{P} : \forall j \in \mathbb{N} : p \neq p_{a_j}, w \notin V'(p)$ [4]

Puisque $\mathcal{M}, w \not\models \theta$, nous avons que la combinaison des ψ_j rend invalide la formule θ dans le modèle $\mathcal{M}$ et le monde w. Comme pour tout ψ_j nous associons un atome p_{a_j} du langage ayant la même valeur de vérité que la formule ψ_j. La combinaison des p_{a_j} rend fausse la formule ϕ dans le modèle $\mathcal{M}'$ et le monde w. Nous avons donc que $\mathcal{M}', w \not\models \phi$. Donc nous venons de prouver par contraposition que la substitution préserve la validité, i.e. si ϕ est valide dans le cadre $\mathcal{C}$ alors sa substitution ψ l'est également dans $\mathcal{C}$.

(2) Prouvons que le modus ponens préserve la validité. Supposons $\vdash \phi$ et $\vdash \phi \Rightarrow \psi$ deux théorèmes du système KBE. Nous avons donc que pour tout cadre $\mathcal{C}$ logique KBE, $\mathcal{C} \models \phi$ et $\mathcal{C} \models (\phi \Rightarrow \psi)$. Donc pour tout modèle $\mathcal{M}$ et pour tout monde w, nous avons $\mathcal{M}, w \models \phi$ et $\mathcal{M}, w \models \phi \Rightarrow \psi$, c'est-à-dire $\mathcal{M}, w \models \phi \wedge (\neg\phi \vee \psi)$, i.e. $\mathcal{M}, w \models (\phi \wedge \neg\phi) \vee (\phi \wedge \psi)$. Donc, $\mathcal{M}, w \models (\phi \wedge \psi)$, i.e. $\mathcal{M}, w \models \phi$ et $\mathcal{M}, w \models \psi$. Ainsi, $\mathcal{M}, w \models \psi$. Nous avons donc prouvé que pour tout modèle $\mathcal{M}$ et tout monde w, ψ est valide, i.e. $\mathcal{C} \models \psi$.

(3) Prouvons que la nécessitation préserve la validité pour tout agent $i \in \mathcal{N}$, pour $(\Box, \mathcal{R}) \in \{(B_i, \mathcal{B}_i), (K_i, \mathcal{K}_i), (E_i, \mathcal{E}_i)\}$. Supposons $\vdash \phi$ un théorème du système KBE. Donc pour tout cadre logique KBE $\mathcal{C}$, nous avons $\mathcal{C} \models \phi$. Donc pour tout modèle $\mathcal{M}$ et $\forall v \in \mathcal{W}, \mathcal{M}, v \models \phi$. Donc $\forall w, v \in \mathcal{W}, w\mathcal{R}v : \mathcal{M}, v \models \phi$, c'est-à-dire $\forall w \in \mathcal{W}, \mathcal{M}, w \models \Box\phi$. Nous avons donc prouvé que si $\mathcal{C} \models \phi$ alors $\mathcal{C} \models \Box\phi$, c'est-à-dire la nécessitation préserve la validité. □

3.2.2 Complétude

Dans cette section, nous prouvons la complétude de notre système logique KBE. Pour ce faire, nous appliquons la méthode de Henkin qui utilise les ensembles maximaux S-consistants. Pour tout rappel sur ces notions, nous renvoyons le lecteur à la section 2.1.1.

Définition du modèle canonique

Nous définissons un modèle canonique pour le système KBE. Ce modèle canonique nous permet alors de prouver la complétude du système KBE. Il repose sur la notion de modèle

4. Puisque p n'est pas un atome concerné par la substitution, peu importe le choix de $w \notin h'(p)$ ou $w \in V'(p)$, cela n'affectera pas la démonstration. Nous nous assurons juste, ici, d'avoir bien défini le modèle.

canonique minimal dans les sémantiques de voisinage. Pour tout détail supplémentaire, le lecteur peut consulter (Pacuit, 2017).

Définition 3.1 - *Modèle canonique pour KBE* : *Considérons un modèle du cadre KBE* $\mathcal{M}^c = (\mathcal{W}^c, \{\mathcal{B}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{K}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^{d^c}\}_{i\in\mathcal{N}}, V^c)$. *Nous appelons* $\mathcal{M}^c$ *un* modèle canonique pour KBE *si, et seulement si,* $\mathcal{M}^c$ *est tel que :*

— $\mathcal{W}^c$ *un ensemble non vide de mondes possibles où chaque monde représente un ensemble maximal de formules KBE-consistant;*

— $\{\mathcal{B}_i^c\}_{i\in\mathcal{N}}$ *un ensemble de relations binaires telles que :*

$$\forall i \in \mathcal{N}, \forall w, v \in \mathcal{W} : w\mathcal{B}_i^c v \text{ ssi } B_i\phi \in w \Rightarrow \phi \in v$$

— $\{\mathcal{K}_i^c\}_{i\in\mathcal{N}}$ *un ensemble de relations binaires telles que :*

$$\forall i \in \mathcal{N}, \forall w, v \in \mathcal{W} : w\mathcal{K}_i^c v \text{ ssi } K_i\phi \in w \Rightarrow \phi \in v$$

— $\{\mathcal{E}_i^c\}_{i\in\mathcal{N}}$ *un ensemble de relations binaires telles que :*

$$\forall i \in \mathcal{N}, \forall w, v \in \mathcal{W} : w\mathcal{E}_i^c v \text{ ssi } E_i\phi \in w \Rightarrow \phi \in v$$

— $\{\mathcal{E}_i^{d^c}\}_{i\in\mathcal{N}}$ *un ensemble de fonctions de voisinage telles que :*

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{E}_i^{d^c}(w) := \{||\phi|| : E_i\phi \in w\} \text{ avec } ||\phi|| := \{w | w \in \mathcal{W}^c \wedge \phi \in w\}$$

— $V^c : \mathcal{P} \to 2^{\mathcal{W}}$ *la fonction d'interprétation telle que :*

$$\forall p \in \mathcal{P}, V^c(p) = ||p||, \text{ avec } ||p|| := \{w | w \in \mathcal{W}^c \wedge p \in w\}$$

Nous remarquons que la définition du modèle canonique pour les fonctions de voisinage repose sur la notion de modèle canonique minimal telle qu'elle est décrite dans (Pacuit, 2017). Dans la suite, nous utilisons les notations suivantes :

1. $\forall i \in \mathcal{N}, \forall w \in \mathcal{W}^c, \mathcal{K}_i^*(w) := \{\phi | K_i\phi \in w\}$;
2. $\forall i \in \mathcal{N}, \forall w \in \mathcal{W}^c, \mathcal{B}_i^*(w) := \{\phi | B_i\phi \in w\}$;
3. $\forall i \in \mathcal{N}, \forall w \in \mathcal{W}^c, \mathcal{E}_i^*(w) := \{\phi | E_i\phi \in w\}$.

Lemme de vérité

Soit un modèle canonique $\mathcal{M}^c = (\mathcal{W}^c, \{\mathcal{B}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{K}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^{d^c}\}_{i\in\mathcal{N}}, V^c)$ pour KBE.

Lemme 3.1 : *Soient* $i, j \in \mathcal{N}$ *et* $\phi \in \mathcal{L}_{KBE}$ *une formule,*

1. $\forall w \in \mathcal{W}^c : \neg K_i\phi \in w \Rightarrow \mathcal{K}_i^*(w) \cup \{\neg\phi\}$ *est KBE-consistant*
2. $\forall w \in \mathcal{W}^c : \neg B_i\phi \in w \Rightarrow \mathcal{B}_i^*(w) \cup \{\neg\phi\}$ *est KBE-consistant*

3. $\forall w \in \mathcal{W}^c : \neg E_i\phi \in w \Rightarrow \mathcal{E}_i^*(w) \cup \{\neg\phi\}$ *est KBE-consistant*

Démonstration. Soient $w \in \mathcal{W}^c$, $i, j \in \mathcal{N}$, $\mathcal{R} \in \{\mathcal{K}_i^c, \mathcal{B}_i^c\}$, $\Box \in \{K_i, B_i, E_i\}$ et $\phi \in \mathcal{L}_{KBE}$ une formule. Raisonnons par contraposition et supposons que $\mathcal{R}^*(w) \cup \{\neg\phi\}$ est KBE-inconsistant. Nous avons donc qu'il existe $n \in \mathbb{N}$ et $\psi_1, \ldots, \psi_n \in \mathcal{R}^*(w)$ tels que :

1. $\vdash \neg(\bigwedge_{k=1}^n \psi_k \wedge \neg\phi)$
2. $\vdash \neg \bigwedge_{k=1}^n \psi_k \vee \neg\neg\phi$
3. $\vdash \bigwedge_{k=1}^n \psi_k \Rightarrow \phi$
4. $\vdash \Box(\bigwedge_{k=1}^n \psi_k \Rightarrow \phi)$
5. $\vdash (\Box \bigwedge_{k=1}^n \psi_k \Rightarrow \Box\phi)$
6. $\vdash (\bigwedge_{k=1}^n \Box\psi_k \Rightarrow \Box\phi)$
7. $\vdash \neg(\bigwedge_{k=1}^n \Box\psi_k \wedge \neg\Box\phi)$

Donc $\{\Box\psi_1, \ldots, \Box\psi_n, \neg\Box\phi\}$ est KBE-inconsistant. Or $\forall k \in \{1, \ldots, n\}$, $\psi_k \in \mathcal{R}^*(w)$ si, et seulement si, $\Box\psi_k \in w$ et w est maximal KBE-consistant. Ainsi $\bigwedge_{k=1}^n \Box\psi_k \in w$ (MC3') et donc $\{\Box\psi_1, ..., \Box\psi_n\}$ est S-consistant. Mais $\{\Box\psi_1, ..., \Box\psi_n\} \cup \{\neg\Box\phi\}$ est KBE-inconsistant. Donc $\neg\Box\phi$ ne peut appartenir à l'ensemble de formules maximal KBE-consistant w. En effet, par l'absurde, si nous avions $\neg\Box\phi \in w$, nous aurions aussi que $\bigwedge_{k=1}^n \Box\psi_k \wedge \neg\Box\phi \in w$ (par MC3'), et $\{\Box\psi_1, ..., \Box\psi_n, \neg\Box\phi\}$ serait KBE-consistant, ce qui est une contradiction. Ainsi, $\neg\Box\phi \notin w$.

Nous avons donc montré que si $\neg\Box\phi \in w$ alors $\mathcal{R}^*(w) \cup \{\neg\phi\}$ est KBE-consistant.

□

Nous avons besoin d'un second lemme pour démontrer la complétude de notre système. Ce lemme montre que toute formule valide du modèle canonique est une formule de l'ensemble maximal KBE-consistant correspondant au monde dans lequel elle est vérifiée et réciproquement.

Lemme 3.2 : *Soit* $\phi \in \mathcal{L}_{KBE}$ *une formule,*

$$\mathcal{M}^c \models \phi \textit{ si, et seulement si, } \forall w \in \mathcal{W}^c, \phi \in w$$

Démonstration. Raisonnons par récurrence sur le degré d'une formule.

(Initialisation) Si $\phi \in \mathcal{L}_{KBE}$ est une formule de degré 0, i.e. il existe $p \in \mathcal{P}, \phi = p$. Par définition du modèle canonique, nous avons donc : $\forall w \in \mathcal{W}^c, w \in V(p)$ si, et seulement si, $p \in w$.

(Hérédité) Supposons que pour toute formule $\phi \in \mathcal{L}_{KBE}$ de degré strictement inférieur à $n \in \mathbb{N}^*$, nous avons $\forall w \in \mathcal{W}^c, \mathcal{M}^c, w \models \phi$ si, et seulement si, $\forall w \in \mathcal{W}^c, \phi \in w$.

Soient $\psi, \theta \in \mathcal{L}_{KBE}$ telles que $max(deg(\psi), deg(\theta)) = n - 1$. Nous avons donc pour tout monde $w \in \mathcal{W}^c$, $\mathcal{M}^c, w \models \psi$ ssi $\psi \in w$ et $\mathcal{M}^c, w \models \theta$ ssi $\theta \in w$.

De plus, nous avons que $\mathcal{M}^c, w \models \neg\psi$ ssi $\mathcal{M}^c, w \not\models \psi$ ssi $\psi \notin w$. Puis $\mathcal{M}^c, w \models \psi \wedge \theta$ ssi $\mathcal{M}^c, w \models \psi$ et $\mathcal{M}^c, w \models \theta$ ssi $\psi \in w$ et $\theta \in w$ ssi $\psi \wedge \theta \in w$ (MC3'). Ensuite $\mathcal{M}^c, w \models \psi \vee \theta$ ssi

$\mathcal{M}^c, w \models \psi$ ou $\mathcal{M}^c, w \models \theta$ ssi $\psi \in w$ ou $\theta \in w$ ssi $\psi \vee \theta \in w$ (MC3). Enfin $\mathcal{M}^c, w \models \psi \Rightarrow \theta$ ssi $\mathcal{M}^c, w \models \neg\psi$ ou $\mathcal{M}^c, w \models \theta$ ssi $\psi \notin w$ ou $\theta \in w$ ssi $\psi \Rightarrow \theta \in w$.

Considérons $(\mathcal{R}, \Box) \in \{(\mathcal{B}_i, B_i), (\mathcal{K}_i, K_i), (\mathcal{E}_i, E_i)\}$ et un monde $w \in \mathcal{W}^c$. Démontrons maintenant l'équivalence pour les modalités normales par double implication.

($\Rightarrow$) Supposons par contraposition que $\Box\psi \notin w$ et comme w est maximal KBE-consistant, nous avons $\neg\Box\psi \in w$. Par le lemme précédent, nous avons $\mathcal{R}^*(w) \cup \{\neg\psi\}$ est KBE-consistant et donc, par le théorème de Lindenbaum, il existe $v \in \mathcal{W}^c : \mathcal{R}^*(w) \cup \{\neg\psi\} \subseteq v$ et v est maximal KBE-consistant. Nous avons donc $\neg\psi \in v$ et, par définition de $\mathcal{R}^c$, nous avons $w\mathcal{R}^c v$. Nous avons donc aussi $\psi \notin v$ et, par hypothèse de récurrence $\mathcal{W}^c, v \nvDash \psi$. Donc puisqu'il existe $v \in \mathcal{W}^c : w\mathcal{R}^c v : v \models \neg\psi$, nous avons $\mathcal{M}^c, w \models \neg\Box\psi$, c'est-à-dire $\mathcal{M}^c, w \nvDash \Box\psi$.

($\Leftarrow$) Par contraposition, supposons que $\mathcal{M}^c, w \nvDash \Box\psi$, i.e, $\mathcal{M}^c, w \models \neg\Box\psi$. Donc il existe $v \in \mathcal{W} : w\mathcal{R}^c v, \mathcal{M}^c, v \models \neg\psi$. Donc $\mathcal{M}^c, v \nvDash \phi$ et par hypothèse de récurrence, nous avons alors $\phi \notin v$. Or, puisque $\phi \notin v$, par définition de $\mathcal{R}^c$, nous avons donc $\Box\phi \notin w$.

Montrons désormais que l'équivalence est aussi vérifiée lorsque la formule $E_i^d\phi$ est de degré n. Supposons que nous avons $\mathcal{M}^c \models E_i^d\phi$, c'est-à-dire $\forall w \in \mathcal{W}^c : \mathcal{M}^c, w \models E_i^d\phi$. Autrement dit, cela est équivalent à $\forall w \in \mathcal{W}^c, \{v | v \in \mathcal{W}^c \wedge \mathcal{M}^c, v \models \phi\} \in \mathcal{E}_i^{d^c}(w)$. Par application de l'hypothèse de récurrence, nous avons donc par équivalence que $\forall w \in \mathcal{W}^c, \{v | v \in \mathcal{W}^c \wedge \phi \in v\} \in \mathcal{E}_i^{d^c}(w)$, c'est-à-dire $\forall w \in \mathcal{W}^c, ||\phi|| \in \mathcal{E}_i^{d^c}(w)$ avec $||\phi|| := \{v | v \in \mathcal{W}^c \wedge \phi \in v\}$. Enfin, par définition du modèle canonique, cela est équivalent à $E_i^d\phi \in w$.

(Conclusion) Nous avons donc montré par récurrence que :

$$\forall \phi \in \mathcal{L}_{KBE}, \forall w \in \mathcal{W}^c : \mathcal{M}^c, w \models \phi \text{ si, et seulement si, } \phi \in w$$

□

Maintenant que le lien entre la validité et les ensembles maximaux KBE-consistants a été démontré, nous pouvons prouver le lien entre notre modèle canonique et les formules prouvées dans notre système axiomatique.

Théorème 3.1 - *Lemme de vérité :* *Soit $\phi \in \mathcal{L}_{KBE}$ une formule,*

$$\mathcal{M}^c \models \phi \textit{ si, et seulement si, } \vdash \phi$$

Démonstration. Soit $\phi \in \mathcal{L}_{KBE}$ une formule. $\mathcal{M}^c \models \phi$ si, et seulement si, $\forall w \in \mathcal{W}^c : \mathcal{M}^c, w \models \phi$ si, et seulement si, par (lemme 3.2), $\forall w \in \mathcal{W}^c, \phi \in w$ si, et seulement si, par (MC5), $\vdash \phi$.

□

Preuve de complétude du système KBE

Nous pouvons désormais prouver que le système KBE est complet.

Théorème 3.2 - : *Le système KBE est complet.*

Démonstration. Soit un modèle canonique $\mathcal{M}^c = (\mathcal{W}^c, \{\mathcal{B}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{K}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^{d^c}\}_{i\in\mathcal{N}}, V^c)$.

(1) Montrons tout d'abord que :

$$\forall w \in \mathcal{W}^c : X \in \mathcal{E}_i^{d^c}(w) \wedge Y \in \mathcal{E}_i^{d^c}(w) \Longrightarrow X \cap Y \in \mathcal{E}_i^{d^c}(w)$$

Soient $w \in \mathcal{W}^c$, $X \in \mathcal{E}_i^c(w)$ et $Y \in \mathcal{E}_i^{d^c}(w)$. Par définition du modèle canonique minimal, il existe donc $\phi, \psi \in \mathcal{L}_{KBE}$ telles que $X = ||\phi||$ et $Y = ||\psi||$. Ainsi, $||\phi|| \in \mathcal{E}_i^{d^c}(w)$ et $||\psi|| \in \mathcal{E}_i^{d^c}(w)$. Ensuite, par définition, nous avons alors $E_i^d\phi \in w$ et $E_i^d\psi \in w$. De plus, par (MC3') nous avons alors $E_i^d\phi \wedge E_i^d\psi \in w$. Or $\vdash E_i^d\phi \wedge E_i^d\psi \Rightarrow E_i^d(\phi \wedge \psi)$. Ainsi, par (MC5) $E_i^d\phi \wedge E_i^d\psi \Rightarrow E_i^d(\phi \wedge \psi) \in w$ et par (MC4), nous déduisons alors que $E_i^d(\phi \wedge \psi) \in w$. Donc $||\phi \wedge \psi|| \in \mathcal{E}_i^{d^c}(w)$. Or $||\phi \wedge \psi|| = ||\phi|| \cap ||\psi|| = X \cap Y$. Par conséquent, nous avons alors prouvé que $X \cap Y \in \mathcal{E}_i^{d^c}(w)$.

(2) Montrons que :

$$\forall w \in \mathcal{W}^c : \mathcal{E}_i^{d^c}(w) \subseteq \bigcap_{v\in\mathcal{W}:w\mathcal{K}_i^c v} \mathcal{E}_i^{d^c}(v)$$

Soient $w \in \mathcal{W}^c$ et $X \in \mathcal{E}_i^{d^c}(w)$. Par définition de $\mathcal{E}_i^{d^c}$, il existe $\phi \in \mathcal{L}_{KBE}$, $X = ||\phi||$. Donc $E_i^d\phi \in w$. Or $\vdash E_i^d\phi \Rightarrow K_iE_i^d\phi$ et, par (MC5), nous avons $E_i^d\phi \Rightarrow K_iE_i^d\phi \in w$, puis, par (MC4) $K_iE_i^d\phi \in w$. Donc, par définition du modèle canonique, nous avons pour tout $v \in \mathcal{W}^c, w\mathcal{K}_i^c v$: $\mathcal{M}^c, v \models E_i^d\phi$. Ainsi, $E_i^d\phi \in v$, c'est-à-dire, $X \in \mathcal{E}_i^{d^c}(v)$. Nous venons donc de montrer que $\forall v \in \mathcal{W}^c, w\mathcal{K}_i^c v : X \in \mathcal{E}_i^{d^c}(v)$, i.e. $X \in \bigcap_{v\in\mathcal{W}:w\mathcal{K}_i^c v} \mathcal{E}_i^{d^c}(v)$.

(3) Montrons que :

$$\forall w, v \in \mathcal{W}^c, \forall X \in 2^{\mathcal{W}} : X \notin \mathcal{E}_i^{d^c}(w) \Longrightarrow (w\mathcal{K}_i^c v \Rightarrow X \notin \mathcal{E}_i^{d^c}(v))$$

Soient $w, v \in \mathcal{W}^c$ tels que $w\mathcal{K}_i^c v$ et $X \notin \mathcal{E}_i^{d^c}(w)$. Par définition de $\mathcal{E}_i^{d^c}(w) = \{||\phi|| : E_i^d\phi \in w\}$, puisque $X \notin \mathcal{E}_i^{d^c}(w)$, il n'existe pas $\phi \in \mathcal{L}_{KBE}$, $X = ||\phi||$ et $E_i^d\phi \in w$. Cela signifie donc que, pour tout $\phi \in \mathcal{L}_{KBE}$, $X = ||\phi|| \Rightarrow E_i^d\phi \notin w$, c'est-à-dire, par (MC2), nous avons pour tout $\phi \in \mathcal{L}_{KBE}$, $X = ||\phi|| \Rightarrow \neg E_i^d\phi \in w$.

Il existe deux cas possibles.

— Soit X s'écrit sous la forme $X = ||\phi||$ et donc $\neg E_i^d\phi \in w$. Or $\vdash \neg E_i^d\phi \Rightarrow K_i\neg E_i^d\phi$ et, par (MC5), $\neg E_i^d\phi \Rightarrow K_i\neg E_i^d\phi \in w$, puis, par (MC4), $K_i\neg E_i^d\phi \in w$. Ainsi, $\forall u \in \mathcal{W}^c$: $w\mathcal{K}_i u, \neg E_i^d\phi \in u$. Enfin, $\forall u \in \mathcal{W}^c : w\mathcal{K}_i u, ||\phi|| \notin \mathcal{E}_i^{d^c}(u)$, et donc $X \notin \mathcal{E}_i^{d^c}(v)$;

— Soit X ne s'écrit pas sous la forme $X = ||\phi||$ et, par simple définition du modèle canonique minimal, puisque $\mathcal{E}_i^{d^c}(w) = \{||\phi|| : E_i^d\phi \in w\}$, nous avons $\forall u \in \mathcal{W}^c : X \notin \mathcal{E}_i^{d^c}(u)$. Par conséquent, $X \notin \mathcal{E}_i^{d^c}(v)$.

(4) Montrons que :

$$\forall w \in \mathcal{W}^c : \mathcal{W}^c \notin \mathcal{E}_i^{d^c}(w)$$

Soit $w \in \mathcal{W}^c$. Puisque $\vdash \neg E_i^d\top$, nous avons par (MC5) que $\neg E_i^d\top \in w$, c'est-à-dire, $||\top|| \notin \mathcal{E}_i^{d^c}(w)$. Or puisque $||\top|| = \mathcal{W}^c$, nous déduisons immédiatement que $\mathcal{W}^c \notin \mathcal{E}_i^{d^c}(w)$.

(5) Montrons que :

$$\forall w \in \mathcal{W} : X \in \mathcal{E}_i^{d^c}(w) \Longrightarrow (\mathcal{E}_i(w) \subseteq X)$$

Soient $w \in \mathcal{W}^c$ et $X \in \mathcal{E}_i^{d^c}(w)$. Par définition de $\mathcal{E}_i^{d^c}(w) = \{||\phi|| : E_i\phi \in w\}$, il existe $\phi \in \mathcal{L}_{KBE}$ tel que $X = ||\phi||$. Donc, $E_i^d\phi \in w$. Or $\vdash E_i^d\phi \Rightarrow E_i\phi$ et par (MC5), $E_i^d\phi \Rightarrow E_i\phi \in w$, puis par (MC4), $E_i\phi \in w$, i.e. $\mathcal{M}^c, w \models E_i\phi$. Ainsi, $\forall u \in \mathcal{W}^c : u \in \mathcal{E}_i^c(w), \mathcal{M}^c, u \models \phi$. Or $X = ||\phi||$ et donc $\forall u \in \mathcal{E}_i^c(w), u \in ||\phi||$. Par conséquent, nous venons de prouver que comme $v \in \mathcal{E}_i^c(w)$, nous avons $v \in ||\phi||$, i.e. $v \in X$ et donc $\mathcal{E}_i(w) \subseteq X$.

(6) Les autres propriétés du modèle canonique à démontrer concernant les relations $\mathcal{K}_i$, $\mathcal{B}_i$ et $\mathcal{E}_i$ sont classiques (Blackburn et al., 2002).

Par conséquent, pour toute formule ϕ valide dans notre cadre $\mathcal{C}$, ϕ est valide quelque soit le modèle $\mathcal{M}$ de $\mathcal{C}$. Donc ϕ est valide dans tout modèle canonique $\mathcal{M}^c$. Par conséquent, $\vdash \phi$. Ainsi, nous venons de prouver que le système KBE est complet.

□

3.2.3 Propriétés fortes du cadre KBE

Dans cette section, nous démontrons quelques propriétés du cadre KBE comme les théorèmes de déduction, la correction forte et la complétude forte du cadre.

Théorèmes de déduction

Tout d'abord, la KBE-déductibilité possède les propriétés de réflexivité, de transitivité et d'affaiblissement à gauche présentées en section 2.1.1. Ces propriétés sont évidentes à démontrer dans KBE. Il est de même standard de démontrer que les théorèmes de déduction sont vérifiés.

Théorème 3.3 - *Théorèmes de déduction dans KBE* : *Soient* Γ *un ensemble de formules prouvées dans KBE,* ϕ *et* ψ *deux théorèmes.*

$$(1) \quad \Gamma \cup \{\psi\} \vdash \phi \text{ si, et seulement si, } \Gamma \vdash \psi \Rightarrow \phi$$

$$(2) \quad \Gamma \cup \{\psi\} \models \phi \text{ si, et seulement si, } \Gamma \models \psi \Rightarrow \phi$$

Démonstration. Soient Γ un ensemble de formules prouvées dans KBE, ϕ et ψ deux théorèmes. Démontrons le sens ($\Rightarrow$) et supposons donc que $\Gamma \cup \{\psi\} \vdash \phi$. Par définition de la KBE-déductibilité, nous avons donc qu'il existe $\Sigma = \{\psi_1, \ldots, \psi_n\}, \Sigma \subseteq \Gamma \cup \{\psi\}$ tel que :

$$\vdash \bigwedge_{i \in \{1,\ldots,n\}} \psi_i \Rightarrow \phi$$

Nous avons donc deux cas à considérer lorsque $\psi \in \Sigma$ et lorsque $\psi \notin \Sigma$.

Si $\psi \in \Sigma$, alors il existe $i \in \{1, \ldots, n\}$ tel que $\psi = \psi_i$. Donc $\vdash (\bigwedge_{k \in \{1,\ldots,n\} \setminus \{i\}} \psi_k) \wedge \psi \Rightarrow \phi$ par commutativité et associativité du $\wedge$. Puis, $\vdash \bigwedge_{k \in \{1,\ldots,n\} \setminus \{i\}} \psi_k \Rightarrow (\psi \Rightarrow \phi)$. Or pour tout $k \in \{1, \ldots, n\} \setminus \{i\}, \psi_k \in \Sigma$ et donc $\psi_k \in \Gamma$ par inclusion. Nous avons donc prouvé dans ce cas que $\Gamma \vdash \psi \Rightarrow \phi$.

Si $\psi \notin \Sigma$, alors pour tout $i \in \{1, \ldots, n\}, \psi_i \neq \psi$. Nous avons donc dans ce cas que :

$$\vdash \bigwedge_{k \in \{1,\ldots,n\}} \psi_k \Rightarrow \phi$$

Or comme $\vdash \phi \Rightarrow (\psi \Rightarrow \phi)$ est un axiome du CP, nous déduisons immédiatement que :

$$\vdash \bigwedge_{k \in \{1,\ldots,n\}} \psi_k \Rightarrow (\psi \Rightarrow \phi)$$

Par conséquent, puisque pour tout $k \in \{1, \ldots, n\}, \psi_k \in \Sigma$ et donc $\psi_k \in \Gamma$ par inclusion. Nous avons donc prouvé dans cet autre cas que $\Gamma \vdash \psi \Rightarrow \phi$.

($\Leftarrow$) Prouvons désormais la réciproque et supposons que $\Gamma \vdash \psi \Rightarrow \phi$.
Il existe donc $\Sigma = \{\psi_1, \ldots, \psi_n\}$, $\Sigma \subseteq \Gamma$ tel que :

$$\vdash \bigwedge_{i \in \{1,\ldots,n\}} \psi_i \Rightarrow (\psi \Rightarrow \phi)$$

Donc $\vdash \bigwedge_{i \in \{1,\ldots,n\}} \psi_i \wedge \psi \Rightarrow \phi$ est aussi un théorème. Ainsi, nous déduisons que $\Sigma \cup \{\psi\} \vdash \phi$. Or comme $\Sigma \subseteq \Gamma$, par affaiblissement à gauche, nous avons immédiatement que $\Gamma \cup \{\psi\} \vdash \phi$.

La version sémantique (2) du théorème de déduction découle immédiatement du fait que KBE est un système correct et complet. Il suffit alors d'appliquer la définition de la conséquence sémantique, puis la complétude et se ramener à la forme du théorème de déduction qui vient d'être démontrée et de repasser par la correction sur la forme sémantique. Sinon une autre façon de le démontrer consiste à faire le raisonnement par double équivalence suivant :

$\Gamma \models \psi \Rightarrow \phi$ ssi $\forall \mathcal{M}$: si $\mathcal{M} \models \Gamma$, alors $\mathcal{M} \models \psi \Rightarrow \phi$ ssi $\forall \mathcal{M} : \mathcal{M} \models \Gamma \Rightarrow (\psi \Rightarrow \phi)$ ssi $\forall \mathcal{M} : \mathcal{M} \models \neg\Gamma \vee \neg\psi \vee \phi$ ssi $\forall \mathcal{M} : \mathcal{M} \models \neg(\Gamma \wedge \psi) \vee \phi$ ssi $\forall \mathcal{M} : \mathcal{M} \models (\Gamma \wedge \psi) \Rightarrow \phi$ ssi $\forall \mathcal{M}$: si $\mathcal{M} \models \Gamma \cup \{\psi\}$, alors $\mathcal{M} \models \phi$ ssi $\Gamma \cup \{\psi\} \models \phi$.

□

Correction forte

Dans la suite, nous démontrons la correction forte de KBE. La correction forte est une conséquence quasiment immédiate de la correction et de l'affaiblissement sémantique que nous redémontrons dans la preuve du théorème.

Théorème 3.4 - *Correction forte de KBE* : *Soient* Γ *un ensemble de formules prouvées dans KBE,* ϕ *un théorème de KBE.*

$$Si\ \Gamma \vdash \phi\ alors\ \Gamma \models \phi$$

Démonstration. Soient Γ un ensemble de formules prouvées dans KBE, ϕ un théorème de KBE.

Supposons que $\Gamma \vdash \phi$, donc il existe $\psi_1, \ldots, \psi_n \in \Gamma$ tq $\vdash \psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \phi$. Donc par correction, nous avons $\models \psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \phi$ i.e. $\models \neg\psi_1 \vee \ldots \vee \neg\psi_n \vee \phi$. Donc $\models \neg\psi_1 \vee \ldots \vee \neg\psi_n \vee \neg \bigwedge_{\theta \in \Gamma} \theta \vee \phi$. Ainsi, en appliquant la loi de De Morgan et en regroupant les termes, nous avons que :

$$\models \neg \bigwedge_{\theta \in \Gamma \cup \{\psi_1, \ldots, \psi_n\}} \theta \vee \phi$$

Or comme $\{\psi_1, \ldots, \psi_n\} \subseteq \Gamma$, nous avons $\Gamma \cup \{\psi_1, \ldots, \psi_n\} = \Gamma$ et donc $\models \neg \bigwedge_{\theta \in \Gamma} \theta \vee \phi$ (affaiblissement sémantique). Par conséquent, nous avons $\forall \mathcal{M}$, si $\mathcal{M} \models \Gamma$, alors $\mathcal{M} \models \phi$. Nous venons donc de montrer que $\Gamma \models \phi$.

□

Complétude forte

Le système KBE est fortement complet. Pour démontrer la complétude forte du système, nous avons besoin du lemme 3.3 suivant.

Lemme 3.3 : *Pour tout ensemble* Γ *de formules KBE-consistant, il existe un modèle canonique* $\mathcal{M}^c$ *tel que :* $\mathcal{M}^c \models \Gamma$*, i.e.* $\forall \phi \in \Gamma : \mathcal{M}^c \models \phi$.

Démonstration. Soit Γ un ensemble de formules de KBE-consistant. Par application du lemme de Lindenbaum, il existe un ensemble maximal de formules KBE-consistant Γ' tel que $\Gamma \subseteq \Gamma'$. Considérons un modèle canonique $\mathcal{M}^c = (\mathcal{W}^c, (\mathcal{K}_i^c)_{i \in \mathcal{N}}, (\mathcal{B}_i^c)_{i \in \mathcal{N}}, (\mathcal{E}_i^c)_{i \in \mathcal{N}}, (\mathcal{E}_i^{d^c})_{i \in \mathcal{N}}, V^c)$. Nous avons $\forall \phi \in \Gamma' : \mathcal{M}^c, \Gamma' \models \phi$, et donc $\forall \phi \in \Gamma : \mathcal{M}^c, \Gamma \models \phi$. Nous avons prouvé que pour tout ensemble de formules Γ KBE-consistant, il existe un modèle canonique satisfaisant toutes les formules de Γ. □

Théorème 3.5 - *Complétude forte de KBE* :

Le système KBE est fortement complet, c'est-à-dire pour toute formule $\phi \in \mathcal{L}_{KBE}$ *et tout ensemble de formules* $\Gamma \subseteq \mathcal{L}_{KBE}$*, si* $\Gamma \models \phi$ *alors* $\Gamma \vdash \phi$.

Démonstration. Démontrons la complétude forte par contraposition. Soit $\Gamma \subseteq \mathcal{L}_{KBE}$ un ensemble de formules tel que $\Gamma \nvdash \phi$, nous avons donc que $\Gamma \cup \{\neg\phi\}$ est un ensemble de formules KBE-consistant. En effet, par l'absurde, si nous avions $\Gamma \cup \{\neg\phi\}$ est KBE-inconsistant, il existerait $\psi_1, \ldots, \psi_n \in \Gamma$ telles que : $\vdash \neg(\psi_1 \wedge \ldots \wedge \psi_n \wedge \neg\phi)$ et donc, nous aurions aussi $\vdash (\psi_1 \wedge \ldots \wedge \psi_n) \Rightarrow \phi$.

Cependant, par le théorème de déduction, nous déduirions que $\Gamma \cup \{\psi_1, \ldots, \psi_n\} \vdash \phi$, i.e. $\Gamma \vdash \phi$, ce qui contredit l'hypothèse que $\Gamma \nvdash \phi$. Ainsi, par le lemme 3.3, il existe un modèle $\mathcal{M}$ tel que $\mathcal{M} \models \Gamma \cup \{\neg\phi\}$, c'est-à-dire $\mathcal{M} \models \Gamma$ et $\mathcal{M} \models \neg\phi$. Par conséquent, nous venons de prouver qu'il existe un modèle canonique et donc un modèle $\mathcal{M}$ tel que $\mathcal{M}, \Gamma \not\models \phi$.

□

3.2.4 Quelques théorèmes déductibles dans KBE

Remarquons que l'axiome (D) est valide pour E_i^d. Il signifie qu'un agent i ne peut pas avoir l'intention délibérée de faire quelque chose et son contraire.

Théorème 3.6 - : $\vdash \neg E_i^d \bot \quad (D_{E_i^d})$

Démonstration.

1. $\vdash E_i^d \bot \Rightarrow E_i \bot$
2. $\vdash (E_i^d \bot \Rightarrow E_i \bot) \Rightarrow (\neg E_i \bot \Rightarrow \neg E_i^d \bot)$
3. $\vdash \neg E_i \bot \Rightarrow \neg E_i^d \bot$
4. $\vdash \neg E_i \bot$
5. $\vdash \neg E_i^d \bot$

□

Ce théorème traduit aussi sémantiquement que l'ensemble vide ne peut appartenir à aucun ensemble de voisinages. Par ailleurs, d'autres théorèmes peuvent être déduits dans KBE. En particulier, nous pouvons montrer que tout agent i, lorsqu'il veille à ce qu'un autre agent j croit quelque chose, cet agent i veille aussi à ce que l'agent j ne puisse pas croire que cet agent i puisse savoir le contraire. Un tel théorème traduit qu'un agent lorsqu'il cherche à véhiculer de nouvelles croyances, qu'elles soient fondées ou fausses dans le cas des mensonges, cet agent veille toujours à être crédible vis-à-vis de l'autre agent.

Théorème 3.7 - :

1. *Un agent qui influence de manière délibérée ne peut avoir l'intention de montrer que d'autres agents, y compris lui-même, puissent détenir la vérité, c'est-à-dire :*

$$\vdash E_i^d B_j \phi \Rightarrow \neg E_i^d B_j K_k \neg\phi$$

2. *Un agent qui veille de manière délibérée qu'un agent ne croit pas une information, ne peut pas aussi veiller que ce dernier sache que des agents détiennent la vérité, c'est-à-dire :*

$$\vdash E_i^d \neg B_j \phi \Rightarrow \neg E_i^d B_j K_k \phi$$

Démonstration. Pour démontrer ces théorèmes, nous présentons dans un premier temps les schémas de preuve, puis dans un second temps, nous donnons la preuve à la Hilbert correspondante.

(1) La première étape est de montrer que $\vdash B_j\phi \Rightarrow \neg B_j K_k \neg\phi$. Ce théorème vient directement d'un raisonnement par l'absurde, en supposant $B_j\phi \wedge B_j K_k \neg\phi$, puis, en appliquant l'axiome (T_{K_k}). La deuxième étape consiste à appliquer la nécessitation $\vdash E_i(B_j\phi \Rightarrow \neg B_j K_k \neg\phi)$, puis d'appliquer le modus ponens (MP) sur l'axiome (K_{E_i}). Immédiatement $\vdash E_i B_j\phi \Rightarrow E_i \neg B_j K_k \neg\phi$. Ensuite, avec $(E_i^d E_i) + (MP)$, nous obtenons alors $\vdash E_i^d B_j\phi \Rightarrow E_i \neg B_j K_k \neg\phi$. Enfin, en appliquant le modus ponens sur (D_{E_i}) et la contraposition de $(E_i^d E_i)$, nous obtenons $\vdash E_i^d B_j\phi \Rightarrow \neg E_i^d B_j K_k \neg\phi$.

La preuve avec un système à la Hilbert est donnée en annexe A.

(2) Tout d'abord avec (Nec_{B_j}) sur (T_{K_k}), nous avons $\vdash B_j K_k\phi \Rightarrow B_j\phi$. Ensuite, en appliquant (Nec_{E_i}) sur la contraposition de ce théorème, nous obtenons $\vdash E_i \neg B_j\phi \Rightarrow E_i \neg B_j K_k\phi$. Ensuite, nous avons $\vdash E_i^d \neg B_j\phi \Rightarrow E_i \neg B_j K_k\phi$ avec $(E_i^d E_i) + (MP)$. Enfin, par un raisonnement analogue au précédent, par application du modus ponens sur l'axiome (D_{E_i}) et la contraposition de l'axiome $(E_i^d E_i)$, nous prouvons alors $\vdash E_i^d \neg B_j\phi \Rightarrow \neg E_i^d B_j K_k\phi$.

La preuve avec un système à la Hilbert est donnée en annexe A.

□

En remarquant que la contraposition de $K_i\phi \Rightarrow B_i\phi$ est $\neg B_i\phi \Rightarrow \neg K_i\phi$, nous pouvons déduire deux corollaires immédiats à ces deux théorèmes :

1. $\vdash E_i B_j\phi \Rightarrow E_i \neg K_j K_i \neg\phi$
2. $\vdash E_i \neg B_j\phi \Rightarrow E_i \neg K_j K_i\phi$

Ces deux théorème nous apprennent qu'un agent qui veille à faire croire quelque chose à quelqu'un, veille aussi à ce qu'il ne sache pas que l'agent qui influence les croyances puisse savoir le contraire. Il en est de même lorsqu'un agent veille à empêcher un autre agent de croire, une conséquence de ses intentions est aussi qu'il l'en empêche de savoir que lui sait.

Par ailleurs, dans le système KBE, nous pouvons aussi déduire le principe *qui facit per alium facit per se*, c'est-à-dire ≪celui qui agit à travers un autre fait acte lui-même≫. Si la manipulation est l'intention délibérée d'instrumentaliser un autre agent sans qu'il ne s'en rende compte, cela signifie qu'un agent manipulant un autre agent à agir de façon illégale fait lui aussi acte lui-même. Un agent manipulateur est donc lui aussi responsable d'actes illégaux perpétrés par un agent manipulé. Le manipulateur, par ce principe, a donc une part de responsabilité dans ces actes commis.

Théorème 3.8 - *Qui facit per alium facit per se* :

$$\vdash (E_i^d E_j\phi \vee E_i^d E_j^d\phi) \Rightarrow E_i\phi$$

Démonstration. Pour la partie gauche de la disjonction avec $(E_i^d E_i)$ et (Nec_{E_i}) sur (T_{E_j}), nous avons immédiatement $\vdash E_i^d E_j\phi \Rightarrow E_i\phi$. Pour la partie droite de la disjonction avec $(E_i^d E_i)$, nous

avons $\vdash E_i^d E_j \phi \Rightarrow E_i E_j \phi$, puis avec (Nec_{E_i}) sur (T_{E_j}), $\vdash E_i^d E_j^d \phi \Rightarrow E_i \phi$. Par conséquent, par élimination de la disjonction, nous montrons que $\vdash (E_i^d E_j \phi \vee E_i^d E_j^d E_j^d \phi) \Rightarrow E_i \phi$.

La preuve avec un système à la Hilbert est donnée en annexe A.

□

3.3 Une théorie de la manipulation

Nous avons vu au chapitre 1 que la manipulation est caractérisée comme l'intention délibérée d'un manipulateur d'instrumentaliser un agent victime, tout veillant à lui dissimuler cette intention. Un agent manipulé peut satisfaire à cette influence soit de façon délibérée, ou soit de façon non délibérée. Pourtant dans les deux cas, nous pouvons parler de manipulation. Par ailleurs, un agent manipulateur peut aussi instrumentaliser un autre agent afin de l'en empêcher de faire quelque chose. Dans ce dernier cas, nous parlons aussi de manipulation. De plus, un agent peut dissimuler ses intentions soit en veillant à ce que l'autre agent ne sache pas ses intentions ou en veillant à ce qu'il ne puisse pas croire que ses intentions de l'instrumentaliser. Dans ces deux derniers cas, nous pouvons encore parler de manipulation. Par conséquent, pour pouvoir définir formellement la manipulation dans KBE, il est nécessaire de séparer la manipulation en huit cas. Ces cas représentent différentes formes que la manipulation peut avoir. Ainsi, une manipulation est dite *forte* si l'agent manipulé veille de façon délibérée à réaliser les intentions du manipulateur, sinon la manipulation est dite *douce*. Nous parlons de *manipulation destructive* si l'agent manipulateur veille à empêcher l'agent manipulé d'agir, sinon la manipulation est dite *constructive*. De plus, suivant la manière dont l'agent dissimule ses intentions, nous parlons de *manipulation avec dissimulation épistémique* lorsque l'agent manipulateur veille à ce que l'agent manipulé ne sache pas et de *manipulation avec dissimulation doxastique* lorsque l'agent manipulé ne croit pas que le manipulateur l'instrumentalise. Enfin, après avoir donné une définition formelle de ces différentes formes de manipulation dans KBE, nous donnons une définition générale de manipulation et nous illustrons en quoi « la manipulation n'est pas exactement de la coercition, ni de la persuasion et n'est pas entièrement similaire à la tromperie » (Handelman, 2009). Nous donnons alors une définition formelle dans KBE de la coercition, la tromperie et de la persuasion.

3.3.1 Différentes formes d'instrumentalisation et de dissimulation

La manipulation est représentée comme la combinaison d'une instrumentalisation et d'une dissimulation. La figure 3.1 et la figure 3.2 présentent les différentes façons d'exprimer l'instrumentalisation et la dissimulation dans les cas de manipulations constructives et les cas de manipulations destructives.

La figure 3.1 montre les différentes composantes d'une manipulation constructive. Par exemple, l'instrumentalisation forte est représentée par la formule $E_i^d E_j^d \phi$. Littéralement, cette formule décrit que l'agent i emploie une stratégie amenant l'agent j a effectué un ensemble d'actions qui mènent à rendre vrai la conséquence ϕ. Ensuite, l'instrumentalisation douce peut être représenté par la formule $E_i^d E_j \phi$. Enfin, dans le cas des manipulations constructives, la dissimulation épis-

Instrumentalisation	**Dissimulation**
Forte ($E_i^d E_j^d \phi$)	Épistémique ($E_i^d \neg K_j E_i^d E_j^d \phi$)
Douce ($E_i^d E_j \phi$)	Doxastique ($E_i^d \neg B_j E_i^d E_j \phi$)

Tableau 3.1 – Formes constructives de manipulation

témique peut être représentée par la formule $E_i^d \neg K_j E_i^d E_j \phi$ et la dissimulation doxastique par la formule $E_i^d \neg B_j E_i^d E_j \phi$.

Instrumentalisation	**Dissimulation**
Forte ($E_i^d \neg E_j^d \phi$)	Épistémique ($E_i^d \neg K_j E_i^d \neg E_j^d \phi$)
Douce ($E_i^d \neg E_j \phi$)	Doxastique ($E_i^d \neg B_j E_i^d \neg E_j \phi$)

Tableau 3.2 – Formes destructives de manipulation

La figure 3.2 décrit les différentes composantes lorsque la manipulation est destructive. Par exemple, dans ces cas de manipulations destructives, l'instrumentalisation douce y est représentée par la formule $E_i^d \neg E_j \phi$, la forte y est représentée par la formule $E_i^d \neg E_j^d \phi$, puis la dissimulation épistémique par la formule $E_i^d \neg K_j E_i^d \neg E_j^d \phi$ et enfin, la dissimulation doxastique y est représentée par la formule $E_i^d \neg B_j E_i^d \neg E_j^d \phi$.

Dans la suite, nous combinons ces différentes formes d'instrumentalisation et de dissimulation pour définir toutes les formes de manipulation pouvant être exprimées dans KBE.

3.3.2 Manipulation constructive, destructive, forte et douce

Il existe donc deux manières d'influencer un agent : soit en le poussant à agir de façon délibérée ou non, soit en l'empêchant d'agir. Ainsi, nous parlons de *manipulation constructive douce* lorsque l'agent manipulateur amène sa victime à réaliser de façon délibérée ou non les intentions du manipulateur. Par ailleurs, la manipulation n'est pas seulement une stratégie d'influencer, elle est aussi une intention de dissimuler son intention. Cette dissimulation dans la manipulation peut donc être caractérisée soit comme le fait de ne pas savoir les stratégies du manipulateur (dissimulation épistémique), soit de ne pas croire (dissimulation doxastique). Ainsi, nous définissons donc les manipulations constructives douces de deux façons, suivant que la dissimulation soit épistémique ou doxastique.

Définition 3.2 - ***Manipulations constructives douces :*** *Soient Σ un ensemble fini et fermé de formules tel que $\top \in \Sigma$ et $\phi \in \Sigma$ une formule. Nous appelons* manipulation constructive douce avec dissimulation épistémique *la formule $MCEK_{i,j}^{\Sigma}\phi$ caractérisée par le prédicat :*

$$MCEK_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(E_j\phi \wedge \neg K_j E_i^d E_j \phi \wedge \psi)$$

Nous appelons manipulation constructive douce avec dissimulation doxastique *la formule* $MCEB^{\Sigma}_{i,j}\phi$ *caractérisée par le prédicat :*

$$MCEB^{\Sigma}_{i,j}\phi = \bigvee_{\psi\in\Sigma} E^d_i(E_j\phi \wedge \neg B_j E^d_i E_j\phi \wedge \psi)$$

Cette formule décrit qu'il y a manipulation lorsqu'un agent influence les intentions de manière délibérée d'un autre agent, tout en veillant de façon délibérée à lui dissimuler cette stratégie de l'influencer. En effet, si $E^d_i E_j\phi \wedge E^d_i \neg K_j E^d_i E_j\phi$ est vraie, puisque $\vdash E^d_i\phi \wedge E^d_i\psi \Rightarrow E^d_i(\phi\wedge\psi) \quad (C)$, nous avons alors que $E^d_i(E_j\phi \wedge \neg K_j E^d_i E_j\phi)$, puis par (RE), $E^d_i(E_j\phi \wedge \neg K_j E^d_i E_j\phi \wedge \top)$ et donc $MCEK^{\Sigma}_{i,j}\phi$. De plus, remarquons que puisque nous modélisons la manipulation avec une modalité non normale (E^d_i) qui ne possède pas la réciproque au théorème (C), il est nécessaire de considérer toutes les autres formules qui peuvent être contenues dans la stratégie courante de l'agent i, qu'elles aient un lien ou non avec la stratégie de manipulation. Le système KBE ne permet pas de considérer le théorème $E^d_i(\phi\wedge\psi) \equiv E^d_i\phi \wedge E^d_i\psi$ et si $E^d_i\phi$ est une manipulation, nous devons aussi considérer $E^d_i(\phi\wedge\psi)$ comme de la manipulation. Ainsi l'ensemble de formules Σ représente toutes les formules sur lesquelles les agents peuvent raisonner. Les ψ considérés sont donc d'une part, les formules de Σ qui ne contredisent pas $E_j\phi\wedge\neg K_j E^d_i E_j\phi$. En effet, si ψ contredit $E_j\phi\wedge\neg K_j E^d_i E_j\phi$, alors nous déduisons immédiatement que $E^d_i\bot$. Or, cela est impossible en raison du théorème $\vdash \neg E^d_i\bot$. D'autre part, puisque nous ne pouvons pas décider de la manipulation sur des ensembles infinis de formules du langage $\mathcal{L}_{KBE}$, nous devons restreindre la manipulation à un ensemble fini Σ. Cet ensemble Σ peut être vu comme un ensemble de formules pour lesquelles les agents ont conscience. Cet ensemble trouve donc son analogie avec la fonction de correspondance de conscience $\mathcal{A}_i : \mathcal{W} \to 2^{\mathcal{L}}$ telle qu'elle est introduite dans (Schipper, 2014) et présentée dans la section 2.3.3. Ici, nous pourrions supposer que tous les agents ont conscience de toutes les formules de Σ et donc pour tout monde $w \in \mathcal{W}$, pour tout $i \in \mathcal{N}$, nous aurions $\mathcal{A}_i(w) = \Sigma$. Un point remarquable de cette définition est que nous pouvons prouver l'existence d'une manipulation que par rapport à ce que nous avons conscience. Ainsi, même si sur un ensemble fermé de formules Σ, nous montrons qu'il n'est pas le cas qu'un agent i manipule un autre agent j. Nous n'en sommes jamais certain, sauf si nous supposons qu'aucun agent ne considère des formules qui ne sont pas dans cet ensemble Σ.

Exemple 3.2 : *Illustrons ce prédicat avec un exemple lié à la publicité. Un publicitaire a toujours pour intentions d'influencer de nouveaux clients pour acheter un produit. Ces intentions ne sont pas dissimulées et de ce fait, nous ne pouvons parler de manipulation. En revanche, cela devient de la manipulation lorsque le publicitaire utilise une technique de vente composée d'une ou plusieurs actions, qu'il cherche à dissimuler aux futurs acheteurs, comme par exemple, l'utilisation d'images subliminales. Ainsi, le publicitaire ne cherche pas à dissimuler son intention de faire que le client achète le produit mais à dissimuler de façon délibérée la technique qu'il utilise pour inciter le client à acheter.*

Si l'agent i est le publicitaire, E^d_i représente sa stratégie d'utiliser les images subliminales afin d'amener le client j à acheter un produit ϕ. Le client ne sait pas que le publicitaire a employé ces

images. Ainsi, cette situation de manipulation est complètement décrite par la formule $E_i^d(E_j\phi \wedge \neg K_j E_i^d E_j\phi)$.

Définition 3.3 - *Manipulations constructives fortes* : *Soient* Σ *un ensemble fini et fermé de formules tel que* $\top \in \Sigma$ *et* $\phi \in \Sigma$ *une formule. Nous appelons* manipulation constructive forte avec dissimulation épistémique *la formule* $MCE^d K_{i,j}^{\Sigma}\phi$ *caractérisée par le prédicat :*

$$MCE^d K_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(E_j^d\phi \wedge \neg K_j E_i^d E_j^d\phi \wedge \psi)$$

Nous appelons manipulation constructive forte avec dissimulation doxastique *la formule* $MCE^d B_{i,j}^{\Sigma}\phi$ *caractérisée par le prédicat :*

$$MCE^d B_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(E_j^d\phi \wedge \neg B_j E_i^d E_j^d\phi \wedge \psi)$$

Cette manipulation représente une influence dissimulée du manipulateur sur des choix stratégiques qu'un autre agent peut faire. Lorsqu'un agent pousse un autre agent à voter pour une certaine option ϕ (i.e $E_i^d E_j^d\phi$) tout en dissimulant son intention de l'influencer dans son choix (i.e. $E_i^d \neg K_j E_i^d E_j^d\phi$ ou $E_i^d \neg B_j E_i^d E_j^d\phi$). Par exemple, une stratégie de l'agent manipulateur peut être de ne pas révéler son véritable profil de préférence afin d'influencer l'autre agent. Cette stratégie est bien une manipulation car pour être fonctionnelle, les autres agents ne doivent pas connaître la stratégie du manipulateur.

Enfin, nous définissons de la même manière les manipulations destructives.

Définition 3.4 - *Manipulations destructives* : *Soient* Σ *un ensemble fini et fermé de formules tel que* $\top \in \Sigma$ *et* $\phi \in \Sigma$ *une formule. Nous appelons* manipulation destructive douce avec dissimulation épistémique *la formule* $MDEK_{i,j}^{\Sigma}\phi$ *caractérisée par le prédicat :*

$$MDEK_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(\neg E_j\phi \wedge \neg K_j E_i^d \neg E_j\phi \wedge \psi)$$

Nous appelons manipulation destructive douce avec dissimulation doxastique *la formule* $MDEB_{i,j}^{\Sigma}\phi$ *caractérisée par le prédicat :*

$$MDEB_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(\neg E_j\phi \wedge \neg B_j E_i^d \neg E_j\phi \wedge \psi)$$

Nous appelons manipulation destructive forte avec dissimulation épistémique *la formule* $MDE^d K_{i,j}^{\Sigma}\phi$ *caractérisée par le prédicat :*

$$MDE^d K_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(\neg E_j^d\phi \wedge \neg K_j E_i^d \neg E_j^d\phi \wedge \psi)$$

Nous appelons manipulation destructive forte avec dissimulation doxastique *la formule* $MDE^d B^{\Sigma}_{i,j}\phi$ *caractérisée par le prédicat :*

$$MDE^d B^{\Sigma}_{i,j}\phi = \bigvee_{\psi \in \Sigma} E^d_i(\neg E^d_j \phi \wedge \neg B_j E^d_i \neg E^d_j \phi \wedge \psi)$$

Exemple 3.3 : *Un exemple de manipulation destructive est illustré par le cas des attaques éclipses (Singh et al., 2006). Ces attaques consistent à isoler un agen afin de l'exclure du réseau (i.e.* ψ*). Ce type de manipulation est capturé par la manipulation destructive. En effet, le pirate informatique s'assure au moment où il sévit que le nœud cible ne puisse plus communiquer avec les autres nœuds du réseau (i.e.* $E^d_i \neg E_j \phi$ *avec* ϕ *toute information communicable) tout en veillant à ce qu'il ne sache pas qu'il est à cet instant sous l'instance d'une attaque (i.e.* $E^d_i \neg K_j E^d_i \neg E_j \phi$*).*

Toutes ces définitions peuvent être fusionnées dans une unique définition générale de la manipulation.

Définition 3.5 - *Manipulation* : *Soient* Σ *un ensemble fini et fermé de formules tel que* $\top \in \Sigma$ *et* $\phi \in \Sigma$ *une formule. Nous appelons* manipulation *d'un agent* i *envers l'agent* j *sur* ϕ *dans le contexte* Σ *la formule* $M^{\Sigma}_{i,j}\phi$ *caractérisée par le prédicat :*

$$M^{\Sigma}_{i,j}\phi = \bigvee_{\Box \in \{B,K\}} (MCE\Box^{\Sigma}_{i,j}\phi \vee MCE^d\Box^{\Sigma}_{i,j}\phi \vee MDE\Box^{\Sigma}_{i,j}\phi \vee MDE^d\Box^{\Sigma}_{i,j}\phi)$$

3.3.3 Des stratégies de manipulation

Les stratégies de manipulation ne sont pas seulement capturées dans la sémantique de l'opérateur E^d_i. Elles peuvent aussi être exprimées explicitement par des formules logiques. Ainsi, en définissant la coercition, la persuasion et la tromperie avec des formules de KBE, nous expliquons en quoi « la manipulation n'est pas exactement de la coercition, ni de la persuasion et n'est pas entièrement similaire à la tromperie » (Handelman, 2009).

Dans la suite de cette section, nous considérons Σ un ensemble fini et fermé de formules tel que $\top \in \Sigma$ et $\phi \in \Sigma$ une formule de Σ. Cet ensemble représente les formules sur lesquelles les agents peuvent raisonner.

Coercition

Si l'influence délibérée sur les intentions est caractérisée par l'intention délibérée d'un agent i d'amener un autre agent à agir de façon délibérée ou non, la coercition quant-à-elle n'est jamais

dissimulée. Nous définissons l'influence délibérée par le prédicat :

$$infd_{i,j}^{\Sigma} = \bigvee_{\psi \in \Sigma} (E_i^d(E_j\phi \wedge \psi) \vee E_i^d(E_j^d\phi \wedge \psi))$$

Lorsqu'un braqueur pointe son arme sur quelqu'un afin qu'il obtienne son portefeuille, il ne manipule pas sa victime mais il exerce une influence non dissimulée sur son comportement. Celui-ci s'assure alors de façon délibérée que la victime sache qu'elle est sous pression (en braquant son arme). La coercition est donc cette influence exercée d'un agent sur un autre par moyen de pression et sans aucune dissimulation. Elle est alors décrite dans KBE comme une intention délibérée d'instrumentaliser tout en veillant à ce que la victime connaisse parfaitement les intentions de l'agent, c'est-à-dire :

$$coed_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(E_j^d\phi \wedge K_j E_i^d E_j^d\phi \wedge \psi)$$

La coercition est une forme d'influence non dissimulée portée sur les intentions d'un autre agent. Il existe d'autres formes d'influence qui ne sont pas portées uniquement sur des intentions mais plutôt sur les désirs d'un agent, les croyances, les connaissances, ou encore la confiance. Parmi l'une d'entre elles, la persuasion.

Persuasion

Un naturopathe cherche à persuader un client des bienfaits de l'huile d'origan sur le système immunitaire a pour intentions de faire croire le client cette information. Ainsi, un agent i persuade de ϕ un autre agent j si l'agent i veille de façon délibérée à ce que l'autre agent j croit ϕ. Ainsi, dans KBE, nous définissons la persuasion comme le prédicat :

$$perd_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(B_j\phi \wedge \psi)$$

Dans l'exemple mentionné, nous pouvons remarquer que le naturopathe ne dissimule pas ses intentions de convaincre son client. La persuasion n'est donc pas nécessairement dissimulée. Cependant, lorsqu'elle est dissimulée, c'est-à-dire lorsque l'agent i veille en plus de façon délibérée à ce que l'agent j ne sache pas les stratégies employées par l'agent i pour le persuader, nous parlons de tromperie.

Tromperie

La tromperie consiste à amener un agent à croire une certaine information ϕ tout en lui cachant une certaine vérité liée à cette information ϕ . Nous considérons deux formes de tromperie. Tout d'abord, la *dissimulation de la source* qui est le fait de cacher ses intentions de faire croire à un agent une information. Ensuite, nous considérons le *mensonge crédible* qui est le fait de veiller à ce que l'agent croit une certaine information tout en veillant à ce que cet agent ne sache pas que l'agent menteur croit le contraire.

La dissimulation de la source représente par exemple la propagation de fausses rumeurs dans

un système. Dans le cas des échanges en bourses, il arrive qu'un agent propage des rumeurs dans le but d'influencer les autres agents à acheter ou revendre certains de leurs produits sans que ces derniers ne sachent que c'est une stratégie appliquée par un agent manipulateur (Aggarwal and Wu, 2006). Ainsi, dans KBE, elle se caractérise par le prédicat :

$$con^{\Sigma}_{i,j}\phi = \bigvee_{\psi \in \Sigma} E^d_i(B_j\phi \wedge \neg K_j E^d_i B_j \phi \wedge \psi)$$

La dissimulation de la source est bien distincte du mensonge. En effet, Mahon définit un mensonge comme « faire croire à une autre personne une déclaration crue comme fausse et avec l'intention que celle-ci soit crue comme vraie par l'autre personne » (Mahon, 2008). Un mensonge est réel si l'agent menteur a aussi une intention de rendre crédible son mensonge en veillant à dissimuler cette intention de mentir. Nous parlons alors de *mensonge crédible* et nous l'exprimons dans KBE comme l'intention délibérée d'un agent i de faire qu'un agent j croit quelque chose alors que l'agent i croit en son contraire.

$$cre^{\Sigma}_{i,j}\phi = B_i\neg\phi \wedge (\bigvee_{\psi \in \Sigma} E^d_i(B_j\phi \wedge \neg K_j B_i \neg\phi \wedge \psi)))$$

Contrairement à (Sakama et al., 2015), nous n'avons pas besoin d'introduire une modalité de communication, la modalité de communication est réduite à l'intention délibérée de faire croire quelque chose. Sakama *et al.* définissent de nombreuses autres formes de tromperie que nous n'avons pas présenté dans cette section mais que nous pourrions définir avec KBE comme la demi-vérité, la tromperie par omission, ou encore le baratinage.

Les définitions formelles données de la coercition, de la persuasion et de la tromperie nous permettent de conclure qu'en effet « la manipulation n'est pas exactement de la coercition, ni de la persuasion et n'est pas entièrement similaire à la tromperie » (Handelman, 2009). Dans la suite de ce chapitre, nous proposons une instanciation du système KBE pour l'illustrer.

3.4 Une mise en situation de raisonnement avec KBE

L'exemple ci-dessous a pour objectif d'instancier le système KBE à une situation dans laquelle il est possible qu'un agent en manipule un autre.

Sur un site de e-commerce, deux agents effectuent une transaction commerciale. Notons i l'agent vendeur et j le client. L'agent i affirme à l'agent j : « Vous pouvez me faire confiance sur la qualité du produit. Vous ne trouverez nul part ailleurs un meilleur rapport qualité prix. Vous êtes libre d'aller vérifier l'information par vous même chez les concurrents. ».

Remarquons tout d'abord que dans la conversation l'agent i utilise des termes comme « vous êtes libre de », qui peuvent laisser entendre l'usage d'une technique de manipulation à des fins commerciales. Ces termes sont à la base d'une technique de manipulation de la théorie de la soumission librement consentie. Cette technique de manipulation consiste à donner l'illusion à l'agent manipulé d'agir librement et permet dans la suite de l'interaction de l'influencer plus

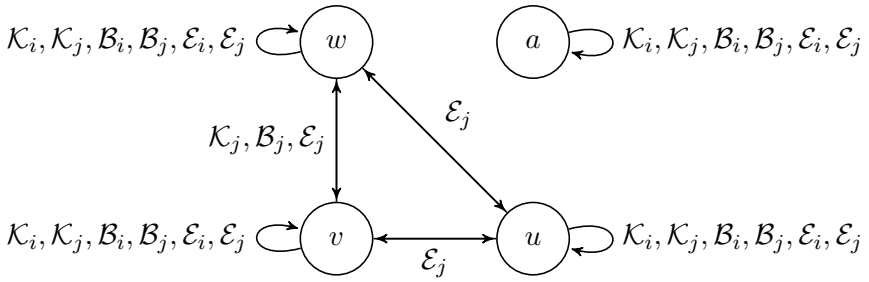

FIGURE 3.2 – États mentaux des agents

facilement [5]. Notons toutefois qu'il est aussi tout à fait possible que l'agent i n'ait utilisé ce terme sans pour autant avoir eu l'intention délibérée d'utiliser une telle stratégie de manipulation.

Pour représenter cette situation, nous considérons deux variables propositionnelles p et q :

p désigne que « l'agent j fait confiance en la sincérité de l'agent i sur la qualité du produit » ;

q désigne que « l'agent j achète le produit ».

Nous considérons alors plusieurs scénarios possibles et les représentons dans l'ensemble des mondes possibles $\mathcal{W} = \{a, w, v, u\}$.

w : « l'agent i a veillé de façon délibérée à instaurer la confiance afin d'amener l'agent j à acheter le produit »

v : « l'agent i n'a pas eu l'intention délibérée d'influencer j dans son achat »

u : « l'agent j achète le produit sans avoir confiance en i sur la qualité du produit et sait que l'agent i a l'intention délibéré de lui faire acheter le produit »

a : « l'agent j n'achète pas le produit et n'a pas confiance en i sur la qualité du produit »

Nous nous donnons un ensemble de formules $\Sigma = Cl(\Gamma)$ où $\Gamma = \{E_i^d(E_i^d p \Rightarrow E_i^d E_j q), E_i^d E_j q \wedge E_i^d \neg K_j E_i^d E_j q \Rightarrow E_i^d(E_j q \wedge \neg K_j E_i^d E_j q), E_i^d E_j^d q \wedge E_i^d K_j E_i^d E_j^d q \Rightarrow E_i^d(E_j^d q \wedge K_j E_i^d E_j^d q)\}$. Cet ensemble Σ, qui est défini comme la fermeture de l'ensemble Γ est donc fini et fermé. Il représente toutes les formules sur lesquelles les agents peuvent raisonner.

La fonction de valuation V du modèle décrivant cette situation est donnée par $V(p) = \{w, v\}$ et $V(q) = \{w, u, v\}$. De plus, les relations sont illustrées par la figure 3.2 et sont telles que :

1. $\mathcal{K}_i(w) = \{w\}$, $\mathcal{K}_i(v) = \{v\}$, $\mathcal{K}_i(u) = \{u\}$, $\mathcal{K}_i(a) = \{a\}$
2. $\mathcal{K}_j(w) = \{w, v\}$, $\mathcal{K}_j(v) = \{w, v\}$, $\mathcal{K}_j(u) = \{u\}$, $\mathcal{K}_j(a) = \{a\}$
3. $\mathcal{B}_i(w) = \{w\}$, $\mathcal{B}_i(v) = \{v\}$, $\mathcal{B}_i(u) = \{u\}$, $\mathcal{B}_i(a) = \{a\}$
4. $\mathcal{B}_j(w) = \{w, v\}$, $\mathcal{B}_j(v) = \{w, v\}$, $\mathcal{B}_j(u) = \{u\}$, $\mathcal{B}_j(a) = \{a\}$
5. $\mathcal{E}_i(w) = \{w\}$, $\mathcal{E}_i(v) = \{u, v\}$, $\mathcal{E}_i(u) = \{u\}$, $\mathcal{E}_i(a) = \{a\}$
6. $\mathcal{E}_j(w) = \{w, u, v\}$, $\mathcal{E}_j(v) = \{w, u, v\}$, $\mathcal{E}_j(u) = \{w, u, v\}$, $\mathcal{E}_j(a) = \{a\}$

5. En effet, il a été observé par les sociopsychologues que l'usage de termes comme « vous êtes libre de », peuvent influencer fortement une personne dans le choix désiré par un manipulateur (Joule et al., 2002).

7. $\mathcal{E}_i^d(w) = \{\{w,v\},\{w,u,v\},\{w,u,a\},\{w,v,a\},\{w\},\{w,a\},\{w,u\}\}$,
   $\mathcal{E}_i^d(v) = \{\{u,v\}\}$, $\mathcal{E}_i^d(u) = \{\{u\},\{w,u,v\}\}$, $\mathcal{E}_i^d(a) = \{\{w,a\}\}$
8. $\mathcal{E}_j^d(w) = \{\{w,u,v\}\}$, $\mathcal{E}_j^d(v) = \{\{w,u,v\}\}$, $\mathcal{E}_j^d(u) = \{\{w,u,v\}\}$, $\mathcal{E}_j^d(a) = \{\{w,a\}\}$

(1) et (2) décrivent le fait que l'agent i sait parfaitement si l'agent j a confiance en lui et si l'agent j achète le produit. De plus, (3) et (4) imposent ici que les agents croient ce qu'ils savent et réciproquement, i.e. $\mathcal{K}_i = \mathcal{B}_i$ et $\mathcal{K}_j = \mathcal{B}_j$.

(5) dans le monde possible w, l'agent veille à ce que p et q sinon dans les mondes v et u, l'agent i veille seulement à ce que le client lui achète le produit.

(6) l'agent j veille à acheter le produit dans $\{w,u,v\}$ mais ne veille pas nécessairement à faire confiance à i.

(7) L'agent i a l'intention délibéré dans w que l'agent j lui fasse confiance et veille de façon délibérée à ce que si l'agent j lui fait confiance, alors l'agent i achète le produit et ce tout en veillant à lui dissimuler sa stratégie pour lui faire acheter le produit.

(8) Enfin, l'agent j, dans w,u,v a seulement l'intention d'acheter le produit.

Remarquons tout d'abord que dans w l'agent i influence de manière délibérée l'agent j pour qu'il achète le produit. En effet, nous avons $|E_i^d p \Rightarrow E_i^d E_j q| = \{w,u,a\}$ et $\{w,u,a\} \in \mathcal{E}_i^d(w)$. Donc $\mathcal{M}, w \models E_i^d(E_i^d p \Rightarrow E_i^d E_j q)$. Ainsi, par application du théorème $\models E_i^d\phi \Rightarrow \phi$, nous déduisons que dans w, nous avons alors $\mathcal{M}, w \models E_i^d p \Rightarrow E_i^d E_j q$. Or $|p| = \{w,v\}$ et $\{w,v\} \in \mathcal{E}_i^d(w)$. Donc $\mathcal{M}, w \models E_i^d p$. Par conséquent, nous avons donc que $\mathcal{M}, w \models E_i^d E_j q$.

De plus, nous pouvons remarquer par la même occasion, que dans w, l'agent i a aussi l'intention de dissimuler sa stratégie pour amener l'agent j à acheter le produit. En effet, nous avons dans v que comme $|E_j q| = \{w,u,v\}$ et $\{w,u,v\} \notin \mathcal{E}_i^d(v)$, $\mathcal{M}, v \models \neg E_i^d E_j q$. Or l'agent j ne peut discerner entre les mondes w et v. Ainsi, nous avons donc que $\mathcal{M}, w \models \neg K_j E_i^d E_j q$. De plus, nous pouvons remarquer que $|\neg K_j E_i^d E_j q| = \{w,v,a\}$ [6] et $\{w,v,a\} \in \mathcal{E}_i^d(w)$. Par conséquent, puisque $|\neg K_j E_i^d E_j q| \in \mathcal{E}_i^d(w)$, nous avons $\mathcal{M}, w \models E_i^d \neg K_j E_i^d E_j q$.

En conclusion, nous avons donc montré que $\mathcal{M}, w \models E_i^d E_j q \wedge E_i^d \neg K_j E_i^d E_j q$. Or, par le théorème $\models E_i^d\phi \wedge E_i^d\psi \Rightarrow E_i^d(\phi \wedge \psi)$, nous déduisons que $\mathcal{M}, w \models E_i^d(E_j q \wedge \neg K_j E_i^d E_j q)$. Ainsi, nous avons donc pu montrer que dans cette situation, il existe un monde possible dans lequel l'agent i est en train de manipuler l'agent j pour l'amener à acheter le produit. De plus, si nous décomposons les intentions délibérées de l'agent i, $\mathcal{E}_i^d(w) = \{\{w,v\},\{w,u,v\},\{w,u,a\},\{w,v,a\},\{w\},\{w,a\},\{w,u\}\}$, nous remarquons que l'agent i a d'une part l'intention de rendre vrai p en considérant l'ensemble $|p| = \{w,v\}$, et d'autre part de rendre vrai q en considérant l'ensemble $|q| = \{w,u,v\}$. Cet agent a aussi la stratégie d'amener l'autre agent à acheter le produit par l'ensemble $|E_i^d p \Rightarrow E_i^d E_j q| = \{w,u,a\}$, et enfin sa stratégie de dissimulation est représentée par l'ensemble $|\neg K_j E_i^d E_j q| = \{w,v,a\}$. Pour finir, les ensembles $\{w\}$, $\{w,a\}$, $\{w,u\}$ sont donnés par la contrainte imposée (CE) sur le cadre pour permettre le théorème $\models E_i^d\phi \wedge E_i^d\psi \Rightarrow E_i^d(\phi \wedge \psi)$. Ces ensembles de mondes possibles traduisent simplement le fait que lorsqu'un agent met en place différents plans indépendants, cet agent considère bien aussi les intersections de mondes possibles entre ces différents plans qui ont été délibérés par l'agent.

6. Pour $a \in |\neg K_j E_i^d E_j q|$, il suffit de remarquer que $|E_j q| \notin \mathcal{E}_i^d(a)$, donc $\mathcal{M}, a \models \neg E_i^d E_j q$. Or, $\forall x \in \mathcal{W}$: $a\mathcal{K}_j x, \mathcal{M}, x \models \neg E_i^d E_j q$, donc $\mathcal{M}, a \models K_j \neg E_i^d E_j q$, et donc $\mathcal{M}, a \models \neg K_j E_i^d E_j q$.

Un dernier point amusant à remarquer sur ce modèle est que dans le monde u, l'agent i effectue de la coercition sur l'agent j afin de le pousser à acheter le produit. En effet, puisque nous avons $|E_i^d E_j^d q| = \{w, u\}$ et $|K_j E_i^d E_j^d q| = \{u\}$ [7] et que $|E_i^d E_j^d q| \cap |K_j E_i^d E_j^d q| = \{u\} \in \mathcal{E}_i^d(u)$, nous déduisons immédiatement que $\mathcal{M}, u \models E_i^d(E_j^d q \wedge K_j E_i^d E_j^d q)$ et donc $\mathcal{M}, u \models coe_{i,j}^d q$.

Nous avons donc proposé un premier cadre logique permettant d'exprimer la manipulation comme une intention délibérée d'instrumentaliser une victime tout en s'assurant à lui dissimuler cette intention. Ce nouveau cadre logique, nommé KBE, introduit une nouvelle modalité permettant d'exprimer l'intention délibérée. Nous avons ensuite prouvé que ce système était correct et complet, et nous permettait de déduire de nouveaux théorèmes comme la dissimulation de ses croyances contraires, ou encore le principe *qui facit per alium facit per se*. Enfin, nous avons donné une définition explicite à la manipulation, et il se trouve que ce n'est ni de la *coercition*, ni de la *persuasion*, ou de la *tromperie*.

Dans la suite, puisque beaucoup de stratégies de manipulation reposent sur la confiance, nous proposons de compléter ce cadre en intégrant une notion de confiance en la sincérité permettant à un agent d'accorder sa confiance sur des propositions émises par d'autres agents.

7. Pour s'en assurer, il suffit de calculer l'ensemble $|E_j^d q| = \{w, u, v\}$ et de remarquer que le seul monde possible x tel que $\forall z \in \mathcal{W}, x \mathcal{K}_j z, \mathcal{M}, z \models E_i^d E_j^d q$ est le monde $x = u$.

Chapitre 4

Le système TB - raisonner sur la confiance en la sincérité

Dans le chapitre 3, nous avons proposé la logique KBE permettant de raisonner sur la notion de manipulation. Dans un contexte où des agents peuvent être manipulateurs, les agents intègrent souvent une notion de confiance pour réguler leurs interactions avec les autres agents. Cependant, nous n'avons pas exprimé cette dernière notion avec KBE. Si des travaux se sont intéressés à représenter la confiance comme nous l'avons vu au chapitre 2, ceux-ci ne représentent souvent que la confiance en la fiabilité des agents et ne considèrent pas de notion de « confiance en l'absence de comportement manipulateur ». Or dans un tel contexte, la notion qui se rapproche au mieux de « l'absence de comportement manipulateur » est la notion de sincérité. Cette notion de confiance en la sincérité est importante dans les systèmes de réputation où les agents utilisent les témoignages d'autres agents pour pouvoir évaluer la réputation d'un agent spécifique. Bien que les témoignages recueillis peuvent être inconsistants d'un agent à un autre, car ils représentent les interactions passées des agents avec un agent spécifique, ces témoignages sont intéressants s'ils proviennent d'agents sincères. Ainsi dans un premier temps, nous définissons la *sincérité*, la notion d'*agents sincères* et ce que nous appelons la *confiance en la sincérité* en section 4.1. Puis, nous proposons dans ce chapitre un nouveau système logique, nommé TB, permettant de raisonner sur cette notion de confiance en la sincérité. En section 4.2, nous présentons ce système logique. Nous prouvons ensuite en section 4.3 que ce système est correct et fortement complet, puis, nous nous intéressons en section 4.4 à quelques propriétés logiques de la confiance en la sincérité. Par la suite, nous proposons d'étendre cette notion de confiance entre deux agents à une notion de confiance collective entre plusieurs groupes d'agents. Enfin, en section 4.5, nous proposons une application de ce système logique.

4.1 La notion de confiance en la sincérité

De façon usuelle, lorsque nous parlons de « sincérité », cela se rapporte au caractère d'un agent « qui exprime, sans les déguiser, ses pensées et ses sentiments ; qui est réellement tel, d'une manière convaincue ; qui est marqué par la franchise, la droiture, qui est réellement éprouvé » (La-

rousse, 1867). Par ailleurs, dans le dictionnaire d'Oxford la sincérité y est définie comme « l'absence de faux-semblants, de tromperie ou d'hypocrisie »[1]. Ainsi, en s'appuyant sur ces deux définitions, nous considérons alors la définition suivante d'un agent sincère.

Définition 4.1 - *Agent sincère* : *Un agent est* sincère *si, et seulement si, cet agent ne transmet aucune information qu'il ne croit pas et dont les intentions ne sont pas de manipuler d'autres agents*[2].

La confiance de manière abstraite caractérise le choix d'un agent de considérer qu'un autre agent respecte une certaine propriété abstraite (cf. définition 1.3). La *confiance en la sincérité* désigne alors le choix d'un agent de considérer qu'un autre agent est sincère. Dans ce chapitre, nous allons nous intéresser à exprimer logiquement la notion de *confiance en la sincérité relative à une proposition.*

Définition 4.2 - *Confiance en la sincérité* : *La* confiance en la sincérité relative à une proposition *désigne le choix d'un agent de considérer qu'un autre agent est sincère sur une proposition*[3].

Ainsi, un agent qui accorde sa confiance en la sincérité à un autre agent, croit nécessairement que ce dernier croit ce qu'il dit. Cependant, puisque la confiance est un choix, nous ne pouvons pas considérer la réciproque. En effet, si un agent propage une rumeur dans le but d'influencer une décision collective, cet agent peut révéler une information qu'il croit vraie, mais cet agent n'est pas sincère car son intention est de manipuler les autres agents.

Dans la suite, nous parlons de confiance en la sincérité d'un agent en faisant toujours référence à la *confiance en la sincérité relative à une proposition.*

4.2 Une logique de la confiance en la sincérité

Nous présentons dans cette section la logique modale TB qui a pour particularité d'intégrer une modalité qui représente la confiance qu'un agent i accorde à un agent j envers sa sincérité *pour un énoncé* ϕ. Pour cela, nous définissons un langage noté $\mathcal{L}_{TB}$.

4.2.1 Le langage $\mathcal{L}_{TB}$

Soient un ensemble fini de variables propositionnelles $\mathcal{P} = \{p, q, r, s, t, \ldots\}$, un ensemble d'agents $\mathcal{N}$, deux agents $i, j \in \mathcal{N}$, et $p \in \mathcal{P}$ une variable propositionnelle. Le langage $\mathcal{L}_{TB}$ est

1. Dans le dictionnaire d'Oxford, la définition anglaise de la sincérité est la suivante : *Sincerity is « the quality of being free from pretense, deceit, or hypocrisy »*. Cette définition est consultable sur le site https://www.lexico.com/en/definition/sincerity.

2. Cette définition diffère légèrement de celle de (Demolombe, 2004) qui considère qu'un agent est *sincère* si, et seulement si, celui-ci ne transmet uniquement des informations qu'il croit nécessairement.

3. Cette définition diffère encore de la définition de la confiance en la sincérité de (Demolombe, 2011) qui considère qu'un agent fait confiance en la sincérité d'un autre agent si, et seulement si, l'agent qui fait confiance croit que lorsque l'autre agent transmet une information, alors nécessairement, ce dernier croit l'information qu'il transmet.

généré par la grammaire sous forme de Backus-Naur suivante :

$$\psi ::= p|\neg\psi|\psi \wedge \psi|\psi \vee \psi|\psi \Rightarrow \psi|T^s_{i,j}\psi|B_i\psi$$

Nous considérons des modalités B_i et $T^s_{i,j}$ pour tout agent $i, j \in \mathcal{N}$. Ainsi, $B_i\phi$ exprime le fait qu'un agent i croit la proposition ϕ tandis que la formule $T^s_{i,j}\phi$ exprime qu'un agent i fait confiance à la sincérité de j sur la proposition ϕ[4]. De plus, remarquons que lorsque $i = j$, la formule $T^s_{i,i}\phi$ signifie que l'agent i a confiance en sa sincérité sur la proposition ϕ.

Par ailleurs, notons que nous n'introduisons pas de modalités pour exprimer les connaissances d'un agent (K_i), ni de modalité d'acquisition d'informations d'un l'agent i émanant d'un autre agent j ($I_{j,i}$), ni de modalité de communication d'un agent i vers un agent j ($C_{i,j}$) comme nous pouvons l'avoir dans (Liau, 2003; Demolombe, 2004). Pour les détails concernant ces travaux, nous renvoyons le lecteur à la section 2.2. Ici, nous considérons le minimum de modalités pour exprimer la notion de confiance en la sincérité. De plus, nous ne les introduisons pas car informer et acquérir une information d'un agent peuvent être vus comme deux choses bien distinctes, et nous ne voulons pas rentrer dans le débat d'utiliser une modalité plutôt qu'une autre. La modalité $C_{i,j}$ peut être vue comme une action de communication tandis que la modalité $I_{j,i}$ est une action d'acquisition de l'information qui est une conséquence d'une action de communication réussie. D'autre part, ces modalités alourdissent le modèle sans pour autant lui apporter plus d'expressivité. Ainsi, plutôt que d'introduire explicitement cette distinction dans notre modèle, nous préférons considérer un cadre plus simple pour nous concentrer uniquement sur l'axiomatique de la confiance en la sincérité. En réalité, cette notion de communication ou d'acquisition d'information est implicite à notre modalité $T^s_{i,j}$. Ainsi, lorsqu'un agent i accorde sa confiance à un agent j sur sa sincérité pour une proposition ϕ, il est bien le cas que l'agent i a acquis une certaine information émanant de l'agent j à cet instant donné où il accorde sa confiance.

4.2.2 Sémantique de la confiance en la sincérité

La confiance en la sincérité a été définie en section 4.1 comme le choix d'un agent d'accorder sa confiance à un autre agent vis-à-vis de sa sincérité. La confiance est donc avant tout un choix d'un agent, contrairement à la sincérité qui est une propriété qu'un agent possède. Ainsi, un agent peut croire qu'un autre agent croit une certaine information sans pour autant faire confiance en sa sincérité lorsqu'il révèle cette information. En effet, un agent manipulateur pourrait, par exemple, révéler une certaine information qu'il croit comme vraie en vue de mettre en place une stratégie de manipulation. De plus, un agent rationnel ne peut faire confiance en la sincérité d'un autre agent si lui-même ne croit pas que l'autre agent croit ce qu'il dit. Ainsi, si un agent accorde sa confiance envers un autre agent, nécessairement l'agent qui accorde sa confiance croit que l'autre agent croit ce qu'il dit, puisque les agents sont supposés rationnels et ne vont donc pas accorder leur confiance si elle n'est pas fondée. Cette propriété de la confiance en la sincérité

4. Remarquons que la formule $\neg T^s_{i,j}\phi$ est littéralement traduite par « il n'est pas le cas que l'agent i fait confiance à la sincérité de j sur la proposition ϕ ». Cependant, dans la suite de ce travail, nous faisons l'abus de langage suivant en considérant que la négation de la confiance est traduite par « l'agent i n'a pas confiance en la sincérité de l'agent j sur la proposition ϕ ».

est appelée *contrainte de sincérité*.

Pour définir sémantiquement la confiance en la sincérité, nous considérons un cadre de Kripke $\mathcal{C} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}})$ pour interpréter les formules du langage $\mathcal{L}_{TB}$ tel que :

— $\mathcal{W}$ est un ensemble de mondes ;

— $\{\mathcal{B}_i\}_{i\in\mathcal{N}}$ est un ensemble de relations binaires tel que :

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{B}_i(w) := \{v \in \mathcal{W} | w\mathcal{B}_i v\}$$

— $\{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}$ est un ensemble de relations binaires tel que :

$$\forall i, j \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{T}_{i,j}(w) := \{v \in \mathcal{W} | w\mathcal{T}_{i,j} v\}$$

Pour tout agent $i, j \in \mathcal{N}$, $\mathcal{B}_i$ représente une relation d'indiscernabilité telle qu'elle est utilisée pour exprimer la sémantique d'opérateurs modaux doxastiques (B_i) tandis qu'une relation $\mathcal{T}^s_{i,j}$ représente les mondes indiscernables pour un agent i par rapport à la confiance qu'il accorde à l'agent j au regard de sa sincérité ($T^s_{i,j}$). Un modèle de Kripke pour le cadre TB est un N-uplet $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$ avec $V : \mathcal{P} \to 2^{\mathcal{W}}$ une fonction d'interprétation. Pour tout monde $w \in \mathcal{W}$, pour toute formule $\phi, \psi \in \mathcal{L}_{TB}$ et pour tout atome $p \in \mathcal{P}$, nous considérons la sémantique :

0. $\mathcal{M}, w \models \top$
1. $\mathcal{M}, w \not\models \bot$
2. $\mathcal{M}, w \models p$ ssi $w \in V(p)$
3. $\mathcal{M}, w \models \neg\phi$ ssi $w \not\models \phi$
4. $\mathcal{M}, w \models \phi \vee \psi$ ssi $\mathcal{M}, w \models \phi$ ou $\mathcal{M}, w \models \psi$
5. $\mathcal{M}, w \models \phi \wedge \psi$ ssi $\mathcal{M}, w \models \phi$ et $\mathcal{M}, w \models \psi$
6. $\mathcal{M}, w \models \phi \Rightarrow \psi$ ssi $\mathcal{M}, w \models \neg\phi$ ou $\mathcal{M}, w \models \psi$
7. $\mathcal{M}, w \models B_i\phi$ ssi $\forall v \in \mathcal{W} : w\mathcal{B}_i v, \mathcal{M}, v \models \phi$
8. $\mathcal{M}, w \models T^s_{i,j}\phi$ ssi $\forall v \in \mathcal{W} : w\mathcal{T}_{i,j} v, \mathcal{M}, v \models \phi$

Nous rappelons les notations duales des modalités, pour pour tout monde $w \in \mathcal{W}$, pour tout $(\Diamond_i, \mathcal{R}_i) \in \{(\langle T^s_{i,j}\rangle, \mathcal{T}_{i,j}), (\langle B_i\rangle, \mathcal{B}_i)\}$, nous avons $\mathcal{M}, w \models \Diamond_i\phi$ si, et seulement si, $\exists v \in \mathcal{W} : w\mathcal{R}_i v, \mathcal{M}, v \models \phi$.

Enfin, ϕ est valide dans un modèle $\mathcal{M}$ de TB (noté $\mathcal{M} \models \phi$) si, et seulement si, pour tout monde $w \in \mathcal{W}$, ϕ est satisfiable dans w i.e. $\mathcal{M}, w \models \phi$ est vraie. Une formule ϕ est valide dans un cadre $\mathcal{C}$ de TB (noté $\models_{\mathcal{C}} \phi$ ou $\mathcal{C} \models \phi$) si, et seulement si, pour tout modèle $\mathcal{M}$ fondé sur $\mathcal{C}$, $\mathcal{M} \models \phi$. Une formule ϕ est la *conséquence sémantique* d'un ensemble de formules Γ dans $\mathcal{C}$ et notons $\Gamma \models_{\mathcal{C}} \phi$ si, et seulement si, pour tout modèle $\mathcal{M}$ de TB $\mathcal{M} \models \Gamma$, implique $\mathcal{M} \models \phi$.

Le cadre de Kripke $\mathcal{C}$ associé à notre modèle est tel que :

1. Pour tout $i, j \in \mathcal{N}$, $\forall w \in \mathcal{W}, \exists v \in \mathcal{W} : w\mathcal{T}_{i,j} v$
2. Pour tout $i, j \in \mathcal{N}$, $\forall w, u, v \in \mathcal{W} : w\mathcal{B}_i u \wedge u\mathcal{T}_{i,j} v \Rightarrow w\mathcal{T}_{i,j} v$
3. Pour tout $i, j \in \mathcal{N}$, $\forall w, u, v \in \mathcal{W} : w\mathcal{B}_i u \wedge w\mathcal{T}_{i,j} v \Rightarrow u\mathcal{T}_{i,j} v$
4. Pour tout $i, j \in \mathcal{N}$, $\forall w, u, v \in \mathcal{W} : w\mathcal{B}_i u \wedge u\mathcal{B}_j v \Rightarrow w\mathcal{T}_{i,j} v$

5. Pour tout $i \in \mathcal{N}$, $\mathcal{B}_i$ est sérielle, transitive et euclidienne.

Lorsque nous faisons confiance en la sincérité du discours d'un autre agent, nous prenons le risque d'être trahi. Ainsi, un moyen de se protéger face à la trahison est de ne pas pouvoir faire confiance en quelque chose et son contraire. En effet, nous ne pouvons pas faire confiance à quelqu'un qui dit une chose et se contredit en même temps. Un exemple flagrant de ce lien entre confiance et non contradiction est très bien illustré par une enquête policière. Les policiers font confiances à la sincérité des témoins tant qu'ils ne se contredisent pas. Par conséquent, un moyen de considérer ce principe consiste à dire qu'il existe toujours un monde accessible par $\mathcal{T}_{i,j}$ depuis n'importe quel monde à l'aide la propriété (1).

Un agent a conscience de la confiance qu'il accorde à un autre agent. La propriété (2), donnée par (Liau, 2003; Dundua and Uridia, 2010), correspond à l'axiome d'introspection positive sur la confiance. Si un agent fait confiance, alors cet agent croit qu'il fait confiance. De plus, nous considérons aussi la réciproque qui est l'introspection négative. Elle décrit le fait que si nous ne faisons pas confiance à quelqu'un alors nous avons bien conscience que nous ne faisons pas confiance à cette personne.

La propriété (4) correspond à la contrainte de sincérité. Elle signifie que s'il est possible que l'agent i croit que l'agent j puisse croire une information ϕ vraie dans v, alors il est nécessairement possible que l'agent i puisse faire confiance à l'agent j vis-à-vis de sa sincérité sur toute information possible ϕ vraie dans v et émise par cet agent j. Nous venons alors de décrire la propriété fondamentale de la confiance en la sincérité qui est $\models T^s_{i,j}\phi \Rightarrow B_i B_j \phi$.

Enfin, les dernières propriétés sont celles usuellement utilisées pour représenter la sémantique d'une modalité doxastique (Liau, 2003; Herzig et al., 2010; Demolombe, 2004).

4.2.3 Système axiomatique

Nous notons $\vdash \phi$ pour dire que ϕ est un théorème qui peut être déduit à partir d'un système de preuves à la Hilbert. Ainsi, la figure C.1 représente l'ensemble des axiomes dans un tel système. Chaque axiome provient directement des contraintes sémantiques imposées sur le cadre TB. Nous considérons les axiomes suivants : les tautologies du calcul propositionnel ($\mathbf{R_1}$ à $\mathbf{R_8}$), les règles d'inférence classiques des logiques modales (**MP**, **Nec**, **Sub**). Notre logique de la confiance en la sincérité est donc une logique normale qui satisfait la nécessitation, la substitution, le modus ponens et l'axiome de Kripke K pour toutes ses modalités[5]. Enfin, la définition de TB-déductibilité est fondée sur la notion de $\mathcal{S}$-déductibilité présentée en section 2.1.1 où $\mathcal{S} = TB$.

L'axiome **K** pour la modalité de confiance traduit la propriété que si un agent i a confiance en un agent j sur ϕ qui implique ψ alors, si i a confiance en j pour ϕ alors i a aussi confiance en

5. Pour rappel, la **nécessitation** signifie que si une proposition ϕ est un théorème ($\vdash \phi$) alors n'importe quel agent i peut avoir confiance en n'importe quel autre agent j pour ce théorème ($\vdash T^s_{i,j}\phi$). La **substitution** signifie que si un théorème est prouvé et que nous substituons uniformément une formule quelconque à une lettre de proposition, la formule résultante est aussi un théorème. Le **modus ponens** signifie que si une proposition ϕ est prouvée et qu'il est aussi prouvé que $\phi \Rightarrow \psi$ alors la formule ψ est prouvée.

(CP)	Tous les théorèmes du CP
$(KD45_{B_i})$	Tous les théorèmes de KD45 sont vérifiés pour B_i
$(KD_{T^s_{i,j}})$	Tous les théorèmes de KD sont vérifiés pour $T^s_{i,j}$
(4_{TB})	$\vdash T^s_{i,j}\phi \Rightarrow B_i T^s_{i,j}\phi$
(5_{TB})	$\vdash \neg T^s_{i,j}\phi \Rightarrow B_i \neg T^s_{i,j}\phi$
(S)	$\vdash T^s_{i,j}\phi \Rightarrow B_i B_j \phi$

FIGURE 4.1 – Système axiomatique du cadre TB

j pour ψ, c'est-à-dire, le système TB vérifie le théorème suivant :

$$\vdash T^s_{i,j}(\phi \Rightarrow \psi) \Rightarrow T^s_{i,j}\phi \Rightarrow T^s_{i,j}\psi \quad (K)$$

À noter que (Liau, 2003), présenté en section 2.2.2, considère une modalité abstraite non normale de la confiance dans le processus d'acquisition de nouvelles informations. Puisqu'il considère une sémantique minimale, l'axiome K n'est pas vérifié dans son système logique. Il justifie sa position par le fait que si un agent i accorde sa confiance envers un autre agent j lorsqu'il affirme que la situation financière d'une compagnie est excellente (ϕ), et que si tel est le cas, il est alors intéressant d'investir dans cette compagnie (ψ). La formule $T_{i,j}(\phi \wedge (\phi \Rightarrow \psi))$ est donc supposée vraie dans cet exemple. Cependant, pour Liau, il n'est pas question de déduire $T_{i,j}\psi$ car selon lui, ce n'est pas parce qu'un agent i a confiance en la fiabilité du jugement de j pour ces deux propositions ϕ et $\phi \Rightarrow \psi$ que i a nécessairement confiance en j sur la proposition ψ. Autrement dit, l'argument de Liau signifie que même si l'agent i fait confiance en la fiabilité de l'agent j pour $\phi \wedge (\phi \Rightarrow \psi)$, l'agent i ne peut pas nécessairement avoir confiance en la fiabilité de la conséquence logique ψ. Cela semble contre-intuitif. En effet, cela signifierait que bien que l'agent i soit supposé rationnel et fiable lorsqu'il accorde sa confiance à l'agent j, cet agent i ne peut déduire que s'il fait confiance en la fiabilité de ses informations, il est donc intéressant d'investir dans cette compagnie et qu'il peut donc aussi lui faire confiance sur l'implication ψ. Ainsi, s'il s'agit de la confiance de i en la fiabilité de j pour les informations qu'il transmet, il est nécessaire de déduire $T_{i,j}\psi$. Par ailleurs, même si Liau considère une notion de confiance en la fiabilité et que dans ce chapitre, nous considérons plutôt une notion de confiance en la sincérité d'un agent, notre argument reste inchangé et est tout à fait applicable à TB.

Non-inconsistance de la confiance

Un agent i qui fait confiance à j pour un énoncé ϕ ne peut pas faire confiance à l'agent j pour son contraire, pour des raisons de cohérence dans le discours. Il n'est donc pas possible de faire confiance à un agent qui raisonne de manière inconsistante.

$$\vdash T^s_{i,j}\phi \Rightarrow \neg T^s_{i,j}\neg\phi \quad (D_{T^s_{i,j}})$$

Ce théorème signifie que si un agent i a confiance en un agent j pour sa sincérité sur ϕ

alors cet agent i n'a pas confiance en j pour sa sincérité sur $\neg\phi$. Cependant, nous ne pouvons généraliser ce théorème à tout autre agent $k \in \mathcal{N}$, $T^s_{i,j}\phi \not\Rightarrow \neg T^s_{i,k}\neg\phi$ car si un agent i a confiance en j sur ϕ, c'est-à-dire, $T^s_{i,j}\phi$, rien ne nous dit et nous empêche d'avoir $T^s_{i,k}\neg\phi$ pour un autre agent $k \in \mathcal{N}$. Contrairement à (Liau, 2003) qui ne permet pas de faire confiance à deux sources qui se contredisent entre elles, ici, deux agents peuvent avoir des propos contradictoires et cela ne signifie pas pour autant qu'ils ne sont pas sincères. De plus, si un tel théorème était vérifié dans TB, nous obtiendrions aussi le théorème $T^s_{i,j}p \wedge T^s_{i,k}\neg p \Rightarrow (\neg T^s_{i,k}\neg p \wedge T^s_{i,k}\neg p)$ qui n'est généralement pas vrai. En effet, si la formule $T^s_{i,j}p \wedge T^s_{i,k}\neg p$ est vraie, alors la formule $T^s_{i,j}p \wedge T^s_{i,k}\neg p \Rightarrow (\neg T^s_{i,k}\neg p \wedge T^s_{i,k}\neg p)$ est fausse car elle est aussi équivalente à $T^s_{i,j}p \wedge T^s_{i,k}\neg p \Rightarrow \bot$.

Lien entre confiance et croyance

(Liau, 2003; Dundua and Uridia, 2010) considèrent un lien entre la confiance et la croyance. La sémantique de notre système amène aussi le théorème qui pour tout agent $i, j \in \mathcal{N}$ si un agent i a confiance en un agent j sur l'énoncé ϕ alors l'agent i croit qu'il a confiance en l'agent j sur l'énoncé ϕ :

$$\vdash T^s_{i,j}\phi \Rightarrow B_i T^s_{i,j}\phi \quad (4_{T,B})$$

De la même manière, nous obtenons le théorème que si un agent i n'a pas confiance en un agent j sur l'énoncé ϕ alors l'agent i croit qu'il n'a pas confiance en l'agent j sur l'énoncé ϕ :

$$\vdash \neg T^s_{i,j}\phi \Rightarrow B_i \neg T^s_{i,j}\phi \quad (5_{T,B})$$

Notons que nous ne pouvons pas considérer un axiome de non-inconsistance entre la confiance d'un agent et ses croyances, c'est-à-dire si un agent croit que quelque chose comme vrai, cela n'implique pas qu'il ne fait pas confiance dans un discours qui annonce le contraire de sa croyance, i.e, $\forall i, j \in \mathcal{N}, B_i\phi \not\Rightarrow \neg T^s_{i,j}\neg\phi$. En effet, si nous prenons le cas de la confiance en la sincérité, nous pouvons croire ϕ et avoir confiance en la sincérité d'un agent pour son contraire $\neg\phi$.

Enfin, un dernier axiome important à souligner est *l'axiome de sincérité*. Cet axiome découle directement de la contrainte de sincérité. Ce dernier est exprimé par le fait que si un agent i fait confiance en la sincérité d'un autre agent j pour ϕ alors i croit que j croit ϕ, c'est-à-dire :

$$\vdash T^s_{i,j}\phi \Rightarrow B_i B_j \phi \quad (S)$$

La réciproque est bien évidemment non correcte puisque ce n'est pas parce que l'agent i croit que l'agent j croit ϕ que i fait confiance à j pour sa sincérité sur ϕ puisque la confiance est un choix de l'agent i. Cette propriété ne peut être exprimée dans le modèle de Liau, même en considérant une sémantique minimale. En effet, même s'il n'est pas contradictoire d'écrire $T_{i,j}p \wedge T_{i,j}\neg p$ dans le modèle de Liau, le fait d'avoir un axiome $T_{i,j}p \Rightarrow B_i B_j p$, implique que nous aurions aussi le théorème $(T_{i,j}p \wedge T_{i,j}\neg p) \Rightarrow (B_i B_j p \wedge B_i B_j \neg p)$. Ce qui n'est généralement pas vrai et ne peut donc être considéré comme un théorème valide. En effet, dans le cadre de Liau, si $(T_{i,j}p \wedge T_{i,j}\neg p)$ est vraie, nous déduisons une contradiction. Cependant, dans notre cadre, nous pouvons considérer ce théorème, car le faux implique le faux est toujours vérifié. De la même

manière si nous considérons la formule $T_{i,j}p \wedge T_{i,k}\neg p \Rightarrow (B_iB_jp \wedge B_iB_k\neg p)$, Liau prouverait dans son système BA (cf. section 2.2.2) qu'un agent ne peut pas faire confiance dans deux sources différentes et contradictoires alors que, dans notre cas, il est tout à fait possible de faire confiance en la sincérité de deux sources qui se contredisent.

4.3 Correction et complétude du système TB

Dans cette section, nous rappelons dans un premier temps, des résultats standards de formules valides des logiques modales, puis nous montrons que chacune des formules données par le système axiomatique préservent bien la validité. Nous prouvons dans un second temps que notre système TB est correct puis, nous montrons que ce système axiomatique est complet. Enfin, nous montrons que TB possède aussi les propriétés de correction forte et de complétude forte.

4.3.1 Correction

Nous considérons un cadre quelconque $\mathcal{C} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}^s_{i,j}\}_{i,j\in\mathcal{N}})$ sur $\mathcal{L}_{TB}$ et rappelons que l'axiome $(D_{T^s_{i,j}})$ correspond à la sérialité de la relation $\mathcal{T}^s_{i,j}$.

Proposition 4.1 : *Pour tout $i, j \in \mathcal{N}$,*

$$\mathcal{C} \models T^s_{i,j}p \Rightarrow \neg T^s_{i,j}\neg p$$

si, et seulement si,

$$\forall w \in \mathcal{W}, \exists v \in \mathcal{W} : w\mathcal{T}^s_{i,j}v$$

Démonstration. La preuve est standard aux systèmes KD (Blackburn et al., 2002). □

De la même manière, nous rappelons que les propriétés correspondantes aux systèmes KD45 sont les relations sérielles, transitives et euclidiennes.

Proposition 4.2 : *Pour tout agent $i \in \mathcal{N}$, tous les axiomes KD45 pour B_i sont vérifiés dans $\mathcal{C}$ si, et seulement si, $\mathcal{C}$ est sériel, transitif et euclidien pour $\mathcal{B}_i$.*

Démonstration. Cette preuve est aussi standard (Blackburn et al., 2002). □

Prouvons ensuite la propriété relationnelle correspondante à l'axiome $(5_{T,B})$.

Proposition 4.3 : *Pour tout* $i, j \in \mathcal{N}$,

$$\mathcal{C} \models \neg T^s_{i,j} p \Rightarrow B_i \neg T^s_{i,j} p$$

si, et seulement si,

$$\forall w, u, v \in \mathcal{W}, w\mathcal{B}_i u \wedge w\mathcal{T}^s_{i,j} v \Rightarrow u\mathcal{T}^s_{i,j} v$$

Démonstration. Soient deux agents $i, j \in \mathcal{N}$.

($\Rightarrow$) Par contraposition, supposons qu'il existe $w, u, v \in \mathcal{W}$: $w\mathcal{B}_i u \wedge w\mathcal{T}^s_{i,j} v \wedge \neg(u\mathcal{T}^s_{i,j} v)$. Définissons un modèle $\mathcal{M}$ où $V(p) = \mathcal{W} \setminus \{v\}$. Puisque $V(p) = \mathcal{W} \setminus \{v\}$ et $w\mathcal{T}^s_{i,j} v$, nous avons $\mathcal{M}, w \models \neg T^s_{i,j} p$. De plus, comme $\neg(u\mathcal{T}^s_{i,j} v)$, nous avons donc $\mathcal{M}, u \models T^s_{i,j} p$. Puisque $w\mathcal{B}_i u$, nous déduisons $\mathcal{M}, w \models \neg B_i \neg T^s_{i,j} p$. Par conséquent, il existe un modèle $\mathcal{M}$ et un monde $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models \neg T^s_{i,j} p \wedge \neg B_i \neg T^s_{i,j} p$ i.e. $\mathcal{C} \not\models \neg T^s_{i,j} p \Rightarrow B_i \neg T^s_{i,j} p$.

($\Leftarrow$) Par contraposition, supposons qu'il existe $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}^s_{i,j}\}_{i,j\in\mathcal{N}}, V)$ et un monde $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models \neg T^s_{i,j} p \wedge \neg B_i \neg T^s_{i,j} p$. Ainsi, il existe $v \in \mathcal{W}$, $w\mathcal{T}^s_{i,j} v$ tel que $\mathcal{M}, v \models \neg p$ et il existe $u \in \mathcal{W}$: $w\mathcal{B}_i u$ tel que $\mathcal{M}, u \models T^s_{i,j} p$. Puisque $v \notin V(p)$ et $\forall u' \in \mathcal{W}$: $u\mathcal{T}^s_{i,j} u', \mathcal{M}, u' \models p$, nous déduisons que $\neg(u\mathcal{T}^s_{i,j} v)$.

Par conséquent, pour tout cadre $\mathcal{C}$, nous avons donc montré que $\mathcal{C} \models \neg T^s_{i,j} p \Rightarrow B_i \neg T^s_{i,j} p$ si, et seulement si, $\forall w, u, v \in \mathcal{W}, w\mathcal{B}_i u \wedge w\mathcal{T}^s_{i,j} v \Rightarrow u\mathcal{T}^s_{i,j} v$. □

Nous caractérisons les propriétés correspondantes aux axiomes $(4_{T,B})$ et (S).

Proposition 4.4 : *Pour tout* $i, j \in \mathcal{N}$ *et* $(\Box, \mathcal{R}) \in \{(T^s_{i,j}, \mathcal{T}^s_{i,j}), (B_j, \mathcal{B}_j)\}$,

$$\mathcal{C} \models T^s_{i,j} p \Rightarrow B_i \Box p$$

si, et seulement si,

$$\forall w, u, v \in \mathcal{W}, w\mathcal{B}_i u \wedge u\mathcal{R} v \Rightarrow w\mathcal{T}^s_{i,j} v$$

Démonstration. Soient $i, j \in \mathcal{N}$ deux agents et $(\Box, \mathcal{R}) \in \{(T^s_{i,j}, \mathcal{T}^s_{i,j}), (B_j, \mathcal{B}_j)\}$.

($\Rightarrow$) Par contraposition supposons un cadre $\mathcal{C}$ tel qu'il existe $w, u, v \in \mathcal{W}$: $w\mathcal{B}_i u \wedge u\mathcal{R} v$ et $\neg(w\mathcal{T}^s_{i,j} v)$. Définissons un modèle $\mathcal{M}$ où $V(p) = \mathcal{W} \setminus \{v\}$. Puisque $\neg(w\mathcal{T}^s_{i,j} v)$, nous avons donc $\mathcal{M}, w \models T^s_{i,j} p$. De plus, comme $u\mathcal{R} v$ et $\mathcal{M}, v \models \neg p$, nous déduisons que $\mathcal{M}, u \models \neg\Box p$ et comme $w\mathcal{B}_i u$, $\mathcal{M}, w \models \neg B_i \Box p$. Nous avons alors $\mathcal{M}, w \models T^s_{i,j} p \wedge \neg B_i \Box p$. Par conséquent, nous venons de prouver que $\not\models_{\mathcal{C}} T^s_{i,j} p \Rightarrow B_i \Box p$.

($\Leftarrow$) Par contraposition, supposons un cadre $\mathcal{C}$ tel que $\not\models_{\mathcal{C}} T^s_{i,j} p \Rightarrow B_i \Box p$. Ainsi, il existe un modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}^s_{i,j}\}_{i,j\in\mathcal{N}}, V)$ et un monde $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models T^s_{i,j} p \wedge \neg B_i \Box p$. Ainsi, pour tout $v' \in \mathcal{W}$: $w\mathcal{T}^s_{i,j} v'$, $\mathcal{M}, v' \models p$ et il existe $u \in \mathcal{W}$: $w\mathcal{B}_i u$, $\mathcal{M}, u \models \neg\Box p$. Par

conséquent, il existe $v \in \mathcal{W} : u\mathcal{R}v, \mathcal{M}, v \models \neg p$. Cependant, pour tout $v' \in \mathcal{W} : w\mathcal{T}^s_{i,j}v', \mathcal{M}, v' \models p$ donc $\neg(w\mathcal{T}^s_{i,j}v)$. Nous avons donc montré qu'il existe $w, u, v \in \mathcal{W}$ tel que $w\mathcal{B}_i u \wedge u\mathcal{R}v$ et $\neg(w\mathcal{T}^s_{i,j}v)$.

Par conséquent, pour tout cadre $\mathcal{C}$, nous avons donc montré que $\mathcal{C} \models T^s_{i,j}p \Rightarrow B_i\Box p$ si, et seulement si, $\forall w, u, v \in \mathcal{W}, w\mathcal{B}_i u \wedge u\mathcal{R}v \Rightarrow w\mathcal{T}^s_{i,j}v$.

□

La preuve de correction du système TB découle alors des preuves précédentes.

Théorème 4.1 - *Correction de TB* : *Le système TB est correct.*

Démonstration. (Schéma de preuve) Puisque nous avons montré dans la section précédente que les propriétés de la relation d'accessibilité preservaient la validité des axiomes de TB, il suffit alors de démontrer que la substitution, le modus ponens et la nécessitation préservent la validité. Or, en appliquant la méthode déjà présentée au chapitre 3, nous avons immédiatement ces résultats. □

4.3.2 Complétude

Pour montrer la complétude du système axiomatique par rapport au cadre TB, nous définissons la notion de modèle canonique pour TB. Enfin, nous prouvons que notre modèle canonique satisfait bien toutes les propriétés nécessaires du système TB.

Modèle canonique

Un modèle canonique pour TB permet de faire une correspondance entre un théorème prouvé dans notre système axiomatique et sa validité dans le cadre TB.

Définition 4.3 - *Modèle canonique de TB* : *Soit* $\mathcal{M}^c = (\mathcal{W}^c, \{\mathcal{B}^c_i\}_{i\in\mathcal{N}}, \{\mathcal{T}^c_{i,j}\}_{i,j\in\mathcal{N}}, V^c)$ *un modèle de Kripke tel que :*

- $\mathcal{W}^c$ *est un ensemble de mondes possibles où chaque monde est un ensemble de formules maximal TB-consistant ;*
- $\{\mathcal{B}^c_i\}_{i\in\mathcal{N}}$ *est un ensemble de relations binaires tel que :*

$$\forall i \in \mathcal{N}, \forall w, v \in \mathcal{W}^c : w\mathcal{B}^c_i v \text{ si, et seulement si, } B_i\phi \in w \Rightarrow \phi \in v$$

- $\{\mathcal{T}^c_{i,j}\}_{i,j\in\mathcal{N}}$ *est un ensemble de relations binaires tel que :*

$$\forall i, j \in \mathcal{N}, \forall w, v \in \mathcal{W}^c : w\mathcal{T}^c_{i,j}v \text{ si, et seulement si, } T^s_{i,j}\phi \in w \Rightarrow \phi \in v$$

- $V^c : \mathcal{P} \to 2^{\mathcal{W}}$ *est une fonction d'interprétation :*

$$\forall p \in \mathcal{P}, w \in V^c(p) \text{ si, et seulement si, } p \in w$$

Nous considérons les notations suivantes :

$$\forall i,j \in \mathcal{N}, \forall w \in \mathcal{W}^c, \mathcal{T}^*_{i,j}(w) := \{\phi \in \mathcal{L}_{TB} | T^s_{i,j}\phi \in w\}$$

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W}^c, \mathcal{B}^*_i(w) := \{\phi \in \mathcal{L}_{TB} | B_i\phi \in w\}$$

Avec ces notations, les relations d'accessibilité du modèle canonique $\mathcal{T}^c_{i,j}$ et $\mathcal{B}^c_i$ deviennent :

$$\forall i,j \in \mathcal{N}, \forall w,v \in \mathcal{W}^c : w\mathcal{T}^c_{i,j}v \text{ si, et seulement si, } \mathcal{T}^*_{i,j}(w) \subseteq v$$

$$\forall i \in \mathcal{N}, \forall w,v \in \mathcal{W}^c : w\mathcal{B}^c_i v \text{ si, et seulement si, } \mathcal{B}^*_i(w) \subseteq v$$

Dans la suite, nous considérons un modèle canonique $\mathcal{M}^c = (\mathcal{W}^c, \{\mathcal{B}^c_i\}_{i\in\mathcal{N}}, \{\mathcal{T}^c_{i,j}\}_{i,j\in\mathcal{N}}, V^c)$.

Lemme 4.1 : *Soit $i,j \in \mathcal{N}$ et $\phi \in \mathcal{L}_{TB}$,*

— $\forall w \in \mathcal{W}^c : \neg T^s_{i,j}\phi \in w \Rightarrow \mathcal{T}^*_{i,j}(w) \cup \{\neg\phi\}$ *est TB-consistant ;*

— $\forall w \in \mathcal{W}^c : \neg\mathcal{B}_i\phi \in w \Rightarrow B^*_i(w) \cup \{\neg\phi\}$ *est TB-consistant.*

Démonstration. Soient $w \in \mathcal{W}^c$, $i,j \in \mathcal{N}$, $(\Box, \mathcal{R}) \in \{(T^s_{i,j}, \mathcal{T}^c_{i,j}), (B_i, \mathcal{B}^c_i)\}$. Supposons par contraposition que $\mathcal{R}^*(w) \cup \{\neg\phi\}$ est TB-inconsistant. Alors, il existe $n \in \mathbb{N}$ et $\psi_1, \ldots, \psi_n \in \mathcal{R}^*(w)$ telles que :

1. $\vdash \neg(\bigwedge_{k=1}^n \psi_k \wedge \neg\phi)$
2. $\vdash \neg \bigwedge_{k=1}^n \psi_k \vee \neg\neg\phi$
3. $\vdash \bigwedge_{k=1}^n \psi_k \Rightarrow \phi$
4. $\vdash \Box(\bigwedge_{k=1}^n \psi_k \Rightarrow \phi)$
5. $\vdash (\Box \bigwedge_{k=1}^n \psi_k \Rightarrow \Box\phi)$
6. $\vdash (\bigwedge_{k=1}^n \Box\psi_k \Rightarrow \Box\phi)$
7. $\vdash \neg(\bigwedge_{k=1}^n \Box\psi_k \wedge \neg\Box\phi)$

Par conséquent, $\{\Box\psi_1, \ldots, \Box\psi_n, \neg\Box\phi\}$ est TB-inconsistant. Cependant, $\forall k \in \{1,\ldots,n\}, \psi_k \in \mathcal{R}^*(w)$ si, et seulement si, $\Box\psi_k \in w$ et w est un ensemble TB-consistent. Ainsi, $\bigwedge_{k=1}^n \Box\psi_k \in w$ (MC3') et alors $\{\Box\psi_1, \ldots, \Box\psi_n\}$ est TB-inconsistent. Comme $\{\Box\psi_1, \ldots, \Box\psi_n\} \cup \{\neg\Box\phi\}$ est TB-inconsistant, $\neg\Box\phi$ n'appartient pas à un ensemble maximal TB-consistent. Supposons par l'absurde que $\neg\Box\phi \in w$, nous aurions aussi $\bigwedge_{k=1}^n \Box\psi_k \wedge \neg\Box\phi \in w$ (par MCS3') et donc $\{\Box\psi_1, \ldots, \Box\psi_n, \neg\Box\phi\}$ serait TB-consistant ce qui est contradictoire. Par conséquent, $\neg\Box\phi \notin w$.

Ainsi, nous avons prouvé que si $\neg\Box\phi \in w$, alors $\mathcal{R}^*(w) \cup \{\neg\phi\}$ est TB-consistant. □

Nous avons besoin d'un troisième lemme pour montrer la complétude de notre système.

Lemme 4.2 : *Soit $w \in \mathcal{W}^c$ et $\phi \in \mathcal{L}_{T,B}$. $\mathcal{M}^c, w \models \phi$ si, et seulement si, $\phi \in w$.*

Démonstration. Raisonnons par récurrence sur le degré d'une formule.

(Initialisation) Si $\phi \in \mathcal{L}_{TB}$ est une formule de degré 0, i.e. il existe $p \in \mathcal{P}, \phi = p$. Par définition du modèle canonique, nous avons $\forall w \in \mathcal{W}^c, w \in V(p)$ si, et seulement si, $p \in w$.

(Hérédité) Supposons que pour toute formule $\phi \in \mathcal{L}_{TB}$ de degré strictement inférieur à $n \in \mathbb{N}^*$, nous avons : $\forall w \in \mathcal{W}^c, \mathcal{M}^c, w \models \phi$ si, et seulement si, $\forall w \in \mathcal{W}^c, \phi \in w$

Soit $\psi, \theta \in \mathcal{L}_{TB}$ telle que $max(deg(\psi), deg(\theta)) = n-1$. Nous avons pour tout monde $w \in \mathcal{W}^c$, $\mathcal{M}^c, w \models \psi$ ssi $\psi \in w$ et $\mathcal{M}^c, w \models \theta$ ssi $\theta \in w$.

De plus, nous avons que $\mathcal{M}^c, w \models \neg\psi$ ssi $\mathcal{M}^c, w \not\models \psi$ ssi $\psi \notin w$. Puis, $\mathcal{M}^c, w \models \psi \wedge \theta$ ssi $\mathcal{M}^c, w \models \psi$ et $\mathcal{M}^c, w \models \theta$ ssi $\psi \in w$ et $\theta \in w$ ssi $\psi \wedge \theta \in w$ (MC3'). Ensuite, $\mathcal{M}^c, w \models \psi \vee \theta$ ssi $\mathcal{M}^c, w \models \psi$ ou $\mathcal{M}^c, w \models \theta$ ssi $\psi \in w$ ou $\theta \in w$ ssi $\psi \vee \theta \in w$ (MC3). Enfin, $\mathcal{M}^c, w \models \psi \Rightarrow \theta$ ssi $\mathcal{M}^c, w \models \neg\psi$ ou $\mathcal{M}^c, w \models \theta$ ssi $\psi \notin w$ ou $\theta \in w$ ssi $\psi \Rightarrow \theta \in w$.

Prenons $(\mathcal{R}, \Box) \in \{(\mathcal{B}_i, B_i), (\mathcal{T}^s_{i,j}, T^s_{i,j})\}$ et considérons un monde $w \in \mathcal{W}^c$.

($\Rightarrow$) Supposons par contraposition que $\Box\psi \notin w$. Comme w est maximal TB-consistant, nous avons $\neg\Box\psi \in w$. Par le lemme 4.1, nous avons $\mathcal{R}^*(w) \cup \{\neg\psi\}$ est TB-consistant et donc, par le théorème de Lindenbaum, il existe $v \in \mathcal{W}^c : \mathcal{R}^*(w) \cup \{\neg\psi\} \subseteq v$ et v est maximal TB-consistant. Nous avons donc $\neg\psi \in v$ et, par définition de $\mathcal{R}^c$, nous avons $w\mathcal{R}^c v$ et donc $\psi \notin v$. Ainsi, par hypothèse de récurrence, nous déduisons $\mathcal{W}^c, v \not\models \psi$. Or, puisqu'il existe $v \in \mathcal{W}^c : w\mathcal{R}^c v : v \models \neg\psi$, nous avons $\mathcal{M}^c, w \models \neg\Box\psi$, c'est-à-dire $\mathcal{M}^c, w \not\models \Box\psi$.

($\Leftarrow$) Par contraposition, supposons que $\mathcal{M}^c, w \not\models \Box\psi$, i.e. $\mathcal{M}^c, w \models \neg\Box\psi$. Donc, il existe $v \in \mathcal{W} : w\mathcal{R}^c v, \mathcal{M}^c, v \models \neg\psi$. Ainsi, $\mathcal{M}^c, v \not\models \phi$ et, par hypothèse de récurrence, nous avons $\phi \notin v$. Or, puisque $\phi \notin v$, par définition de $\mathcal{R}^c$, nous avons donc $\Box\phi \notin w$.

(Conclusion) Nous avons donc montré par récurrence que :

$$\forall \phi \in \mathcal{L}_{TB}, \forall w \in \mathcal{W}^c : \mathcal{M}^c, w \models \phi \text{ si, et seulement si, } \phi \in w$$

□

Nous prouvons ensuite la connexion entre notre modèle canonique et les formules prouvées du système.

Théorème 4.2 - *Lemme de vérité* : *Soit $\phi \in \mathcal{L}_{TB}$. $\mathcal{M}^c \models \phi$ si, et seulement si, $\vdash \phi$.*

Démonstration.

1. Par définition $\mathcal{M}^c \models \phi$ si, et seulement si, $\forall w \in \mathcal{W}^c : \mathcal{M}^c, w \models \phi$;
2. Par le lemme 4.2 $\forall w \in \mathcal{W}^c : \mathcal{M}^c, w \models \phi$ si, et seulement si, $\forall w \in \mathcal{W}^c, \phi \in w$;
3. Par MC5, $\forall w \in \mathcal{W}^c, \phi \in w$ si, et seulement si, $\vdash \phi$;
4. Par conséquent, $\mathcal{M}^c \models \phi$ si, et seulement si, $\vdash \phi$.

□

Preuve de complétude

Nous avons jusqu'à présent montré que les résultats principaux autour des modèles canoniques étaient vérifiés pour le cadre TB. En particulier, ces résultats nous ont permis de montrer que toute formule valide dans le modèle canonique était une formule valide de TB, et réciproquement. Nous devons désormais démontrer que le modèle canonique satisfait toutes les contraintes sémantiques de TB. En démontrant qu'il satisfait toutes ses contraintes, la preuve de complétude du système TB devient alors immédiate.

Lemme 4.3 : *Soit* $\mathcal{M}^c = (\mathcal{W}^c, \{\mathcal{B}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}^c\}_{i,j\in\mathcal{N}}, i^c)$ *un modèle canonique de TB. Ce modèle canonique* $\mathcal{M}^c$ *est tel que :*

1. $\forall i,j \in \mathcal{N}, \mathcal{T}_{i,j}^c$ *est sérielle ;*
2. $\forall i,j \in \mathcal{N}, \forall w,u,v \in \mathcal{W}^c, w\mathcal{B}_i^c u \wedge w\mathcal{T}_{i,j}^c v \Rightarrow u\mathcal{T}_{i,j}^c v$;
3. $\forall i,j \in \mathcal{N}, \forall w,u,v \in \mathcal{W}^c, w\mathcal{B}_i^c u \wedge u\mathcal{T}_{i,j}^c v \Rightarrow w\mathcal{T}_{i,j}^c v$;
4. $\forall i,j \in \mathcal{N}, \forall w,u,v \in \mathcal{W}^c, w\mathcal{B}_i^c u \wedge u\mathcal{B}_j^c v \Rightarrow w\mathcal{T}_{i,j}^c v$;
5. $\forall i \in \mathcal{N}, \mathcal{B}_i^c$ *est transitive, sérielle et euclidienne.*

Démonstration. Soient $i,j \in \mathcal{N}$, $w \in \mathcal{W}^c$ et $T_{i,j}^s\phi \in w$.

(1) Il s'agit d'une preuve standard pour les systèmes KD (Blackburn et al., 2002).

(2) Soient $i,j \in \mathcal{N}$. Pour tout $w,u,v \in \mathcal{W}^c$: $w\mathcal{B}_i^c u \wedge w\mathcal{T}_{i,j}^c v$ et $\phi \notin v$. Par MC2, $\neg\phi \in v$ et, puisque $w\mathcal{T}_{i,j}^c v$, nous avons $\neg T_{i,j}^s\phi \in w$. Cependant $\vdash \neg T_{i,j}^s\phi \Rightarrow B_i\neg T_{i,j}^s\phi$. Ainsi, par MC5, nous avons $\neg T_{i,j}^s\phi \Rightarrow B_i\neg T_{i,j}^s\phi \in w$. De plus, par MC4, nous déduisons que $B_i\neg T_{i,j}^s\phi \in w$ et puisque $w\mathcal{B}_i^c u$, nous avons $\neg T_{i,j}^s\phi \in u$. Donc, par MC2, $T_{i,j}^s\phi \notin u$. Nous venons donc de montrer par contraposition que $T_{i,j}^s\phi \in u \Rightarrow \phi \in v$, et donc $u\mathcal{T}_{i,j}^c v$. Ainsi, le modèle canonique satisfait :

$$\forall i,j \in \mathcal{N}, \forall w,u,v \in \mathcal{W}^c, w\mathcal{B}_i^c u \wedge w\mathcal{T}_{i,j}^c v \Rightarrow u\mathcal{T}_{i,j}^c v$$

(3) Soient $i,j \in \mathcal{N}$. Pour tout $w,u,v \in \mathcal{W}^c$: $w\mathcal{B}_i^c u \wedge u\mathcal{T}_{i,j}^c v$ et $T_{i,j}^s\phi \in w$. Cependant $\vdash T_{i,j}^s\phi \Rightarrow B_iT_{i,j}^s\phi$. Ainsi, par MC5, $T_{i,j}^s\phi \Rightarrow B_iT_{i,j}^s\phi \in w$ et, par MC4, $B_iT_{i,j}^s\phi \in w$. Par conséquent, $T_{i,j}^s\phi \in u$, et donc $\phi \in v$. Ainsi, par définition de $\mathcal{T}_{i,j}^c$, nous avons $w\mathcal{T}_{i,j}^c v$. Par conséquent, le modèle canonique vérifie la propriété :

$$\forall i,j \in \mathcal{N}, \forall w,u,v \in \mathcal{W}^c, w\mathcal{B}_i^c u \wedge u\mathcal{T}_{i,j}^c v \Rightarrow w\mathcal{T}_{i,j}^c v$$

(4) Soient $i,j \in \mathcal{N}$. Pour tout $w,u,v \in \mathcal{W}^c$: $w\mathcal{B}_i^c u \wedge u\mathcal{B}_j^c v$ et $T_{i,j}^s\phi \in w$. Cependant $\vdash T_{i,j}^s\phi \Rightarrow B_iB_j\phi$. Ainsi, par MC5, $T_{i,j}^s\phi \Rightarrow B_iB_j\phi \in w$ et, par MC4, $B_iB_j\phi \in w$. Par conséquent, $B_j\phi \in u$ et donc $\phi \in v$. Ainsi, par définition de $\mathcal{T}_{i,j}^c$, nous avons $w\mathcal{T}_{i,j}^c v$, i.e. nous avons prouvé que :

$$\forall i,j \in \mathcal{N}, \forall w,u,v \in \mathcal{W}^c, w\mathcal{B}_i^c u \wedge u\mathcal{B}_j^c v \Rightarrow w\mathcal{T}_{i,j}^c v$$

(5) Il s'agit d'une preuve standard pour les systèmes KD45 (Blackburn et al., 2002). □

Théorème 4.3 - *Complétude de TB* : *Le système TB est complet.*

Démonstration. Par raisonnement par synthèse, nous avons :

1. $\mathcal{C} \models \phi \Longrightarrow \mathcal{M}^c \models \phi$
2. $\mathcal{M}^c \models \phi \Longleftrightarrow \vdash \phi$

Par conséquent, $\mathcal{C} \models \phi \Longrightarrow \vdash \phi$. □

Nous avons donc montré que le système TB était correct et complet.

4.3.3 Propriétés fortes du cadre TB

Dans cette section, nous démontrons quelques propriétés du cadre TB comme les théorèmes de déduction, la correction forte et la complétude forte du cadre.

Théorèmes de déduction

Tout d'abord, la TB-déductibilité possède les propriétés de réflexivité, de transitivité et d'affaiblissement à gauche présentées en section 2.1.1. Ces propriétés sont évidentes à démontrer dans TB. Il est de même standard de démontrer que les théorèmes de déduction sont vérifiés.

Théorème 4.4 - *Théorèmes de déduction* : *Soient* Γ *un ensemble de formules prouvées dans TB,* ϕ *et* ψ *deux théorèmes.*

$$(1) \quad \Gamma \cup \{\psi\} \vdash \phi \text{ si, et seulement si, } \Gamma \vdash \psi \Rightarrow \phi$$

$$(2) \quad \Gamma \cup \{\psi\} \models \phi \text{ si, et seulement si, } \Gamma \models \psi \Rightarrow \phi$$

Démonstration. Soient Γ un ensemble de formules prouvées dans TB, ϕ et ψ deux théorèmes. Démontrons le sens ($\Rightarrow$) et supposons donc que $\Gamma \cup \{\psi\} \vdash \phi$. Par définition de la TB-déductibilité, nous avons donc qu'il existe $\Sigma = \{\psi_1, \ldots, \psi_n\}, \Sigma \subseteq \Gamma \cup \{\psi\}$ tel que :

$$\vdash \bigwedge_{i \in \{1,\ldots,n\}} \psi_i \Rightarrow \phi$$

Nous avons donc deux cas à considérer lorsque $\psi \in \Sigma$ et lorsque $\psi \notin \Sigma$.

Si $\psi \in \Sigma$, alors il existe $i \in \{1, \ldots, n\}$ tel que $\psi = \psi_i$. Donc $\vdash (\bigwedge_{k \in \{1,\ldots,n\} \setminus \{i\}} \psi_k) \wedge \psi \Rightarrow \phi$ par commutativité et associativité du $\wedge$. Puis, $\vdash \bigwedge_{k \in \{1,\ldots,n\} \setminus \{i\}} \psi_k \Rightarrow (\psi \Rightarrow \phi)$. Or pour tout $k \in \{1, \ldots, n\} \setminus \{i\}, \psi_k \in \Sigma$ et donc $\psi_k \in \Gamma$ par inclusion. Nous avons donc prouvé dans ce cas que $\Gamma \vdash \psi \Rightarrow \phi$.

Si $\psi \notin \Sigma$, alors pour tout $i \in \{1, \ldots, n\}, \psi_i \neq \psi$. Nous avons donc dans ce cas que :

$$\vdash \bigwedge_{k \in \{1,\ldots,n\}} \psi_k \Rightarrow \phi$$

Or comme $\vdash \phi \Rightarrow (\psi \Rightarrow \phi)$ est un axiome du CP, nous déduisons immédiatement que :

$$\vdash \bigwedge_{k \in \{1,\ldots,n\}} \psi_k \Rightarrow (\psi \Rightarrow \phi)$$

Par conséquent, puisque pour tout $k \in \{1, \ldots, n\}, \psi_k \in \Sigma$ et donc $\psi_k \in \Gamma$ par inclusion. Nous avons donc prouvé dans cet autre cas que $\Gamma \vdash \psi \Rightarrow \phi$.

($\Leftarrow$) Prouvons désormais la réciproque et supposons que $\Gamma \vdash \psi \Rightarrow \phi$.
Il existe donc $\Sigma = \{\psi_1, \ldots, \psi_n\}$, $\Sigma \subseteq \Gamma$ tel que :

$$\vdash \bigwedge_{i \in \{1,\ldots,n\}} \psi_i \Rightarrow (\psi \Rightarrow \phi)$$

Donc $\vdash \bigwedge_{i \in \{1,\ldots,n\}} \psi_i \wedge \psi \Rightarrow \phi$ est aussi un théorème. Ainsi, nous déduisons que $\Sigma \cup \{\psi\} \vdash \phi$. Or comme $\Sigma \subseteq \Gamma$, par affaiblissement à gauche, nous avons immédiatement que $\Gamma \cup \{\psi\} \vdash \phi$.

La version sémantique (2) peut être prouvée par double équivalence :
$\Gamma \models \psi \Rightarrow \phi$ ssi $\forall \mathcal{M}$: si $\mathcal{M} \models \Gamma$, alors $\mathcal{M} \models \psi \Rightarrow \phi$ ssi $\forall \mathcal{M} : \mathcal{M} \models \Gamma \Rightarrow (\psi \Rightarrow \phi)$ ssi $\forall \mathcal{M} : \mathcal{M} \models \neg\Gamma \vee \neg\psi \vee \phi$ ssi $\forall \mathcal{M} : \mathcal{M} \models \neg(\Gamma \wedge \psi) \vee \phi$ ssi $\forall \mathcal{M} : \mathcal{M} \models (\Gamma \wedge \psi) \Rightarrow \phi$ ssi $\forall \mathcal{M}$: si $\mathcal{M} \models \Gamma \cup \{\psi\}$, alors $\mathcal{M} \models \phi$ ssi $\Gamma \cup \{\psi\} \models \phi$.

□

Correction forte

La correction forte est une conséquence quasiment immédiate de la correction et de l'affaiblissement sémantique que nous redémontrons dans la preuve du théorème.

Théorème 4.5 - *Correction forte de TB :* *Soient Γ un ensemble de formules prouvées dans TB, ϕ un théorème de TB.*

Si $\Gamma \vdash \phi$ alors $\Gamma \models \phi$

Démonstration. Soient Γ un ensemble de formules prouvées dans TB, ϕ un théorème de TB.

Supposons que $\Gamma \vdash \phi$, donc il existe $\psi_1, \ldots, \psi_n \in \Gamma$ tq $\vdash \psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \phi$. Donc par correction, nous avons $\models \psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \phi$ i.e. $\models \neg\psi_1 \vee \ldots \vee \neg\psi_n \vee \phi$. Donc $\models \neg\psi_1 \vee \ldots \vee \neg\psi_n \vee \neg \bigwedge_{\theta \in \Gamma} \theta \vee \phi$.

Ainsi, en appliquant la loi de De Morgan et en regroupant les termes, nous avons que :

$$\models \neg \bigwedge_{\theta \in \Gamma \cup \{\psi_1, \ldots, \psi_n\}} \theta \vee \phi$$

Or comme $\{\psi_1, \ldots, \psi_n\} \subseteq \Gamma$, nous avons $\Gamma \cup \{\psi_1, \ldots, \psi_n\} = \Gamma$ et donc $\models \neg \bigwedge_{\theta \in \Gamma} \theta \vee \phi$ (affaiblissement sémantique). Par conséquent, nous avons $\forall \mathcal{M}$, si $\mathcal{M} \models \Gamma$, alors $\mathcal{M} \models \phi$. Nous venons donc de montrer que $\Gamma \models \phi$.

□

Complétude forte

Le système TB est fortement complet.

Théorème 4.6 - *Complétude forte de TB* : *Le système TB est* fortement complet.

Démonstration. Par contraposition, considérons $\Gamma \subset \mathcal{L}_{TB}$ un ensemble de formules tel que $\Gamma \nvdash \phi$. Nous avons que $\Gamma \cup \{\neg\phi\}$ est un ensemble TB-consistant de formules [6]. Comme $\Gamma \cup \{\neg\phi\}$ est consistant, par le lemme de Lindenbaum, il existe un ensemble maximal TB-consistant Γ' tel que $\Gamma \cup \{\neg\phi\} \subseteq \Gamma'$ et $\mathcal{M}^c \models \Gamma'$ avec $\mathcal{M}^c$ un modèle canonique. Ainsi, $\mathcal{M}^c \models \Gamma \cup \{\neg\phi\}$. Donc il existe un modèle $\mathcal{M}$ tel que $\mathcal{M} \models \Gamma \cup \{\neg\phi\}$. Donc $\mathcal{M} \models \Gamma$ et $\mathcal{M} \models \neg\phi$. Ainsi, ce modèle $\mathcal{M}$ est tel que $\mathcal{M}, \Gamma \not\models \phi$. En conclusion, nous venons donc de montrer que le système TB est fortement complet.

□

Nous avons donc montré que le système TB est correct, complet, fortement correct et fortement complet. Dans la suite, nous nous intéressons aux propriétés logiques de la confiance en la sincérité que nous pouvons déduire dans ce système logique.

4.4 Confiance individuelle et confiance collective

Si la notion de confiance a certaines propriétés logiques déjà axiomatisées, comme l'axiome de sincérité, ou encore l'introspection de la confiance par rapport à ses croyances, ces propriétés ne sont pas les seules. Dans cette section, nous étudions dans un premier temps, des propriétés comme la distributivité de la confiance en la sincérité, ou encore nous montrons pourquoi la confiance en la sincérité n'est pas transitive. Par la même occasion, nous montrons que l'axiome D est en réalité déductible par l'axiome de sincérité. Dans un second temps, nous étendons la confiance en la sincérité à la notion de confiance partagée par un groupe.

6. La preuve est immédiate avec une preuve par l'absurde telle qu'elle a été faite en section 3.2.3.

4.4.1 Des propriétés de distributivité

Les premières propriétés intéressantes à considérer sont liées au caractère normal de la modalité de confiance.

Proposition 4.5 : *Soient $i, j \in \mathcal{N}$ deux agents.*

1. $\vdash T^s_{i,j}\phi \wedge T^s_{i,j}\psi \equiv T^s_{i,j}(\phi \wedge \psi) \quad (\wedge_T)$
2. $\vdash (T^s_{i,j}\phi \vee T^s_{i,j}\psi) \Rightarrow T^s_{i,j}(\phi \vee \psi) \quad (\vee_T)$

Démonstration. Puisque $T^s_{i,j}$ est une modalité normale, il en résulte immédiatement cette propriété par la proposition démontrée et issue de l'exemple 2.4. La preuve est standard. □

La première propriété traduit le fait qu'un agent i fait confiance à la sincérité d'un autre agent j sur l'ensemble des propositions énoncées par l'agent j et sur lesquelles l'agent i a accordé sa confiance. La seconde propriété signifie que si l'agent i accorde sa confiance à un agent j sur une certaine proposition ϕ, ou s'il l'accorde sur une autre proposition ψ, alors l'agent i a confiance en la sincérité de l'agent j sur la disjonction des deux propositions. Cependant, la réciproque ne peut être considérée. En effet, lorsqu'un agent i fait confiance à la sincérité d'un agent j pour deux choix possibles $\phi \vee \psi$, si l'agent j énonce le choix ϕ ou s'il énonce le choix ψ à l'agent i, l'agent i peut ne pas faire confiance à j pour ϕ et ne pas non plus lui faire confiance pour ψ. Cet agent j pourrait par exemple être en train de manipuler l'agent i sur un choix spécifique.

4.4.2 Des propriétés liées avec la croyance

Les réciproques des axiomes (4_{TB}) et (5_{TB}) sont prouvées dans le système TB.

Proposition 4.6 : *Soient $i, j \in \mathcal{N}$ deux agents.*

1. $\vdash B_i T^s_{i,j}p \Rightarrow T^s_{i,j}p \quad (C4_{TB})$
2. $\vdash B_i \neg T^s_{i,j}p \Rightarrow \neg T^s_{i,j}p \quad (C5_{TB})$

Démonstration. Soient deux agents $i, j \in \mathcal{N}$. Nous prouvons la première propriété :

1. $\vdash \neg T^s_{i,j}p \Rightarrow B_i \neg T^s_{i,j}p \quad (5_{TB})$
2. $\vdash B_i \neg T^s_{i,j}p \Rightarrow \neg B_i T^s_{i,j}p \quad (D_B)$
3. $\vdash \neg T^s_{i,j}p \Rightarrow B_i \neg T^s_{i,j}p \Rightarrow \neg B_i T^s_{i,j}p$
4. $\vdash (\neg T^s_{i,j}p \Rightarrow B_i \neg T^s_{i,j}p \Rightarrow \neg B_i T^s_{i,j}p) \Rightarrow (\neg T^s_{i,j}p \Rightarrow B_i \neg T^s_{i,j}p) \Rightarrow (\neg T^s_{i,j}p \Rightarrow \neg B_i T^s_{i,j}p)$
5. $\vdash \neg T^s_{i,j}p \Rightarrow \neg B_i T^s_{i,j}p$
6. $\vdash (\neg T^s_{i,j}p \Rightarrow \neg B_i T^s_{i,j}p) \Rightarrow (B_i T^s_{i,j}p \Rightarrow T^s_{i,j}p)$
7. $\vdash B_i T^s_{i,j}p \Rightarrow T^s_{i,j}p$

Nous prouvons la seconde propriété :

1. $\vdash B_i \neg T^s_{i,j}p \Rightarrow \neg B_i T^s_{i,j}p \quad (D_B)$

2. $\vdash T^s_{i,j}p \Rightarrow B_i T^s_{i,j}p \quad (4_{TB})$
3. $\vdash (T^s_{i,j}p \Rightarrow B_i T^s_{i,j}p) \Rightarrow (\neg B_i T^s_{i,j}p \Rightarrow \neg T^s_{i,j}p)$
4. $\vdash \neg B_i T^s_{i,j}p \Rightarrow \neg T^s_{i,j}p$
5. $\vdash B_i \neg T^s_{i,j}p \Rightarrow \neg B_i T^s_{i,j}p \Rightarrow \neg T^s_{i,j}p$
6. $\vdash (B_i \neg T^s_{i,j}p \Rightarrow \neg B_i T^s_{i,j}p \Rightarrow \neg T^s_{i,j}p) \Rightarrow (B_i \neg T^s_{i,j}p \Rightarrow \neg B_i T^s_{i,j}p) \Rightarrow (B_i \neg T^s_{i,j}p \Rightarrow \neg T^s_{i,j}p)$
7. $\vdash B_i \neg T^s_{i,j}p \Rightarrow \neg T^s_{i,j}p$

□

Ces deux propriétés soulignent (1) que quand il est le cas qu'un agent croit qu'il fait confiance, alors il est bien le cas qu'il fait confiance ; (2) quand il est le cas qu'il croit qu'il ne fait pas confiance, alors il est le cas qu'il ne fait pas confiance. Enfin, nous considérons une dernière propriété liée à la croyance.

Proposition 4.7 : *Pour tout agent* $i, j \in \mathcal{N}$,

$$\vdash B_i B_j \phi \Rightarrow \neg T^s_{i,j} \neg \phi$$

Démonstration. Soient $i, j \in \mathcal{N}$ deux agents.

1. $\vdash B_j \phi \Rightarrow \neg B_j \neg \phi$
2. $\vdash B_i(B_j \phi \Rightarrow \neg B_j \neg \phi)$
3. $\vdash B_i(B_j \phi \Rightarrow \neg B_j \neg \phi) \Rightarrow (B_i B_j \phi \Rightarrow B_i \neg B_j \neg \phi)$
4. $\vdash B_i B_j \phi \Rightarrow B_i \neg B_j \neg \phi$
5. $\vdash B_i \neg B_j \neg \phi \Rightarrow \neg B_i B_j \neg \phi$
6. $\vdash B_i B_j \phi \Rightarrow B_i \neg B_j \neg \phi \Rightarrow \neg B_i B_j \neg \phi$
7. $\vdash T^s_{i,j} \neg \phi \Rightarrow B_i B_j \neg \phi$
8. $\vdash (T^s_{i,j} \neg \phi \Rightarrow B_i B_j \neg \phi) \Rightarrow (\neg B_i B_j \neg \phi \Rightarrow \neg T^s_{i,j} \neg \phi)$
9. $\vdash \neg B_i B_j \neg \phi \Rightarrow \neg T^s_{i,j} \neg \phi$
10. $\vdash (B_i B_j \phi \Rightarrow B_i \neg B_j \neg \phi \Rightarrow \neg B_i B_j \neg \phi) \Rightarrow (B_i B_j \phi \Rightarrow B_i \neg B_j \neg \phi) \Rightarrow (B_i B_j \phi \Rightarrow \neg B_i B_j \neg \phi)$
11. $\vdash B_i B_j \phi \Rightarrow \neg B_i B_j \neg \phi$
12. $\vdash B_i B_j \phi \Rightarrow \neg B_i B_j \neg \phi \Rightarrow \neg T^s_{i,j} \neg \phi$
13. $\vdash (B_i B_j \phi \Rightarrow \neg B_i B_j \neg \phi \Rightarrow \neg T^s_{i,j} \neg \phi) \Rightarrow (B_i B_j \phi \Rightarrow \neg B_i B_j \neg \phi) \Rightarrow (B_i B_j \phi \Rightarrow \neg T^s_{i,j} \neg \phi)$
14. $\vdash B_i B_j \phi \Rightarrow \neg T^s_{i,j} \neg \phi$

□

Cette dernière propriété énonce que si un agent i croit qu'un autre agent j croit une certaine proposition ϕ, alors il n'est pas le cas que l'agent i puisse faire confiance à l'agent j pour le contraire. Une application directe de cette propriété permet alors de déduire sur quelles propositions un agent n'a pas confiance.

4.4.3 Des propriétés de pseudo-transitivité

Des travaux ont pointé du doigt le fait que la confiance pouvait ne pas être transitive (Christianson and Harbison, 1997) (cf. section 2.2.1). La confiance en la sincérité ne l'est pas. Par transitivité, nous entendons que nous n'avons pas le théorème $T^s_{i,j}T^s_{j,k}\phi \Rightarrow T^s_{i,k}\phi$. En effet, ce n'est pas parce qu'un agent i fait confiance en la sincérité d'un agent j quand j déclare qu'il a confiance en la sincérité d'un autre agent k que l'agent i fait nécessairement confiance à la sincérité de k pour la même proposition. L'agent j peut être sincère lorsqu'il donne son avis mais peut en réalité avoir tort sur la sincérité de l'agent k sur ϕ. Cependant, nous avons une propriété de pseudo-transitivité sur les croyances des agents comme le montre la proposition suivante.

Proposition 4.8 : *Pour tout agent* $i, j, k \in \mathcal{N}$ *:*

$$\vdash T^s_{i,j}T^s_{j,k}\phi \Rightarrow B_iB_jB_k\phi$$

Démonstration. Soient $i, j, k \in \mathcal{N}$ trois agents.

1. $\vdash T^s_{i,j}T^s_{j,k}\phi \Rightarrow B_iB_jT^s_{j,k}\phi$
2. $\vdash T^s_{j,k}\phi \Rightarrow B_jB_k\phi$
3. $\vdash B_jT^s_{j,k}\phi \Rightarrow T^s_{j,k}\phi$
4. $\vdash B_i(B_jT^s_{j,k}\phi \Rightarrow T^s_{j,k}\phi)$
5. $\vdash B_i(B_jT^s_{j,k}\phi \Rightarrow T^s_{j,k}\phi) \Rightarrow B_iB_jT^s_{j,k}\phi \Rightarrow B_iT^s_{j,k}\phi$
6. $\vdash B_iB_jT^s_{j,k}\phi \Rightarrow B_iT^s_{j,k}\phi$
7. $\vdash T^s_{i,j}T^s_{j,k}\phi \Rightarrow B_iB_jB_k\phi$

□

4.4.4 Une propriété déduite du système : l'axiome D

L'axiome D du système de preuve est une conséquence de l'axiome de sincérité. Cela signifie que le théorème $\vdash T^s_{i,j}\phi \Rightarrow \neg T^s_{i,j}\neg\phi$ n'a pas besoin d'être considéré comme un axiome du système de preuve. Ce dernier peut être déduit à partir de l'axiome de sincérité.

Proposition 4.9 : *L'axiome D est une conséquence de l'axiome de sincérité.*

Démonstration.

1. $\vdash T^s_{i,j}\phi \Rightarrow B_iB_j\phi$
2. $\vdash B_iB_j\phi \Rightarrow \neg T^s_{i,j}\neg\phi$
3. $\vdash T^s_{i,j}\phi \Rightarrow B_iB_j\phi \Rightarrow \neg T^s_{i,j}\neg\phi$
4. $\vdash (T^s_{i,j}\phi \Rightarrow B_iB_j\phi \Rightarrow \neg T^s_{i,j}\neg\phi) \Rightarrow (T^s_{i,j}\phi \Rightarrow B_iB_j\phi) \Rightarrow (T^s_{i,j}\phi \Rightarrow \neg T^s_{i,j}\neg\phi)$

5. $\vdash T^s_{i,j}\phi \Rightarrow \neg T^s_{i,j}\neg\phi$

□

Le système de preuve résumé par la figure C.1 peut donc être simplifié en retirant l'axiome ($D_{T^s_{i,j}}$). Cette proposition signifie que la contrainte de sérialité imposée par le cadre TB sur la relation $\mathcal{T}_{i,j}$ est redondante. Elle est une conséquence de la contrainte de sincérité. En effet, puisque les relations $\mathcal{B}_i$ et $\mathcal{B}_j$ sont sérielles, comme $\mathcal{T}_{i,j}$ est transitive par rapport à ces deux relations, elle est donc aussi sérielle.

4.4.5 Confiance en la sincérité de soi-même

Si nous avons présenté la confiance entre deux agents $i, j \in \mathcal{N}$, nous ne nous sommes pas intéressés au cas lorsque $i = j$, c'est-à-dire lorsque l'agent i a confiance en sa propre sincérité sur une proposition ϕ, i.e. $T^s_{i,i}\phi$. Par exemple, un agent i qui manipule un autre agent j sur une proposition ϕ sait qu'il ne peut pas avoir confiance en lui-même sur une proposition ϕ. Même si le cadre TB ne permet pas d'exprimer la manipulation, nous pouvons déduire en revanche qu'un agent i ne peut pas se faire confiance à lui-même s'il ne croit pas une proposition ϕ, i.e $\vdash \neg B_i\phi \Rightarrow \neg T^s_{i,i}\phi$. La proposition suivante considère la contraposée, i.e. $\vdash T^s_{i,i}\phi \Rightarrow B_i\phi$. Un agent i qui s'accorde sa confiance en sa sincérité croit nécessairement ce qu'il dit.

Proposition 4.10 : *Soit $i \in \mathcal{N}$ un agent :*

$$\vdash T^s_{i,i}\phi \Rightarrow B_i\phi$$

Démonstration.

1. $\vdash T^s_{i,i}\phi \Rightarrow B_iB_i\phi$
2. $\vdash B_iB_i\phi \Rightarrow B_i\phi$
3. $\vdash T^s_{i,i}\phi \Rightarrow B_iB_i\phi \Rightarrow B_i\phi$
4. $\vdash (T^s_{i,i}\phi \Rightarrow B_iB_i\phi \Rightarrow B_i\phi) \Rightarrow (T^s_{i,i}\phi \Rightarrow B_iB_i\phi) \Rightarrow T^s_{i,i}\phi \Rightarrow B_i\phi$
5. $\vdash T^s_{i,i}\phi \Rightarrow B_i\phi$

□

4.4.6 Confiance partagée

La confiance en la sincérité est définie de façon individuelle d'un agent envers un autre agent. Cependant, dans la littérature, il existe d'autres notions de confiance comme la réputation (cf. section 1.1.3) qui représente la confiance qu'un groupe d'agents accorde à un agent du système. Dans cette section, nous définissons une notion de confiance collective sur la sincérité, la *confiance partagée en la sincérité au sein d'un groupe.*

Confiance partagée en la sincérité

Pour définir la *confiance partagée*, nous reprenons la définition de (Smith et al., 2011). Un groupe d'agents fait confiance à un autre groupe d'agents si, et seulement si, tous les agents du premier groupe font confiance à tous les agents du second groupe :

$$\forall I, J \subseteq \mathcal{N} : Tc_{I,J}\phi \stackrel{\triangle}{=} \bigwedge_{(i,j)\in I\times J} T^s_{i,j}\phi$$

Cette définition de confiance collective représente un consensus dans le sens que tous les agents de I doivent faire confiance à tous les agents de J pour la même déclaration. De plus, nous considérons une notion duale à la confiance partagée, notée $Tc^*_{I,J}$:

$$\forall I, J \subseteq \mathcal{N} : Tc^*_{I,J}\phi \stackrel{\triangle}{=} \bigvee_{(i,j)\in I\times J} T^s_{i,j}\phi$$

Ce prédicat exprime qu'au moins un agent de I fait confiance à un agent de J. En effet, si aucun agent de I ne fait confiance à des agents de J pour une déclaration ϕ alors $\neg Tc^*_{I,J}\phi$ est vraie. Remarquons que la confiance partagée peut être définie différemment de celle considérée ici. Par exemple, (Herzig et al., 2010) considèrent que la *reputation* est un prédicat indiquant qu'au moins la *majorité* des agents de I ont une confiance dispositionnelle envers les agents de J. Par souci de simplicité, nous n'introduisons pas cette notion de majorité et, par conséquent, nous ne considérons pas cette notion de réputation.

La confiance partagée est un système KD

La confiance partagée possède toutes les propriétés d'un système KD.

Proposition 4.11 : *Pour tout groupe d'agents* $I, J, K \subseteq \mathcal{N}$ *:*

1. $\vdash Tc_{I,J}\phi \wedge Tc_{I,J}\psi \equiv Tc_{I,J}(\phi \wedge \psi)$
2. $\vdash (Tc_{I,J}\phi \vee Tc_{I,J}\psi) \Rightarrow Tc_{I,J}(\phi \vee \psi)$
3. $\vdash (Tc_{I,J}\phi \wedge Tc_{I,J}(\phi \Rightarrow \psi)) \Rightarrow Tc_{I,J}\psi$
4. $\vdash Tc_{I,J}\phi \Rightarrow \neg Tc^*_{I,J}\neg\phi$
5. $\vdash Tc_{I,J}\phi \Rightarrow \neg Tc_{I,J}\neg\phi$

Démonstration. (Schémas de preuve) Pour tout groupe d'agents $I, J, K \subseteq \mathcal{N}$,

(1) est obtenue par :

$$\vdash \bigwedge_{(i,j)\in I\times J} (T^s_{i,j}\phi \wedge T^s_{i,j}\psi) \equiv \bigwedge_{(i,j)\in I\times J} T^s_{i,j}(\phi \wedge \psi)$$

(2) est obtenue par :

$$\vdash \bigwedge_{(i,j)\in I\times J} (T^s_{i,j}\phi \vee T^s_{i,j}\psi) \Rightarrow \bigwedge_{(i,j)\in I\times J} T^s_{i,j}(\phi \vee \psi)$$

(3) est obtenue par :

1. $\{Tc_{I,J}\phi \wedge Tc_{I,J}(\phi \Rightarrow \psi)\} \models \bigwedge_{(i,j)\in I\times J} (T^s_{i,j}\phi \wedge (T^s_{i,j}(\phi \Rightarrow \psi)))$
2. $\{Tc_{I,J}\phi \wedge Tc_{I,J}(\phi \Rightarrow \psi)\} \models \bigwedge_{(i,j)\in I\times J} T^s_{i,j}\psi$
3. $\vdash (Tc_{I,J}\phi \wedge Tc_{I,J}(\phi \Rightarrow \psi)) \Rightarrow Tc_{I,J}\psi$

(4) est obtenue par :

1. $\{Tc_{I,J}\phi\} \models \bigwedge_{(i,j)\in I\times J} (T^s_{i,j}\phi \wedge (T^s_{i,j}\phi \Rightarrow \neg T^s_{i,j}\neg\phi))$
2. $\{Tc_{I,J}\phi\} \models \bigwedge_{(i,j)\in I\times J} \neg T^s_{i,j}\neg\phi$
3. $\{Tc_{I,J}\phi\} \models \neg \bigvee_{(i,j)\in I\times J} T^s_{i,j}\neg\phi$
4. $\vdash Tc_{I,J}\phi \Rightarrow \neg Tc^*_{I,J}\neg\phi$

(5) est obtenue par :

1. $\{Tc_{I,J}\phi\} \models \bigwedge_{(i,j)\in I\times J} (T^s_{i,j}\phi \wedge (T^s_{i,j}\phi \Rightarrow \neg T^s_{i,j}\neg\phi))$
2. $\{Tc_{I,J}\phi\} \models \bigwedge_{(i,j)\in I\times J} \neg T^s_{i,j}\neg\phi$
3. $\{Tc_{I,J}\phi\} \models \bigwedge_{(i,j)\in I\times J} \neg T^s_{i,j}\neg\phi \Rightarrow \bigvee_{(i,j)\in I\times J} \neg T^s_{i,j}\neg\phi$
4. $\{Tc_{I,J}\phi\} \models \bigvee_{(i,j)\in I\times J} \neg T^s_{i,j}\neg\phi \Rightarrow \neg \bigwedge_{(i,j)\in I\times J} T^s_{i,j}\neg\phi$
5. $\{Tc_{I,J}\phi\} \models \neg \bigwedge_{(i,j)\in I\times J} T^s_{i,j}\neg\phi$
6. $\vdash Tc_{I,J}\phi \Rightarrow \neg Tc_{I,J}\neg\phi$

□

Ainsi, la confiance partagée se comporte comme un système KD : les axiomes vérifiés au niveau de la confiance en la sincérité sont aussi vérifiés au niveau de la confiance partagée au sein d'un groupe d'agents. Cependant, nous remarquons que les axiomes d'introspection ne peuvent pas être vérifiés au niveau de la confiance partagée. En effet, les agents du groupe n'ont pas nécessairement conscience qu'ils partagent la même confiance en la sincérité pour un autre groupe d'agents. Pour avoir l'introspection sur la confiance collective, nous avons besoin de considérer une autre notion de confiance partagée que nous appelons *confiance commune partagée*. Cette notion peut être définie comme :

$$\forall I, J \subseteq \mathcal{N} : Tcp_{I,J}\phi \stackrel{\triangle}{=} \bigwedge_{i\in I} B_i Tc_{I,J}\phi \wedge \ldots \wedge \bigwedge_{i\in I} B_i \ldots \bigwedge_{j\in I} B_j Tc_{I,J}\phi$$

Ce prédicat exprime le fait que tous les agents du groupe I croient qu'ils ont une confiance partagée pour un groupe J et que tous les agents du groupe I croient que tous les agents du groupe I ont une confiance partagée pour le groupe J, etc.

La confiance partagée implique les croyances communes

Cependant, des agents qui ont une confiance partagée pour un groupe J, partagent aussi les mêmes croyances sur les croyances des agents du groupe J. Ainsi, l'axiome de sincérité est vérifié au niveau de la confiance partagée ainsi que la propriété de pseudo-transitivité sur les croyances.

Proposition 4.12 : *Pour tout groupe d'agents* $I, J, K \subseteq \mathcal{N}$,

1. $\vdash Tc_{I,J}\phi \Rightarrow \bigwedge_{(i,j)\in I\times J} B_iB_j\phi$
2. $\vdash Tc_{I,J}Tc_{J,K}\phi \Rightarrow \bigwedge_{(i,j,k)\in I\times J\times K} B_iB_jB_k\phi$

Démonstration. (Schémas de preuve) Pour tout $I, J, K \subseteq \mathcal{N}$:

(1) est obtenue par :

1. $\{Tc_{I,J}\phi\} \models \bigwedge\limits_{(i,j)\in I\times J} (T^s_{i,j}\phi \wedge (T^s_{i,j}\phi \Rightarrow B_iB_j\phi))$
2. $\{Tc_{I,J}\phi\} \models \bigwedge\limits_{(i,j)\in I\times J} B_iB_j\phi$
3. $\vdash Tc_{I,J}\phi \Rightarrow \bigwedge_{(i,j)\in I\times J} B_iB_j\phi$

(2) est obtenue par :

1. $\{Tc_{I,J}Tc_{J,K}\phi\} \models \bigwedge\limits_{(i,j)\in I\times J} (T^s_{i,j} \bigwedge\limits_{k\in K} T^s_{j,k}\phi \wedge (T^s_{i,j}T^s_{j,k}\phi \Rightarrow B_iB_jB_k\phi))$
2. $\{Tc_{I,J}Tc_{J,K}\phi\} \models \bigwedge\limits_{(i,j,k)\in I\times J\times K} B_iB_jB_k\phi$
3. $\vdash Tc_{I,J}Tc_{J,K}\phi \Rightarrow \bigwedge_{(i,j,k)\in I\times J\times K} B_iB_jB_k\phi$

□

Remarquons que ces preuves reposent sur les théorèmes $(\wedge_T)$ et $\forall k \in \mathcal{N}, \vdash B_k(p \wedge q) \equiv B_kp \wedge B_kq$. Par ces propriétés, nous prouvons alors que si deux groupes d'agents ont confiance en la sincérité de l'autre, cela implique que chaque agent du groupe I croient que chaque agent de J croient la déclaration ϕ.

4.5 Application de la confiance en la sincérité

L'exemple ci-dessous a pour objectif de présenter une instanciation de la logique TB.

Considérons une situation où un programme doit raisonner sur une situation dans laquelle un client et un vendeur sont en train de marchander. Le vendeur affirme qu'il est très intéressant d'investir dans ce produit. Nous pouvons imaginer qu'il est possible que ce vendeur soit payé

par la compagnie et donc ne soit pas sincère dans son affirmation. Contrairement à l'exemple 3.2 nous cherchons à représenter ici tous les mondes possibles. Ainsi, nous notons c l'agent client, et s l'agent vendeur, et $\mathcal{P} = \{p, q\}$ un ensemble de variables propositionnelles pour :

— $p :=$ « il est intéressant d'investir dans tel produit » ;

— $q :=$ « s est payé par la compagnie du produit pour le promouvoir ».

Nous faisons l'hypothèse que si l'agent s est payé par la compagnie, alors l'agent s croit que c'est le cas. De la même manière, si l'agent s n'est pas payé par la compagnie, alors l'agent s croit que ce n'est pas le cas. Cependant, nous pouvons remarquer qu'il est possible que l'agent s croit qu'il est intéressant d'investir dans le produit (i.e. $B_s p$) mais aussi il est possible que l'agent s croit le contraire (i.e $B_s \neg p$).

Du côté client, il est possible que ce dernier croit qu'il est intéressant d'investir dans le produit (i.e $B_c p$) mais aussi que le client croit qu'il n'est pas intéressant d'investir dedans (i.e. $B_c \neg p$). De la même manière, il est possible que ce client croit que le vendeur est payé par la compagnie, ou croit le contraire (i.e. $B_c q$ ou $B_c \neg q$). Par ailleurs, il est aussi possible que le client ne sache rien sur la situation $\neg B_c \neg p \wedge \neg B_c p$ et $\neg B_c \neg q \wedge \neg B_c q$.

Par souci de clarté, nous supposons tout d'abord que l'agent c informe le programme qu'il ne croit pas qu'il soit intéressant d'investir dans le produit, ni du contraire (i.e. $\neg B_c p \wedge \neg B_c \neg p$). Ensuite, nous supposons que l'agent c informe le programme qu'il croit que le vendeur s travaille pour la compagnie mais n'en est pas certain (i.e. $B_c q$).

Nous supposons alors deux configurations possibles d'états mentaux pour le vendeur, soit le vendeur sait s'il est intéressant d'investir dans le produit (i.e. $(B_s p \Rightarrow p) \wedge (B_s \neg p \Rightarrow \neg p)$), soit il ne sait pas s'il est intéressant d'investir dans le produit (i.e. $\neg B_s p \wedge \neg B_s \neg p$). Nous ne considérons pas les autres cas comme le fait que le vendeur croit qu'il est intéressant d'investir dans le produit sans pour autant être réellement le cas.

Supposons que le premier cas soit vérifié, c'est-à-dire si le vendeur sait s'il est intéressant ou non d'investir dans le produit. La structure de Kripke représentant les états mentaux des agents pour ce premier cas est donnée par la figure 4.2. Au sein de chaque noeud du graphe de Kripke de la figure 4.2, nous décrivons les variables qui sont vraies dans le monde correspondant au noeud. Ainsi, si $\mathcal{M}$ est le modèle de cette première situation, nous déduisons par exemple que l'agent c fait confiance en la sincérité de l'agent s s'il travaille pour la compagnie, i.e. $\mathcal{M} \models T^s_{c,s} q$. Or, par l'axiome de sincérité, nous avons $\mathcal{M} \models T^s_{c,s} q \Rightarrow B_c B_s q$, et donc $\mathcal{M} \models B_c B_s q$, i.e. l'agent c croit que l'agent s croit qu'il travaille pour la compagnie.

Le second cas est représenté par la structure de Kripke donnée par la figure 4.3. Le vendeur ne croit pas que son produit est bon, ni ne croit qu'il est mauvais. Nous déduisons alors $\mathcal{M} \models \neg B_s p \wedge \neg B_s \neg p$ et $\mathcal{M} \models B_c \neg B_s p$, c'est-à-dire le vendeur s ne croit pas qu'il est intéressant d'investir dans ce produit, ni ne croit du contraire et que le client croit qu'il n'est pas le cas que le vendeur croit que son produit est de qualité. Or, par l'axiome (D_{B_c}), nous avons $\mathcal{M} \models B_c \neg B_s p \Rightarrow \neg B_c B_s p$ et par la contraposition sur l'axiome de sincérité $\mathcal{M} \models \neg B_c B_s p \Rightarrow \neg T^s_{c,s} p$. Nous déduisons par

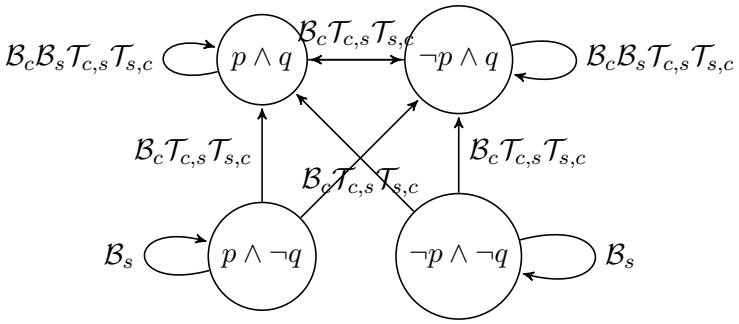

FIGURE 4.2 – Cas 1 : le vendeur sait s'il est intéressant d'investir dans le produit.

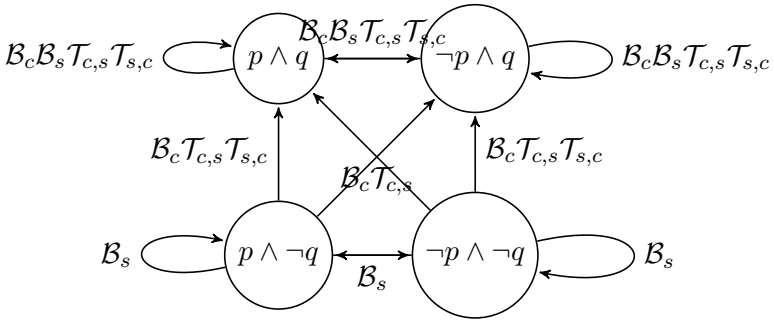

FIGURE 4.3 – Cas 2 : le vendeur ne sait pas s'il est intéressant d'investir dans le produit.

modus ponens que dans une telle situation, $\mathcal{M} \models \neg T^s_{c,s}p$, c'est-à-dire qu'il n'est pas le cas que le client fasse confiance en la sincérité du vendeur lorsqu'il dit que son produit est de qualité.

En conclusion, ce chapitre nous a permis de définir un nouveau cadre logique, nommé TB, permettant d'exprimer la notion de confiance en la sincérité. Nous introduisons alors dans ce système une modalité $T^s_{i,j}$ signifiant qu'un agent i a confiance à la sincérité de j lorsqu'il émet une proposition. Lorsqu'un agent i accorde sa confiance en la sincérité d'un agent j, cet agent i va nécessairement croire que l'agent j croit ce qu'il dit. Cette relation logique de la confiance en la sincérité est appelée *axiome de sincérité*. Cependant, un agent a toujours le choix de ne pas accorder sa confiance, comme par exemple, s'il croit que l'autre agent est en train de le manipuler. La réciproque à cet axiome de sincérité ne peut être considérée. Dans la suite de ce chapitre, nous avons démontré que ce système était correct, complet et nous avons étudié certaines propriétés logiques de la confiance en la sincérité comme, par exemple, qu'elle n'est pas transitive. Nous avons étendu cette confiance individuelle à une notion de confiance collective nommée *confiance partagée*. Par ailleurs, nous montrons que cette dernière notion de confiance partagée se comportait comme un système KD.

Si nous avons défini des systèmes logiques permettant de raisonner sur la manipulation ou sur la confiance en la sincérité d'un agent, nous n'avons aucune méthode algorithmique permettant de raisonner automatiquement dans ces systèmes logiques. Dans la partie suivante, nous pré-

sentons un ensemble de méthodes pour raisonner dans les logiques modales et appliquons l'une d'entre elles, appelée la *méthode des tableaux*. Cette méthode nous permet alors de construire un algorithme pour vérifier la satisfiabilité d'une formule du cadre logique TB.

Chapitre 5

Une méthode des tableaux dédiée au cadre TB

Nous avons vu aux chapitres 3 et 4 deux systèmes logiques, KBE et TB. Le système KBE est un système non-normal et multi-modal permettant de raisonner sur la notion de manipulation. Le système TB est un système normal et multi-modal permettant de raisonner sur la notion de confiance en la sincérité. Toutefois, nous n'avons pas présenté de méthode pour construire des algorithmes permettant de raisonner automatiquement à partir de ces systèmes. Ce chapitre a pour objectif de présenter des méthodes usuelles de résolution du problème de satisfiabilité dans les logiques modales, puis de proposer une méthode de résolution adaptée à TB qui exploite les propriétés de ce cadre pour faciliter la résolution. Dans un premier temps, en section 5.1, nous définissons les problèmes de satisfiabilité, de validité et de vérification de modèles dans les logiques modales. Nous présentons ensuite des méthodes usuelles pour résoudre la validité d'une formule dans les systèmes K, KD ou KD45. Dans un second temps, en section 5.2, nous présentons une nouvelle méthode des tableaux dédiée à la résolution du problème SAT dans TB et prouvons la correction et la complétude de la méthode. Enfin, en section 5.3, nous présentons un algorithme pour résoudre ce problème et son implémentation.

5.1 Méthodes des tableaux et arbres labellisés

Généralement, trois problèmes sont considérés pour étudier la complexité des systèmes logiques : le problème de satisfiabilité qui consiste à vérifier l'existence d'un modèle satisfaisant une formule du langage, le problème de validité qui consiste à vérifier si une formule est valide dans le système logique, et le problème de la vérification formelle de modèles[1] qui consiste à vérifier si une formule est satisfiable dans un modèle donné.

1. *Model-checking*

5.1.1 Des problèmes dans les logiques modales

Pour un cadre de logique modale quelconque M, les trois problèmes mentionnés précédemment se définissent de la façon suivante :

Définition 5.1 - *M-satisfiabilité* : *Le problème de* M-satisfiabilité *(M-SAT) est le problème de décision qui consiste, pour une formule ϕ de M en entrée, à vérifier s'il existe un modèle $\mathcal{M}$ du cadre M et un monde w tel que $\mathcal{M}, w \models \phi$.*

Définition 5.2 - *M-validité* : *Le problème de* M-validité *(M-VAL) est le problème de décision qui consiste pour une formule ϕ de M en entrée, à vérifier si pour tout modèle $\mathcal{M}$ du cadre M et pour tout monde w, nous avons $\mathcal{M}, w \models \phi$.*

Définition 5.3 - *M-vérification* : *Le problème de la* vérification de modèle *(M-MC) est le problème de décision qui consiste pour une formule ϕ de M, un modèle $\mathcal{M}$ du cadre M, et $I \subseteq \mathcal{W}$ un sous-ensemble de mondes, à vérifier si pour tout monde $w \in I$, nous avons $\mathcal{M}, w \models \phi$.*

Lorsque M est le cadre TB, nous notons TB-SAT, TB-VAL et TB-MC pour désigner respectivement le problème de satisfiabilité, le problème de validité et le problème de la vérification dans le cadre TB. Comme le problème de satisfiabilité est le dual du problème de validité, nous nous intéressons à l'implémentation d'une méthode des tableaux permettant de résoudre le problème de satisfiabilité. Le problème de la vérification de modèle est un problème plus simple à décider, nous ne l'aborderons pas dans ce chapitre. Il existe deux méthodes distinctes pour résoudre le problème de satisfiabilité ainsi que le problème de validité :

1. la méthode des tableaux, destinée davantage à la résolution du problème SAT ;
2. la méthode des arbres, plutôt destinée à la résolution du problème de la validité.

5.1.2 La méthode des tableaux

Dans cette sous-section, nous nous plaçons dans un cadre général à une modalité et présentons tout d'abord la méthode des tableaux destinée à la construction d'un modèle permettant de prouver la satisfiabilité d'une formule dans le système K.

Méthodes des tableaux pour les logiques propositionnelles

La méthode des tableaux (Fitting, 1983; Goré, 1999) permet de résoudre le problème de satisfiabilité par construction d'ensembles de formules fermés sur les sous-formules, appelés tableaux. L'exemple 5.1 illustre une méthode reposant sur les tableaux pour vérifier la satisfiabilité ou la validité d'une formule.

Exemple 5.1 : *Lorsque nous voulons vérifier la satisfiabilité d'une formule propositionnelle $\phi = (p \Rightarrow (q \Rightarrow r))$, nous pouvons naïvement énumérer l'ensemble des interprétations possibles pour les variables p, q et r comme illustré sur le tableau de la figure 5.1, puis vérifier si pour*

une interprétation possible, la formule est vraie. Ainsi, pour la formule ϕ, l'interprétation $I = \{(p,0);(q,0);(r,0)\}$ satisfait la formule ϕ. Cependant, pour vérifier la validité, il est nécessaire de parcourir l'ensemble des interprétations possibles et qu'aucune interprétation possible ne rende fausse la formule. Par exemple, remarquons que l'interprétation $I = \{(p,1);(q,1);(r,0)\}$ invalide la formule ϕ. En revanche, la formule $\psi = (p \Rightarrow (q \Rightarrow r)) \Rightarrow (p \Rightarrow q) \Rightarrow (q \Rightarrow r)$ est valide car il n'existe aucune interprétation dans laquelle ψ est fausse.

p	q	r	$q \Rightarrow r$	$p \Rightarrow q$	$p \Rightarrow q$	$p \Rightarrow (q \Rightarrow r)$	$(p \Rightarrow q) \Rightarrow (q \Rightarrow r)$	$(p \Rightarrow (q \Rightarrow r)) \Rightarrow (p \Rightarrow q) \Rightarrow (q \Rightarrow r)$
0	0	0	1	1	1	1	1	1
0	0	1	1	1	1	1	1	1
0	1	0	0	1	1	1	1	1
0	1	1	1	1	1	1	1	1
1	0	0	1	0	0	1	1	1
1	0	1	1	1	0	1	1	1
1	1	0	0	0	1	0	0	1
1	1	1	1	1	1	1	1	1

Tableau 5.1 – Exemple de tableau pour décider de la satisfiabilité d'une formule.

Cependant, cette façon de représenter les formules de la logique propositionnelle sous la forme d'un tableau où chaque ligne est associée à une fonction d'interprétation n'est pas la plus efficace. D'autres méthodes de représentation des formules existent comme les diagramme de décisions binaires et ordonnés (Huth and Ryan, 2004). Dans ce chapitre, une autre méthode de représentation des formules va nous intéresser car elle permet aussi de représenter les formules de logique modale : *les ensembles de Hintikka* (Blackburn et al., 2002).

Méthodes des tableaux pour les logiques modales

La méthode de l'exemple 5.1 ne suffit plus lorsque nous voulons interpréter une formule $\Box\theta$ ou $\Diamond\phi$ de logique modale. Pour pouvoir interpréter ces formules dans un tableau, il est alors nécessaire de construire un autre tableau vérifiant la formule ϕ et toutes les formules θ. Une technique consiste à construire un ensemble de tableaux que nous appelons *ensemble témoin.* Cet ensemble témoin est alors construit de telle sorte que pour chaque tableau I de cet ensemble et à chaque formule $\Diamond\phi \in I$ de ce tableau, il existe un autre tableau J vérifiant la formule ϕ ainsi que toutes les formules θ où $\Box\theta \in I$. Pour représenter aisément un tel tableau, nous avons recourt à la notion d'ensemble de Hintikka qui contient toutes les formules vérifiées pour un tableau donné mais aussi toutes ses sous-formules. L'ensemble $\{\psi\} \cup \{\theta : \Box\theta \in H\}$ de formule inclus dans J est appelé la *demande suscitée par $\Diamond\psi$ dans H* et noté $Dem(H, \Diamond\psi)$. Cette notion de demande représente intuitivement une contrainte imposée sur un modèle lorsqu'une relation d'accessibilité existe.

Exemple 5.2 : *Par exemple, considérons l'ensemble de formules $H = \{p, \Diamond q, \Box p\}$ et vérifions si cet ensemble est satisfiable dans le système K. Construisons un ensemble témoin, il*

existe une $\Diamond p$ dans H. Ainsi, pour vérifier si $\Diamond p$ est satisfiable, il est nécessaire de définir un nouvel ensemble de Hintikka I contenant la formule p. Cet ensemble représente l'instanciation d'un arc dans le modèle. Mais puisqu'il existe aussi une formule $\Box p \in H$, cette formule p doit aussi être vérifiée dans l'ensemble I. Ainsi, l'ensemble $I = \{p, q\}$ et un ensemble témoin, constitué des deux ensembles est $\{H, I\}$. Cet ensemble témoin décrit indirectement le modèle $\mathcal{M} = (\{w_0, w_1\}, \{(w_0, w_1)\}, \{(w_0, \{p, q\}), (w_1, \{p, q\}\})$. Nous avons bien $\mathcal{M}, w_0 \models H$, donc l'ensemble H est satisfiable.

Par conséquent, l'idée de cette méthode est que pour vérifier qu'une formule ϕ est satisfiable, il suffit alors de construire un ensemble de Hintikka H contenant ϕ et montrer qu'à partir de cet ensemble, nous pouvons construire un ensemble témoin.

Dans la suite, pour une formule ϕ, nous notons :

$$\sim \phi = \begin{cases} \psi & \text{si } \phi = \neg\psi, \\ \neg\phi & \text{sinon.} \end{cases}$$

Nous considérons comme équivalentes les notations suivantes :

1. $\Diamond \equiv \sim \Box \sim$
2. $\Box \equiv \sim \Diamond \sim$
3. $\sim \Box \equiv \Diamond \sim$
4. $\sim \Diamond \equiv \Box \sim$

Nous pouvons définir un ensemble de Hintikka.

Définition 5.4 - *Ensembles de Hintikka* : *Soit Σ un ensemble fermé de formules*[2]*. H est un* ensemble de Hintikka sur Σ *si, et seulement si, H est le plus grand sous-ensemble de Σ tel que :*

1. $\bot \notin H$
2. *si $\neg\phi \in \Sigma$, alors $\neg\phi \in H$ ssi $\phi \notin H$*
3. *si $\phi \wedge \psi \in \Sigma$, alors $\phi \wedge \psi \in \Sigma$ ssi $\phi \in \Sigma$ et $\psi \in \Sigma$*

Nous appelons un ensemble de Hintikka H sur Σ un atome *si, et seulement si, H est un ensemble de Hintikka sur Σ satisfiable.*

Exemple 5.3 : *Considérons par exemple, un ensemble de formules $\Gamma = \{p \wedge \Diamond\Diamond q \vee \Box p, \Diamond p\}$ et calculons un ensemble de Hintikka contenant l'ensemble des formules Γ. Notons $\Sigma = Cl(\Gamma)$ la fermeture de cet ensemble Γ. Nous avons $\Sigma = \{p \wedge \Diamond\Diamond q \vee \Box p, \Diamond p, p, \Diamond\Diamond q, \Box p, \neg\Box p, \neg(p \wedge \Diamond\Diamond q \vee \Box p), \neg p, \neg\Diamond\Diamond q, \neg\Diamond q, \Diamond q, q, \neg q, \neg\Diamond p\}$. Ainsi, un ensemble de Hintikka H sur Σ et contenant Γ est :*

$$H = \{p \wedge \Diamond\Diamond q \vee \Box p, \Diamond p, p, \Diamond\Diamond q, \Diamond q, q, \neg\Box p\}$$

2. La définition d'ensembles fermés est donnée en section 2.1.1.

Remarquons que pour un ensemble de Hintikka H sur Σ, où Σ est un ensemble fermé de formules, nous avons toujours $Cl(H) = \Sigma$. Ainsi, par abus de langage, nous pouvons parler d'ensemble de Hintikka H sans avoir à préciser par rapport à quel ensemble de formules fermé celui-ci est défini car il peut être obtenu en calculant $Cl(H)$. Il convient toutefois de souligner qu'il s'agit bien d'un abus de langage car un ensemble de Hintikka n'a de sens que par rapport à son ensemble fermé de formules Σ. De la même manière, il est aussi important de remarquer que tous les ensembles de Hintikka ne sont pas nécessairement satisfiables comme le montre l'exemple 5.4.

Pour vérifier qu'un ensemble de Hintikka est satisfiable, une méthode consiste à prouver l'existence d'un ensemble témoin généré par cet ensemble. Un ensemble témoin est une structure permettant de s'assurer qu'un ensemble de Hintikka est satisfiable. Nous avons parlé précédemment de la notion de demande dans les ensembles témoins. Nous donnons tout d'abord la définition de la demande, pour ensuite définir la notion d'ensemble témoins dans le système K.

Définition 5.5 - *Demande suscitée* : *Soit J un ensemble de Hintikka sur Σ, où Σ est un ensemble de formules fermé. Nous appelons, la* demande suscitée par $\Diamond\phi$ dans J, *l'ensemble $Dem(J, \Diamond\phi)$ tel que :*

$$Dem(J, \Diamond\phi) = \{\phi\} \cup \{\sim \theta | \neg\Diamond\theta \in J\} \text{ avec } \Diamond\phi \in J$$

Les ensembles de Hintikka contenant la demande suscitée par un $\Diamond\phi$ dans J sont représentés par l'ensemble $J_{\Diamond\phi}$ tel que :

$$J_{\Diamond\phi} := \{I \subseteq Cl(Dem(J, \Diamond\phi)) | I \text{ est un ensemble de Hintikka tel que } Dem(J, \Diamond\phi) \subseteq I\}$$

Définition 5.6 - *Ensemble témoin généré par un ensemble de Hintikka* : *Soient Σ un ensemble de formules fermé et H un ensemble de Hintikka sur Σ. Nous appelons l'ensemble $\mathcal{H}_H \subseteq 2^\Sigma$,* un ensemble témoin généré par un ensemble de Hintikka H *sur Σ si, et seulement si :*

1. $H \in \mathcal{H}_H$
2. *si $I \in \mathcal{H}_H$, alors pour tout $\Diamond\psi \in I$, il existe $J \in I_{\Diamond\psi}$ tel que $J \in \mathcal{H}_H$*
3. *si $J \in \mathcal{H}_H$ et $J \neq H$, alors il existe $I^0, \ldots, I^n \in \mathcal{H}_H$ avec $n \in \mathbb{N}^*$ tel que $H = I^0$, $J = I^n$ et pour tout $0 \leq i < n$, il existe $\Diamond\psi \in I^i$ tel que $I^{i+1} \in I^i_{\Diamond\psi}$.*

La relation entre un ensemble de Hintikka et les ensembles témoins générés est donnée par le théorème 5.1.

Théorème 5.1 - : *Un ensemble de Hintikka H sur Σ, où Σ est un ensemble de formules fermé, est satisfiable si, et seulement si, il existe un ensemble témoin $\mathcal{H}_H$ généré par H.*

Démonstration. La preuve est donnée dans (Blackburn et al., 2002). □

L'exemple 5.4 montre deux situations d'ensembles de formules $\Gamma_1 = \{\Box(p \wedge q), \Diamond\neg q\}$ et $\Gamma_2 = \{\Box\Box p, \Diamond\neg p, \Diamond(p \wedge q)\}$ avec lesquels nous construisons deux ensembles de Hintikka H_1 et H_2 contenant respectivement les ensembles Γ_1 et Γ_2. Même si nous pouvons construire un ensemble de Hintikka contenant Γ_2, celui-ci est non satisfiable dans K et nous ne pouvons donc construire un ensemble témoin. Le second ensemble de formules Γ_2 est satisfiable dans K et nous construisons alors l'ensemble témoin généré par l'ensemble de Hintikka H_2 contenant les formules de Γ_2.

Chaque élément d'un ensemble témoin représente la valuation associée à un type de monde possible. La relation du modèle satisfaisant un ensemble de Hintikka est construite en même temps que l'application du calcul de la demande. Ainsi, la construction d'un ensemble témoin nous donne un modèle et un monde dans lequel l'ensemble de Hintikka de départ H est satisfait.

Exemple 5.4 : *Considérons un premier exemple d'ensemble de formules* $\Gamma_1 = \{\Box(p \wedge q), \Diamond\neg q\}$. *La fermeture de l'ensemble* Γ_1 *est donnée par l'ensemble* $Cl(\Gamma_1) = \{\Box(p \wedge q), \neg\Box(p \wedge q), \Diamond\neg q, \neg\Diamond\neg q, p \wedge q, \neg(p \wedge q), p, q, \neg p, \neg q\}$. *Nous pouvons construire un ensemble de Hintikka* $H_1 = \{\Box(p \wedge q), \Diamond\neg q, p \wedge q, p, q\}$. *Cependant, nous remarquons que* H_1 *n'est pas satisfaisable dans K. En effet, puisque* $Dem(H_1, \Diamond\neg q) = \{\neg q\} \cup \{p \wedge q\}$ *et* $Cl(Dem(H_1, \Diamond\neg q)) = \{p \wedge q, p, q\}$, *il n'existe pas d'ensemble de Hintikka* $I \in (H_1)_{\Diamond\neg q}$ *contenant* $Dem(H_1, \Diamond\neg q)$ *puisqu'un ensemble de Hintikka ne peut contenir à la fois* q *et* $\neg q$. *Par conséquent, pour cet ensemble* Γ_1, *il est impossible de construire un ensemble témoin généré par* H_1 *sur la fermeture de* $Cl(\Gamma_1)$ *et contenant* Γ_1.

Considérons un second exemple d'ensemble de formules $\Gamma_2 = \{\Box\Box p, \Diamond\neg p, \Diamond(p \wedge q)\}$. *Nous avons* $Cl(\Gamma_2) = \{\Box\Box p, \neg\Box\Box p, \Box p, \neg\Box p, \Diamond\neg p, \neg\Diamond\neg p, \neg p, p, \Diamond(p\wedge q), \neg\Diamond(p\wedge q), p\wedge q, \neg(p\wedge q), q, \neg q\}$. *L'ensemble* $H_2 = \{\Box\Box p, \Diamond\neg p, \Diamond(p\wedge q), \neg\Box\neg p, p\wedge q, p, q\}$ *est un ensemble de Hintikka sur* $Cl(\Gamma_2)$ *contenant* Γ_2. H_2 *est satisfiable dans K. En effet, nous pouvons construire un ensemble témoin* $\mathcal{H}_{H_2} = \{H_2, I, K, J\}$ *généré par l'ensemble de Hintikka* H_2 *sur* $Cl(\Gamma_2)$ *et contenant* Γ_2. *Nous avons* $I = \{\neg p, \Box p\}$ *l'ensemble de Hintikka dans* $I \in (H_2)_{\Diamond\neg p}$ *sur* $Cl(Dem(H_2, \Diamond\neg p))$ *et contenant la demande* $Dem(H_2, \Diamond\neg p) = \{\neg p\} \cup \{\Box p\}$. *De plus, l'ensemble* $J = \{\Box p, p \wedge q, p, q\}$ *est un ensemble de Hintikka* $J \in (H_2)_{\Diamond(p\wedge q)}$ *sur* $Cl(Dem(H_2, \Diamond(p\wedge q)))$ *contenant* $Dem(H_2, \Diamond(p\wedge q)) = \{p \wedge q\} \cup \{\Box p\}$. *Enfin, puisque* $\sim\Box\sim p \equiv \Diamond p$, *l'ensemble* $K = \{\Box p, p\}$ *est un ensemble de Hintikka* $K \in (H_2)_{\Diamond p}$ *sur* $Cl(Dem(H_2, \Diamond p))$ *contenant* $Dem(H_2, \Diamond p) = \{p\} \cup \{\Box p\}$. *La figure 5.1 représente le modèle satisfaisant* Γ_2 *dans le système K. Cependant, si nous avions considéré un autre système comme* $KD45$, Γ_2 *ne pourrait être satisfiable en raison de* $\Box\Box p$ *et* $\Diamond\neg p$. *L'inconsistance provient alors du théorème* $\vdash_{KD45} \Box\Box p \Rightarrow \Box p$. *Nous aurions dans ce cas à satisfaire* $\neg p$ *et* p *à la fois, ce qui est impossible.*

Si cette méthode de construction fonctionne dans le système K, ce n'est pas nécessairement le cas lorsque nous considérons d'autres systèmes comme KD ou S5. Pour considérer les contraintes liées à ces cadres logiques, il est alors nécessaire d'ajouter des contraintes sur ces ensembles témoins reflétant chacune des contraintes du cadre. Par la suite, en section 5.2, nous faisons cela

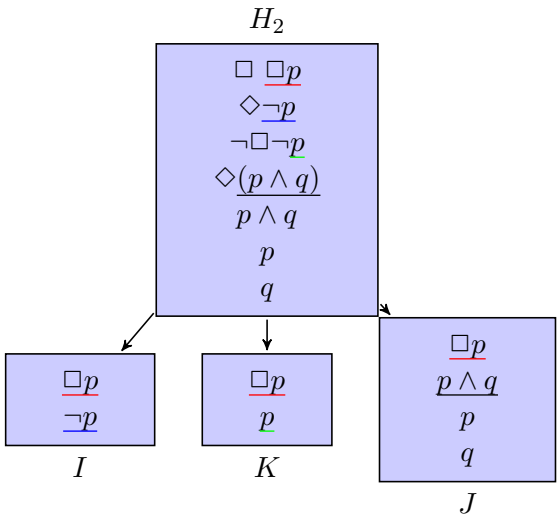

FIGURE 5.1 – Application de la méthode des tableaux sur H_2

en construisant une nouvelle définition d'ensemble témoins dédiée au cadre TB.

5.1.3 Les arbres labellisés

Nous présentons maintenant une seconde méthode, *la méthode des arbres labellisés*, permettant de décider de la validité d'une formule dans un système modal. Avec cette méthode, nous donnons un exemple de construction d'un contre-modèle lorsqu'une formule n'est pas valide. Pour plus de détails sur la méthode des arbres labellisés, le lecteur intéressé peut se référer à (Fitting, 1983) et (Girle, 2014).

Méthode des arbres pour la logique propositionnelle

La méthode des arbres labellisés représente les formules de logique propositionnelle sous la forme d'un arbre. Dans un tel arbre, une formule n'est pas valide s'il existe une branche de l'arbre contenant une lettre propositionnelle et son contraire. Nous appelons une telle branche, *une branche fermée*. Ainsi, pour tester si une formule ϕ est valide, il suffit alors de vérifier que toutes les branches de l'arbre sont fermées pour l'arbre généré par sa négation $\neg\phi$. Pour ce faire, cette méthode considère alors deux types de règles syntaxiques (voir tableau 5.2 et 5.3) pour construire un arbre : les règles α et les règles β. Les règles de construction $\alpha(\phi, \psi)$ sont attachées au connecteur logique $\wedge$ et consistent à ajouter sur une même branche les formules qui constituent une conjonction.

Les règles $\beta(\phi, \psi)$ sont quant à elles associées aux opérations sur la disjonction. Elles servent à construire dans l'arbre deux branchements qui représentent chaque partie de la disjonction.

L'exemple 5.5 illustre une application de ces règles pour construire un arbre afin de mettre en évidence qu'une formule n'est pas valide. Remarquons que chaque chemin qui n'est pas une

	$\alpha(\phi, \psi)$	Description
1	$\frac{\phi \wedge \psi}{\begin{array}{c}\phi \\ \psi\end{array}}$	Ajouter sur la branche de $\phi \wedge \psi$ les formules ϕ puis ψ.
2	$\frac{\neg(\phi \vee \psi)}{\begin{array}{c}\neg\phi \\ \neg\psi\end{array}}$	Ajouter sur la branche de $\neg(\phi \vee \psi)$ les formules $\neg\phi$ puis $\neg\psi$.
3	$\frac{\neg(\phi \Rightarrow \psi)}{\begin{array}{c}\phi \\ \neg\psi\end{array}}$	Ajouter sur la branche de $\neg(\phi \Rightarrow \psi)$ les formules ϕ puis $\neg\psi$.
4	$\frac{\neg\neg\phi}{\phi}$	Ajouter sur la branche de $\neg\neg\phi$ la formule ϕ.

Tableau 5.2 – Tableau des règles $\alpha(\phi, \psi)$

	$\beta(\phi, \psi)$	Description
1	$\frac{\phi \vee \psi}{\begin{array}{cc}\phi & \psi\end{array}}$	Construire deux nouvelles branches partant de $\phi \vee \psi$: l'une avec ϕ, l'autre avec ψ
2	$\frac{\phi \Rightarrow \psi}{\begin{array}{cc}\neg\phi & \psi\end{array}}$	Construire deux nouvelles branches partant de $\phi \Rightarrow \psi$: l'une avec $\neg\phi$, l'autre avec ψ.
3	$\frac{\neg(\phi \wedge \psi)}{\begin{array}{cc}\neg\phi & \neg\psi\end{array}}$	Construire deux nouvelles branches partant de $\neg(\phi \wedge \psi)$: l'une avec $\neg\phi$, l'autre avec $\neg\psi$.

Tableau 5.3 – Tableau des règles $\beta(\phi, \psi)$

branche fermée dans un tel arbre nous donne un ensemble de Hintikka (cf. définition 5.4).

Exemple 5.5 :

Considérons la formule $(a \vee \neg b) \wedge b$ avec a, b deux variables propositionnelles. Par applications des règles présentées à la section 5.1.3, nous obtenons l'arbre de la figure 5.2. Nous remarquons que dans cet exemple, il existe une branche fermée dans laquelle, nous retrouvons au moins une variable propositionnelle et son contraire (la variable b). Par conséquent, la formule $(a \vee \neg b) \wedge b$ n'est pas valide. En effet, pour tout modèle tel que $V(a) = \bot$, $\mathcal{M} \not\models \phi$. La branche $\{(a \vee \neg b) \wedge b, (a \vee \neg b), b, a\}$ donnée est bien un ensemble de Hintikka sur la fermeture $\Sigma = \{(a \vee \neg b) \wedge b, \neg((a \vee \neg b) \wedge b), (a \vee \neg b), \neg(a \vee \neg b), b, \neg b, a, \neg a\}$

Cette méthode de construction d'arbre est un système de preuve syntaxique, au même titre que les systèmes de preuve à la Hilbert comme nous avons pu les présenter en section 2.1.1. Ce système de preuve syntaxique est correct et complet par rapport à la sémantique associée à la logique propositionnelle.

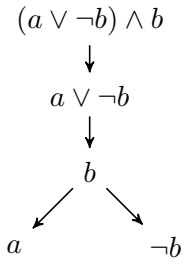

FIGURE 5.2 – Exemple d'arbre montrant que la formule $(a \vee \neg b) \wedge b$ n'est pas valide.

Méthodes des arbres pour les logiques modales

La méthode des arbres doit être complétée pour pouvoir vérifier la validité d'une formule en logique modale. En effet, une formule modale est vraie par rapport au monde dans lequel elle est située, mais aussi par rapport aux relations d'accessibilité. Ainsi, à chaque nœud de l'arbre, nous devons préciser par rapport à quel monde nous nous situons et comment accéder à ce monde. Pour ce faire, considérons un cadre de Kripke $\mathcal{C} = (\mathcal{W}, \mathcal{R}_1, \ldots, \mathcal{R}_n)$ avec $n \in \mathbb{N}^*$ et notons $\mathcal{Z} = \{1, 2, 3, \ldots\}$ un ensemble de symboles que nous appelons *ensemble de labels* et $\mathcal{X} = \{:_{\mathcal{R}_i}\}_{i \in [|1,n|]}$ un ensemble de symboles que nous appelons *ensemble de symboles relationnels*. À chaque relation d'accessibilité de $\mathcal{C}$ est associée un symbole de $\mathcal{X}$. Pour $n = 1$, nous notons simplement $\mathcal{X} = \{:\}$. Ces deux ensembles $\mathcal{Z}$ et $\mathcal{X}$ combinés, nous permettent de définir un ensemble de mondes possibles et comment accéder à ces mondes possibles comme présenté dans l'exemple 5.6.

Exemple 5.6 : *Par exemple, prenons un cadre de Kripke* $\mathcal{C} = (\{w_0, w_1, w_2, w_3, w_4\}, \mathcal{R}_1, \mathcal{R}_2)$ *avec* $w_0 = 1$, $w_1 = (w_0 :_{\mathcal{R}_1} 1)$, $w_2 = (w_0 :_{\mathcal{R}_1} 2)$, $w_3 = (w_0 :_{\mathcal{R}_2} 1)$, $w_4 = (w_1 :_{\mathcal{R}_2} 1)$, *deux relations d'accessibilité* $\mathcal{R}_1$ *et* $\mathcal{R}_2$. *Les relations d'accessibilité* $\mathcal{R}_1$ *et* $\mathcal{R}_2$ *sont alors telles que* $\mathcal{R}_1 = \{(w_0, w_1), (w_0, w_2)\}$ *et* $\mathcal{R}_2 = \{(w_0, w_3), (w_1, w_4)\}$. La méthode des arbres labellisés utilise alors ces ensembles $\mathcal{Z}$ et $\mathcal{X}$ pour construire un arbre de formules qui à chaque nœud de l'arbre associe sa formule, le monde dans lequel cette formule est vraie, et comment accéder à ce monde possible.

Dans la suite de cette section, nous considérons un cadre de Kripke $\mathcal{C} = (\mathcal{W}, \mathcal{R})$ et $w \in \mathcal{W}$ un monde possible. Pour le système K, il est nécessaire de considérer deux nouvelles règles : les règles de type ν (voir tableau 5.5) et les règles de type π (voir tableau 5.4), associées respectivement aux opérateurs $\Box$ et $\Diamond$. Les règles $\pi(\phi)$ expriment les opérations à appliquer sur l'arbre lorsqu'un opérateur $\Diamond$ apparaît. Tandis que les règles $\nu(\phi)$ expriment les opérations à effectuer sur l'arbre lorsque l'opérateur $\Box$ apparaît.

Cette méthode des arbres permet donc, non seulement de décider de la validité d'une formule ϕ en vérifiant que tous les chemins parcourus de l'arbre généré par la négation de ϕ sont fermés, mais aussi, en cas d'invalidité, de construire un contre-modèle satisfaisant la négation de ϕ. L'exemple 5.7 présente une application de cette méthode pour construire un contre-modèle dans

	$\pi(\phi)$		Description
1	w	$\Diamond\phi$	Ajouter sur la branche la formule ϕ, associée au monde $(w:x)$
	$w:x$	ϕ	avec $x \in \mathcal{Z}$ un label tel que $w:x$ n'est pas déjà utilisé sur la branche.
2	w	$\neg\Box\phi$	Ajouter sur la branche la formule $\neg\phi$, associée au monde $(w:x)$
	$w:x$	$\neg\phi$	avec $x \in \mathcal{Z}$ un label tel que $w:x$ n'est pas déjà utilisé sur la branche.

Tableau 5.4 – Tableau des règles $\pi(\phi)$

	$\nu(\phi)$		Description
1	w	$\Box\phi$	Pour tout label $x \in \mathcal{Z}$ tel que $(w:x)$ apparaît sur la branche,
	$w:x$	ϕ	ajouter la formule ϕ sur la branche, associée au monde $(w:x)$.
2	w	$\neg\Diamond\phi$	Pour tout label $x \in \mathcal{Z}$ tel que $(w:x)$ apparaît sur la branche,
	$w:x$	$\neg\phi$	ajouter la formule $\neg\phi$ sur la branche, associée au monde $(w:x)$.

Tableau 5.5 – Tableau des règles $\nu(\phi)$

le système K.

Exemple 5.7 : *La figure 5.3 représente un arbre construit à partir de la négation de la formule* $\phi = \Box(p \vee \Diamond q) \Rightarrow (\Box p \vee \Diamond q)$ *dans le système K. Nous pouvons remarquer que cette formule n'est pas valide puisque toutes les branches de l'arbre ne sont pas fermées. Ainsi, nous pouvons construire un modèle* $\mathcal{M}$ *satisfaisant la formule* $\neg\phi$ *en considérant une branche non fermée de cet arbre. Considérons alors le modèle* $\mathcal{M} = (\mathcal{W}, \mathcal{R}, V)$ *tel que :*

— $\mathcal{W} = \{1, 1:1, 1:1:1\}$;

— $\mathcal{R}\{(1, 1:1); (1:1, 1:1:1)\}$;

— $V(p) = \emptyset$ *et* $V(q) = \{1:1:1\}$.

Nous avons $\mathcal{M}, 1 \models \neg\phi$. *Ce modèle constitue donc un contre-exemple à la validité de la formule* ϕ *dans le système K.*

Concernant les autres systèmes modaux usuels comme KT, KD, KB ou S5, il est possible d'ajouter de nouvelles contraintes sur les arbres pour vérifier la validité des formules. La figure 5.6 présente l'ensemble des règles à ajouter aux règles précédentes pour vérifier la validité des formules dans ces différents systèmes. Par exemple, pour construire un arbre permettant de vérifier la validité d'une formule dans le système KD, il est nécessaire d'ajouter, en plus des règles du système K, les règles D1 et D2. Le système S4 qui est composé des axiomes K, T et 4 doit considérer les règles du système K, mais aussi celles du système KT, qui sont les règles supplémentaires T1 et T2. Enfin, pour le système S5, les règles à considérer sont celles de K, mais aussi les règles T1, T2, 41, 42, 51 et 52. Il est à noter que les règles sont appliquées dans l'ordre suivant :

1. les règles sur les opérateurs binaires α et β ;
2. les règles sur les opérateurs $\Diamond$ et $\neg\Box$, i.e. les règles π ;

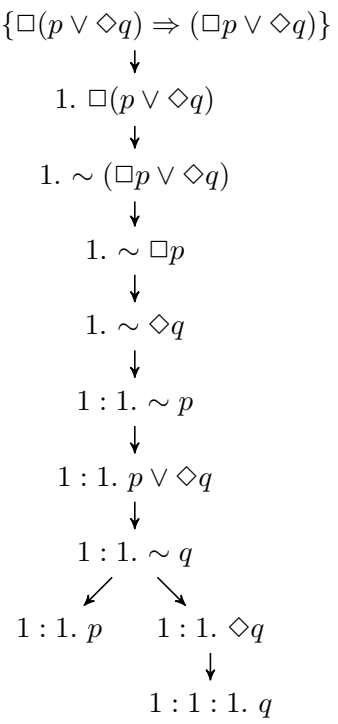

FIGURE 5.3 – Exemple d'arbre pour la logique modale K (Schmitz, 2019)

3. les règles sur les opérateurs □ et $\neg\Diamond$, i.e. ν, D1, D2, T1, T2, 41, 42, B1, B2, 51 et 52.

Nous avons donc présenté des méthodes de calcul permettant de résoudre le problème de validité dans les systèmes usuels unimodaux. Cette méthode des arbres labellisés peut s'étendre aux logiques multi-modales et normales, en combinant et adaptant ces différentes contraintes au système considéré. Par exemple, (Baldoni, 2000) présente un ensemble de règles à considérer pour les systèmes multi-modaux avec des axiomes d'interaction incestueux (cf. section 2.1.2).

5.2 Résoudre TB-SAT avec les ensembles de Hintikka

Dans cette section, nous présentons une méthode des tableaux fondée sur la notion d'ensembles témoins telle que nous l'avons présenté en section 5.1. Cependant, si la méthode des ensembles témoins fonctionne pour le système K, il est nécessaire de l'adapter aux autres systèmes. Ainsi, nous présentons une méthode permettant de construire un ensemble témoin pour vérifier la satisfiabilité d'une formule dans TB.

5.2.1 Ensembles témoins pour le cadre TB

Dans le cadre TB, puisque toutes les relations d'accessibilité sont sérielles, nous présentons tout d'abord une notion d'ensembles témoins adaptée pour ces cadres en considérant aucune dépendance logique entre modalités, puis nous définissons la notion d'ensembles témoins pour le

Système	Nom	Règles	Description
KT (K+T)	T1	$\frac{w \quad \Box\phi}{w \quad \phi}$	Sur la branche de $\Box\phi$ associée au monde w, ajouter la formule ϕ associée au même monde w.
	T2	$\frac{w \quad \neg\Diamond\phi}{w \quad \neg\phi}$	Sur la branche de $\neg\Diamond\phi$ associée au monde w, ajouter la formule $\neg\phi$ associée au même monde w.
KD (K+D)	D1	$\frac{w \quad \Box\phi}{w \quad \Diamond\phi}$	Sur la branche de $\Box\phi$ associée au monde w, ajouter la formule $\Diamond\phi$ associée au même monde w.
	D2	$\frac{w \quad \neg\Diamond\phi}{w \quad \neg\Box\phi}$	Sur la branche de $\neg\Diamond\phi$ associée au monde w, ajouter la formule $\neg\Box\phi$ associée au même monde w.
S4 (K+T+4)	41	$\frac{w \quad \Box\phi}{w:x \quad \Box\phi}$	Pour tout label $x \in \mathcal{Z}$ tel que $(w:x)$ apparaît sur la branche, ajouter la formule $\Box\phi$ associée au monde $(w:x)$.
	42	$\frac{w \quad \neg\Diamond\phi}{w:x \quad \neg\Diamond\phi}$	Pour tout label $x \in \mathcal{Z}$ tel que $(w:x)$ apparaît sur la branche, ajouter la formule $\neg\Diamond\phi$ associée au monde $(w:x)$.
KB (K+T+B)	B1	$\frac{w:x \quad \Box\phi}{w \quad \phi}$	Sur la branche de $\Box\phi$ associée au monde$(w:x)$, ajouter la formule ϕ associée au monde w.
	B2	$\frac{w:x \quad \neg\Diamond\phi}{w \quad \neg\phi}$	Sur la branche de $\neg\Diamond\phi$ associée au monde $(w:x)$, on ajoute la formule $\neg\phi$ associée au monde w.
S5 (K+T+4+5)	51	$\frac{w:x \quad \Box\phi}{w \quad \Box\phi}$	Sur la branche de $\Box\phi$ associée au monde $(w:x)$, ajouter la formule $\Box\phi$ associée au monde w.
	52	$\frac{w:x \quad \neg\Diamond\phi}{w \quad \neg\Diamond\phi}$	Sur la branche de $\neg\Diamond\phi$ associée au monde $(w:x)$, ajouter la formule $\neg\Diamond\phi$ associée au monde w.

Tableau 5.6 – Tableau des règles pour les systèmes KT, KD, S4, KB et S5

cadre TB. De la même façon que pour les ensembles témoins dans le système K, nous donnons une définition de la demande suscitée dans les cadres sériels, puis nous donnons la définition des ensembles témoins dans ces cadres.

Définition 5.7 - *Demande suscitée dans les cadres sériels* : *Soit J un ensemble de Hintikka sur Σ. La notion de demande suscitée par $\Diamond\phi$ ou $\Box\phi$ dans J est telle que :*

$$Dem(J, \Diamond\phi) = \{\phi\} \cup \{\sim\theta | \neg\Diamond\theta \in J\} \text{ avec } \Diamond\phi \in J$$

$$Dem(J, \Box\phi) = \{\phi\} \cup \{\sim\theta | \neg\Diamond\theta \in J\} \text{ avec } \Box\phi \in J$$

Ainsi, les ensembles de Hintikka contenant la demande suscitée par un $\Diamond\phi$ et par un $\Box\phi$ dans J sont représentés par les ensembles $J_{\Diamond\phi}$ et $J_{\Box\phi}$ tels que :

$$J_{\Diamond\phi} := \{I \subseteq Cl(Dem(J, \Diamond\phi)) | I \text{ est un ensemble de Hintikka tel que } Dem(J, \Diamond\phi) \subseteq I\}$$

$$J_{\Box\phi} := \{I \subseteq Cl(Dem(J, \Box\phi)) | I \text{ est un ensemble de Hintikka tel que } Dem(J, \Box\phi) \subseteq I\}$$

Définition 5.8 - *Ensemble témoin pour un cadre sériel généré par un ensemble de Hintikka* : *Soient Σ un ensemble fermé de formules et H un ensemble de Hintikka sur Σ où Σ est un ensemble de formules fermé.Nous appelons $\mathcal{H}_H \subseteq 2^\Sigma$* un ensemble témoin pour un cadre sériel généré par un ensemble de Hintikka *H :*

1. *$H \in \mathcal{H}_H$;*
2. *si $I \in \mathcal{H}_H$, alors pour tout $\Diamond\psi \in I$, il existe $J \in I_{\Diamond\psi}$ tel que $J \in \mathcal{H}_H$;*
3. *si $I \in \mathcal{H}_H$, alors pour tout $\Box\psi \in I$, il existe $J \in I_{\Box\psi}$ tel que $J \in \mathcal{H}_H$;*
4. *si $J \in \mathcal{H}_H$ et $J \neq H$, alors il existe $I^0,\ldots, I^n \in \mathcal{H}_H$ avec $n \in \mathbb{N}^*$ tel que $H = I^0$, $J = I^n$ et pour tout $0 \leq i < n$, il existe $\Diamond \in \tau$ et $\Diamond\psi \in I^i$ tel que $I^{i+1} \in I^i_{\Diamond\psi}$.*

Pour préciser le caractère sériel d'une relation d'accessibilité dans les ensembles témoins, la définition 5.8 incorpore une nouvelle contrainte (3) dans les ensembles témoins pour considérer le cas où, dans un ensemble de Hintikka $I \in \mathcal{H}_H$, il n'existe pas $\Diamond\phi \in I$ mais il existe au moins une formule $\Box\phi \in I$. Cette nouvelle règle est une conséquence directe de l'axiome D ($\models \Box\phi \Rightarrow \Diamond\phi$), i.e. si $\Box\phi \in I$ alors, puisque $\Diamond\phi$ doit être vérifié, il doit nécessairement exister un ensemble de Hintikka dans l'ensemble témoin contenant la demande suscitée par $\Box\phi$ sur I. Ainsi, dans un cadre sériel, les $\Box/\neg\Diamond$ suscitent eux-aussi une demande sur les ensembles de Hintikka I tels qu'il n'existe pas de $\Diamond\phi \in I$ mais uniquement des formules $\Box\phi \in I$.

Dans la suite de cette section, nous revenons au cadre TB et nous considérons l'ensemble des modalités de type $\Diamond$ du langage $\mathcal{L}_{TB}$ comme l'ensemble $\tau = \{\langle B_i\rangle : i \in \mathcal{N}\} \cup \{\langle T^s_{i,j}\rangle : i, j \in \mathcal{N}\}$. Afin d'étendre la notion d'ensembles témoins aux cadres logiques mélangeant plusieurs modalités avec des dépendances logiques entre ces modalités, nous introduisons dans la définition 5.9, une nouvelle notion que nous nommons *extension de Hintikka*. Cette notion caractérise une notation pratique nous permettant d'ajouter à un ensemble de Hintikka de nouvelles formules provenant des contraintes sémantiques associées à notre cadre logique.

Définition 5.9 - *Extensions de Hintikka* : *Soient J un ensemble de formules, $\Diamond \in \tau$ une modalité, I un ensemble de Hintikka et Γ un ensemble de formules.*

— *Si $\Diamond$ n'est pas une modalité sérielle alors nous appelons l'ensemble J une* extension simple de Hintikka *générée par I sur $\Diamond$ si, et seulement si, J est un ensemble de Hintikka sur $Cl(J)$ tel que :*

$$\exists\Diamond\phi \in I \wedge \exists J' \subseteq J : J' \in I_{\Diamond\phi}$$

— *Si $\Diamond$ est une modalité sérielle, alors nous appelons l'ensemble J une* extension sérielle de Hintikka *générée par I sur $\Diamond$ si, et seulement si, J est un ensemble de Hintikka sur $Cl(J)$ tel que $\exists\Diamond\phi \in I \wedge \exists J' \subseteq J : J' \in I_{\Diamond\phi}$ ou $\exists\Box\phi \in I \wedge \exists J' \subseteq J : J' \in I_{\Box\phi}$.*

— *J est une* Γ-extension de Hintikka *générée par I, notée $I\langle\Gamma\rangle J$ si, et seulement si, J est un ensemble de Hintikka sur $Cl(J)$ tel que $\Gamma \subseteq J$.*

Dans les deux premiers cas, nous notons $I\Diamond J$ ou $I\Diamond_\psi J$, lorsque la formule ψ est fixée.

La notion de demande reste inchangée dans le cadre TB par rapport à celle des cadres sériels puisque toutes les modalités de TB sont sérielles. Pour la notion d'ensemble témoin, nous avons

besoin tout d'abord de caractériser les ensembles témoins pour chaque sous-système, i.e. les systèmes KD45 associés aux modalités doxastiques B_i et les systèmes KD associés aux modalités $T^s_{i,j}$. Dans un second temps, puisque TB est la combinaison de ces deux sous-systèmes, nous avons besoin d'introduire une notion de générateurs d'ensembles témoins. Cette notion de générateur caractérise les interactions entre ces sous-systèmes et nous permet de construire un ensemble témoin pour ce cadre si l'ensemble de Hintikka initial est satisfiable.

La définition 5.10, présentée plus bas, caractérise les ensembles témoins associés au cadre KD45, nommé *ensemble $\langle B_i\rangle$-témoin $\mathcal{H}^{B_i}_H$ généré par H* avec H un ensemble de Hintikka. Cette notion d'ensembles témoins considère la nature des propriétés sur les relations $\mathcal{B}_i$. Nous rappelons que ces relations sont transitives, sérielles et euclidiennes, ce qui signifie que la géométrie formée par ces structures peut être décomposée en deux parties : un nœud racine qui atteint tous les autres nœuds de la structure et une clique. Ainsi, si dans un ensemble $\langle B_i\rangle$-témoin $\mathcal{H}^{B_i}_H$ il existe un ensemble $I \in \mathcal{H}^{B_i}_H$ et une formule $B_i\psi \in I$, alors nécessairement ψ doit être dans tous les ensembles $J \in \mathcal{H}^{B_i}_H$, $J \neq H$, i.e. tous les nœuds qui sont dans la clique. Ainsi, la définition 5.10 reprend celle des ensembles témoins pour les cadres sériels et ajoute les propriétés des relations d'accessibilité $\mathcal{B}_i$. De plus, puisque ces ensembles témoins sont des sous-systèmes du cadre TB, nous ajoutons à la définition un ensemble de formules, noté Γ, qui représente des contraintes extérieures à ce sous-système et provenant des interactions avec les autres systèmes du cadre.

Exemple 5.8 : *Par exemple, s'il existe un ensemble $I \in \mathcal{H}^{B_i}_H : I \neq H$ tel que dans cet ensemble I, nous avons une formule $B_j\theta \in I$, alors il est nécessaire de construire un autre ensemble $\langle B_j\rangle$-témoin $\mathcal{H}^{B_j}_I$ à partir de l'ensemble de Hintikka I. Puisque $\mathcal{B}_i \circ \mathcal{B}_j \subseteq \mathcal{T}_{i,j}$, tous les $J \in \mathcal{H}^{B_j}_I : J \neq I$ doivent contenir les θ provenant des $T^s_{i,j}\theta$ des ensembles de $\mathcal{H}^{B_i}_H$. Ainsi, cet ensemble $\Gamma = \{\theta : T^s_{i,j}\theta \in K, K \in \mathcal{H}^{B_i}_H\}$. De plus, puisque les relations $\mathcal{B}_j$ sont sérielles, il doit exister dans $\mathcal{H}^{B_j}_I$ au moins un ensemble de $I \in \mathcal{H}^{B_j}_I$ qui contienne Γ et pour tous les $J \in \mathcal{H}^{B_j}_I : J \neq I$, $\Gamma \subseteq J$.*

Définition 5.10 - *Ensemble $\langle B_i\rangle$-témoin* : *Soient Σ un ensemble fermé de formules, H un ensemble de Hintikka sur Σ et Γ un ensemble de formules. Nous appelons $\mathcal{H}^{B_i}_{H,\Gamma} \subseteq 2^\Sigma$* un ensemble $\langle B_i\rangle$-témoin *contraint par Γ et généré par un ensemble de Hintikka H si, et seulement si, les propriétés suivantes sont vérifiées :*

1. $H \in \mathcal{H}_{H,\Gamma}$
2. *si $I \in \mathcal{H}^{B_i}_{H,\Gamma}$, alors pour tout $\langle B_i\rangle\psi \in I$, il existe $J \in \mathcal{H}^{B_i}_{H,\Gamma}$ tel que $I\langle B_i\rangle_\psi J$*
3. *si $I \in \mathcal{H}^{B_i}_{H,\Gamma}$, alors pour tout $B_i\psi \in I$, il existe $J \in \mathcal{H}^{B_i}_{H,\Gamma}$ tel que $I\langle B_i\rangle_\psi J$*
4. *si $I \in \mathcal{H}^{B_i}_{H,\Gamma}$, alors pour tout $B_i\theta \in I$, pour tout $J \in \mathcal{H}^{B_i}_{H,\Gamma} \setminus \{H\}$, $\theta \in J$*
5. *pour tout $I \in \mathcal{H}^{B_i}_{H,\Gamma} \setminus \{H\}$, $\Gamma \subseteq I$*
6. *il existe $I \in \mathcal{H}^{B_i}_{H,\Gamma}$ tel que $\Gamma \subseteq I$*
7. *si $J \in \mathcal{H}^{B_i}_{H,\Gamma} \setminus \{H\}$, alors il existe $I^0, \ldots, I^n \in \mathcal{H}^{B_i}_{H,\Gamma}$ avec $n \in \mathbb{N}^*$ tel que $H = I^0$, $J = I^n$, $I^0\langle\Gamma\rangle I^1$ ou $I^0\langle B_i\rangle I^1$, et pour tout $1 \leq i < n$, il existe $\langle B_i\rangle\psi \in I^i$ tel que $I^i\langle B_i\rangle_\psi I^{i+1}$*

La définition 5.11 définit les ensembles $\langle T^s_{i,j}\rangle$-témoins pour décrire les sous-systèmes KD, associés aux modalités $T^s_{i,j}$. Nous notons $\mathcal{H}^{T^s_{i,j}}_{H,\Gamma}$ un ensemble $\langle T^s_{i,j}\rangle$-témoin généré par un ensemble de Hintikka H et contraint par un Γ. Cet ensemble Γ permet de capturer les contraintes de transitivité et de caractère euclidien des $\mathcal{B}_i$ par rapport aux $\mathcal{T}_{i,j}$, et donc de considérer les formules θ des $T^s_{i,j}\theta$ appartenant aux éléments d'un ensemble $\mathcal{H}^{B_i}_{H,\Gamma}$ ou $\mathcal{H}^{B_i}_{H',\Gamma}$ où $H \in \mathcal{H}^{B_i}_{H',\Gamma}$. Il est à noter que ce Γ ne s'applique que sur les ensembles I directement suscités à partir des demandes de la racine H.

Définition 5.11 - *Ensemble* $\langle T^s_{i,j}\rangle$*-témoin* : *Soient Σ un ensemble fermé de formules, H un ensemble de Hintikka sur Σ et Γ un ensemble de formules.Nous appelons $\mathcal{H}^{T^s_{i,j}}_{H,\Gamma} \subseteq 2^{\Sigma}$* un ensemble $\langle T^s_{i,j}\rangle$-témoin *contraint par Γ et généré par un ensemble de Hintikka H si, et seulement si, les propriétés suivantes sont vérifiées :*

1. $H \in \mathcal{H}^{T^s_{i,j}}_{H,\Gamma}$
2. *Si* $I \in \mathcal{H}^{T^s_{i,j}}_{H,\Gamma}$, *alors pour tout* $\langle T^s_{i,j}\rangle\psi \in I$, *il existe* $J \in \mathcal{H}^{T^s_{i,j}}_{H,\Gamma}$ *tel que* $I\langle T^s_{i,j}\rangle_\psi J$
3. *Si* $I \in \mathcal{H}^{T^s_{i,j}}_{H,\Gamma}$, *alors pour tout* $T^s_{i,j}\psi \in I$, *il existe* $J \in \mathcal{H}^{T^s_{i,j}}_{H,\Gamma}$ *tel que* $I\langle T^s_{i,j}\rangle_\psi J$
4. *Pour tout* $I \in \mathcal{H}^{T^s_{i,j}}_{H,\Gamma} \setminus \{H\}$ *tel que* $H\langle T^s_{i,j}\rangle I$, $\Gamma \subseteq I$;
5. *Il existe* $I \in \mathcal{H}^{T^s_{i,j}}_{H,\Gamma}$ *tel que* $\Gamma \subseteq I$
6. *Si* $J \in \mathcal{H}^{T^s_{i,j}}_{H,\Gamma} \setminus \{H\}$, *alors il existe* $I^0,\ldots, I^n \in \mathcal{H}^{T^s_{i,j}}_{H,\Gamma}$ *avec* $n \in \mathbb{N}^*$ *tel que* $H = I^0$, $J = I^n$, $I^0\langle\Gamma\rangle I^1$ *ou* $I^0\langle T^s_{i,j}\rangle I^1$, *et pour tout* $1 \leq i < n$, *il existe* $\langle T^s_{i,j}\rangle\psi \in I^i$ *tel que* $I^i\langle T^s_{i,j}\rangle_\psi I^{i+1}$

Maintenant que les ensembles témoins ont été définis pour les sous-systèmes KD45 pour les B_i et KD pour les $T^s_{i,j}$, nous pouvons désormais définir la structure d'ensembles témoins générés par un ensemble de Hintikka dans le cadre TB. Pour générer ces ensembles, dits *TB-témoins*, nous avons besoin de considérer un ensemble d'ensembles témoins, appelés *générateurs d'ensembles témoins générés par un ensemble de Hintikka H* et noté $\mathcal{G}^{TB}_H$. Cet ensemble $\mathcal{G}^{TB}_H$ permet de construire tous les ensembles témoins associés aux sous-systèmes B_i et $T^s_{i,j}$ en tenant compte des contraintes du cadre. L'ensemble TB-témoin est alors obtenu en faisant l'union des ensembles contenus dans tous les ensembles témoins contenus dans ce générateur.

La définition 5.12 définit ce qui est formellement un générateur d'ensembles TB-témoins. Elle est constituée d'un ensemble de départ $\{H\}$ appelé *structure initiale du générateur* et permettant d'initialiser la construction de l'ensemble témoin. Initialement, pour chaque modalité $\Diamond$ et formule $\Diamond\phi \in H$, nous construisons un ensemble $\mathcal{H}^{\Diamond}_{H,\Gamma'}$ $\Diamond$-témoin généré à partir de H. Pour les $\Diamond = B_i$, les $\Gamma' = \emptyset$ sont vides car aucun autre sous-système n'interfère avec les ensembles $\langle B_i\rangle$-témoins. Cependant, en raison du caractère euclidien des $\mathcal{B}_i$ par rapport aux $\mathcal{T}_{i,j}$, les ensembles $\mathcal{H}^{B_i}_{H,\Gamma'}$ ont une incidence sur les ensembles $\mathcal{H}^{T^s_{i,j}}_{H,\Gamma'}$ $\langle T^s_{i,j}\rangle$-témoins. Ainsi, le Γ' des $\langle T^s_{i,j}\rangle$-témoins vaut $\Gamma = \{\theta : T^s_{i,j}\theta \in I, I \in \mathcal{H}^{B_i}_{H,\Gamma'}\}$. Ensuite, par récursivité, nous construisons chaque ensemble témoin $\mathcal{H}^{\Diamond}_{H',\Gamma''}$, $\Diamond$-témoin en tenant compte des contraintes sémantiques imposées par le cadre TB.

Définition 5.12 - *Générateur d'ensembles TB-témoins générés par un ensemble de Hintikka* : *Soient H un ensemble de Hintikka sur Σ, où Σ est un ensemble de formules fermé. Nous appelons $\mathcal{G}_H^{TB} \subseteq 2^\Sigma$ un* générateur d'ensembles TB-témoins fondé sur H *si, et seulement si, toutes les propriétés suivantes sont vérifiées :*

1. $\mathcal{H}_H^\emptyset \in \mathcal{G}_H^{TB}$ *où* $\mathcal{H}_H^\emptyset = \{H\}$ *et est appelée* structure initiale du générateur $\mathcal{G}_H^{TB}$
2. *Si* $\Diamond \in \tau \cup \{\emptyset\}$, $\mathcal{H}_{H',\Gamma'}^\Diamond \in \mathcal{G}_H^{TB}$, *et* $I \in \mathcal{H}_{H',\Gamma'}^\Diamond$, *alors pour tout* $\Diamond'\psi \in I$ *avec* $\Diamond \neq \Diamond'$, *il existe* $\Gamma \subseteq \Sigma$ *et* $\mathcal{H}_{I,\Gamma}^{\Diamond'} \in \mathcal{G}_H^{TB}$ *tel que* $I\Diamond'_\psi J$ *et* $J \in \mathcal{H}_{I,\Gamma}^{\Diamond'}$ *où* $\mathcal{H}_{I,\Gamma}^{\Diamond'}$ *est un ensemble* $\Diamond'$*-témoin généré par* I *et contraint par* Γ. *Nous notons alors* $\mathcal{H}_{H',\Gamma'}^\Diamond \Diamond' \mathcal{H}_{I,\Gamma}^{\Diamond'}$
3. *Si* $\Diamond \in \tau \cup \{\emptyset\}$, $\mathcal{H}_{H',\Gamma'}^\Diamond \in \mathcal{G}_H^{TB}$, *et* $I \in \mathcal{H}_{H',\Gamma'}^\Diamond$, *alors pour tout* $\neg\Diamond'\psi \in I$ *avec* $\Diamond \neq \Diamond'$, *il existe* $\Gamma \subseteq \Sigma$ *et* $\mathcal{H}_{I,\Gamma}^{\Diamond'} \in \mathcal{G}_H^{TB}$ *tel que* $I\Diamond'_{\sim\psi} J$ *et* $J \in \mathcal{H}_{I,\Gamma}^{\Diamond'}$ *où* $\mathcal{H}_{I,\Gamma}^{\Diamond'}$ *est un ensemble* $\Diamond'$*-témoin généré par* I *et contraint par* Γ. *Nous notons alors* $\mathcal{H}_{H',\Gamma'}^\Diamond \Diamond' \mathcal{H}_{I,\Gamma}^{\Diamond'}$
4. *Si* $\Diamond \in \tau \cup \{\emptyset\}, \Diamond' \in \tau, \Diamond \neq \Diamond'$, $\mathcal{H}_{H',\Gamma'}^\Diamond \in \mathcal{G}_H^{TB}$, $I \in \mathcal{H}_{H',\Gamma'}^\Diamond$, $\nexists\Diamond'\psi \in I$ *et* $\nexists\neg\Diamond'\psi \in I$, *alors il existe* $\Gamma \subseteq \Sigma$ *et* $\mathcal{H}_{I,\Gamma}^{\Diamond'} \in \mathcal{G}_H^{TB}$ *tel que* $\mathcal{H}_{I,\Gamma}^{\Diamond'}$ *est un ensemble* $\Diamond'$*-témoin généré par* I *et contraint par* Γ. *Nous notons alors* $\mathcal{H}_{H',\Gamma'}^\Diamond \Diamond' \mathcal{H}_{I,\Gamma}^{\Diamond'}$
5. *Si* $\Diamond \in \tau$, $\mathcal{H}_{H',\Gamma'}^\Diamond \in \mathcal{G}_H^{TB}$ *et* $\mathcal{H}_{H'}^\Diamond \neq \mathcal{H}_H^\emptyset$, *alors il existe* $\mathcal{H}_{I^0,\Gamma^0}^{\Diamond_0}, \ldots, \mathcal{H}_{I^n,\Gamma^n}^{\Diamond_n} \in \mathcal{G}_H^{TB}$ *avec* $n \in \mathbb{N}^*$ *tel que* $\mathcal{H}_H^\emptyset = \mathcal{H}_{I^0,\emptyset}^{\Diamond_0}$, $\mathcal{H}_{H',\Gamma'}^\Diamond = \mathcal{H}_{I^n,\Gamma^n}^{\Diamond_n}$ *et pour tout* $0 \leq i < n$, *nous avons* $\mathcal{H}_{I^i,\Gamma^i}^{\Diamond_i} \Diamond_{i+1} \mathcal{H}_{I^{i+1},\Gamma^{i+1}}^{\Diamond_{i+1}}$
6. *Si* $\mathcal{H}_{H',\Gamma'}^{B_i}, \mathcal{H}_{H'',\Gamma''}^{B_j} \in \mathcal{G}_H^{TB}$ *tel que* $\mathcal{H}_{H',\Gamma'}^{B_i} \langle B_j \rangle \mathcal{H}_{H'',\Gamma''}^{B_j}$ *alors*

$$\Gamma'' = \bigcup_{I \in \mathcal{H}_{H',\Gamma'}^{B_i}} \{\theta : T_{i,j}^s \theta \in I\} \bigcup_{I \in \mathcal{H}_{H'',\Gamma''}^{B_j}} \{\theta : T_{j,j}^s \theta \in I\}$$

7. *Si* $\mathcal{H}_{H',\Gamma'}^{B_i}, \mathcal{H}_{H'',\Gamma''}^{T_{i,j}^s} \in \mathcal{G}_H^{TB}$ *tel que* $\mathcal{H}_{H',\Gamma'}^{B_i} \langle T_{i,j}^s \rangle \mathcal{H}_{H'',\Gamma''}^{T_{i,j}^s}$, *alors :*

$$\Gamma'' = \bigcup_{I \in \mathcal{H}_{H',\Gamma'}^{B_i}} \{\theta : T_{i,j}^s \theta \in I\}$$

8. *Si* $\mathcal{H}_{H',\Gamma'}^{B_i}, \mathcal{H}_{H'',\Gamma''}^{T_{i,j}^s} \in \mathcal{G}_H^{TB}$ *tel que* $\mathcal{H}_{H'',\Gamma''}^{T_{i,j}^s} \langle B_i \rangle \mathcal{H}_{H',\Gamma'}^{B_i}$, *alors pour tout* $J \in \mathcal{H}_{H'',\Gamma''}^{T_{i,j}^s}$, *tel que* $H' \langle T_{i,j}^s \rangle J$, *nous avons* $\bigcup_{I \in \mathcal{H}_{H',\Gamma'}^{B_i} \setminus \{H'\}} \{\theta : T_{i,j}^s \theta \in I\} \subseteq J$

Si H est un ensemble de Hintikka sur Σ et $\mathcal{G}_H^{TB}$ un générateur d'un ensemble TB-témoin fondé sur H, nous remarquons qu'un générateur d'ensemble TB-témoins $\mathcal{G}_H^{TB}$ forme un arbre de profondeur au plus $\deg(\Sigma)$. Les contraintes 6, 7 et 8 correspondent respectivement aux contraintes $\mathcal{B}_i \circ \mathcal{B}_j \subseteq \mathcal{T}_{i,j}$, $\mathcal{B}_i \circ \mathcal{T}_{i,j} \subseteq \mathcal{T}_{i,j}$ et $\mathcal{T}_{i,j} \subseteq \mathcal{B}_i \circ \mathcal{T}_{i,j}$. Enfin, pour définir un ensemble TB-témoin, il suffit alors de faire l'union de tous les ensembles contenus dans ces ensembles témoins.

Définition 5.13 - *Ensemble TB-témoin généré par un ensemble de Hintikka* : *Un ensemble $\mathcal{H}_H$ est un* ensemble TB-témoin généré par un ensemble de Hintikka H *sur* Σ, *un*

ensemble formules fermé si, et seulement si, il existe un générateur $\mathcal{G}_H^{TB}$ d'ensembles TB-témoins fondé sur H tel que :

$$\mathcal{H}_H = \bigcup_{\mathcal{H}^{\Diamond}_{H',\Gamma'} \in \mathcal{G}_H^{TB}} \{I : I \in \mathcal{H}^{\Diamond}_{H',\Gamma'}\}$$

Nous donnons maintenant le théorème de la méthode des tableaux dédiée pour TB ainsi que le schéma de la preuve. Ce schéma de preuve présente les différentes étapes permettant de démontrer le théorème. La suite de la section consiste à définir chacune des notions données dans ce schéma et la méthode complète pour démontrer ce théorème.

Théorème 5.2 - :
Soit H un ensemble de Hintikka sur Σ, un ensemble fini de formules fermé.

H est un TB-atome si, et seulement si, il existe un ensemble TB-témoin généré par H sur Σ.

Démonstration. (Schéma de la preuve) Soit H un ensemble de Hintikka sur Σ un ensemble fermé de formules.
($\Rightarrow$) (Hypothèse) Supposons qu'il existe un modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$ et un monde $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models H$.

1. Nous montrons que la restriction $\mathcal{M}_w$ de $\mathcal{M}$ aux mondes accessibles w est telle que $\mathcal{M}_w, w \models H$, ce modèle $\mathcal{M}_w$ est appelé *modèle TB-connexe depuis w* ;
2. Puis nous définissons une filtration, nommée *TB-filtration* $\mathcal{M}^f_{w,\Sigma}$ de $\mathcal{M}_w$ à travers Σ où Σ est un sous ensemble fermé de formules de H et nous montrons que celle-ci préserve bien toutes les propriétés sur les relations d'accessibilité ;
3. Enfin, nous montrons qu'à partir des classes d'équivalence de $\mathcal{W}^f_{w,\Sigma}$ nous pouvons construire un générateur d'ensembles TB-témoins à partir de l'ensemble H sur Σ et donc un ensemble TB-témoin.

($\Leftarrow$) Soit H un ensemble de Hintikka sur Σ et $\mathcal{G}_H^{TB}$ un générateur d'ensemble TB-témoin fondé sur H. La réciproque est plus simple à démontrer, il suffit de construire un cadre $(\mathcal{F}_n)_{n\in\mathbb{N}} = (\mathcal{W}_n, \{\mathcal{B}_i^n\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}^n\}_{i,j\in\mathcal{N}}, f_n)_{n\in\mathbb{N}}$ de manière récursive qui suit la construction de l'arbre formé par le générateur $\mathcal{G}_H^{TB}$ et d'appliquer les contraintes du cadre TB. Initialement, $\mathcal{W}_0 = \{w_0\}$, pour tout $i, j \in \mathcal{N}$, $\mathcal{B}_i^0 = \emptyset$, $\mathcal{T}_{i,j}^0 = \emptyset$ et $f_0(w_0) = H$. La définition complète d'un tel cadre est donné dans la preuve 5.2.3. Enfin, si $m \in \mathbb{N}$ représente l'étape à partir de laquelle $\mathcal{F}_m$ a terminé d'être construit, i.e. l'ensemble du générateur a été parcouru et toutes les contraintes du cadre TB ont été appliquées, il suffit alors de vérifier que le modèle $\mathcal{M} = (\mathcal{F}_m, V)$ tel que $w \in V(p)$ si, et seulement si, $p \in f_n(w)$ vérifie que $\mathcal{M}, w_0 \models H$. Ainsi, H est satisfiable.

□

Exemple 5.9 : *Considérons une formule $\phi = B_i\langle B_j\rangle q \wedge T^s_{i,j} q \wedge p$ et construisons avec la méthode un ensemble TB-témoin pour décider de la satisfiabilité de la formule ϕ.*

La première étape consiste à calculer un premier ensemble de Hintikka contenant ϕ. Calculons la fermeture de l'ensemble $\{\phi\}$, notée $\Sigma = \{B_i\langle B_j\rangle q \wedge T^s_{i,j}q \wedge p,\ \neg B_i\langle B_j\rangle q, B_i\langle B_j\rangle q,\ \neg T^s_{i,j}q, T^s_{i,j}q, \neg p,\ p,\ \langle B_j\rangle q, \neg\langle B_j\rangle q,\ q, \neg q,\ \neg(B_i\langle B_j\rangle q \wedge T^s_{i,j}q \wedge p)\}$ et considérons l'ensemble $H = \{B_i\langle B_j\rangle q \wedge T^s_{i,j}q \wedge p,\ B_i\langle B_j\rangle q, T^s_{i,j}q,\ p, \langle B_j\rangle q, q\}$.

À partir de cet ensemble H contenant ϕ, nous calculons un générateur $\mathcal{G}^{TB}_H$ fondé sur cet ensemble H. Nous avons tout d'abord une structure $\mathcal{H}^{B_i}_H$, $\mathcal{H}^{B_j}_H$ et une structure $\mathcal{H}^{T^s_{i,j}}_H$ à calculer. Nous avons $\mathcal{H}^{B_i}_H = \{H, I_0 = \{\langle B_j\rangle q, q\}\}$, $\mathcal{H}^{B_j}_H = \{H, J_0 = \{q\}\}$ et $\mathcal{H}^{T^s_{i,j}}_H = \{H, K_0 = \{q\}\}$. Dans la structure $\mathcal{H}^{B_i}_H$, il existe un $\langle B_j\rangle q \in I_0$. Ainsi, il est nécessaire de construire une nouvelle structure $\mathcal{H}^{B_j}_{I_0} = \{I_0, J_{01} = \{q\}\}$. Nous avons alors :

$$\mathcal{G}^{TB}_H = \{\{H\}, \{H, I_0\}, \{H, J_0\}, \{H, K_0\}, \{I_0, J_{01}\}\}$$

L'ensemble $\mathcal{G}^{TB}_H$ est donc un générateur d'ensemble TB-témoins fondé sur H. Par conséquent, $\mathcal{H}_H = \{H, I_0, J_0, K_0, J_{01}\}$ est un ensemble TB-témoin. Ainsi par le théorème 5.2, nous venons de prouver que la formule ϕ est satisfiable dans TB.

5.2.2 Sous modèles TB-connexes et TB-filtration

Dans cette section, nous présentons la notion de *TB-connexité* et de *TB-filtration.*

Étape 1 : extraction d'un modèle TB-connexe La première étape pour démontrer ce théorème consiste à extraire les mondes accessibles de $\mathcal{M}$ depuis w. Pour ce faire, nous définissons la notion de modèle connexe depuis un monde racine.

Définition 5.14 - *TB-chemin* : *Soient un modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$, deux mondes possibles $w, v \in \mathcal{W}$ et $n \in \mathbb{N}^*$. Un $(n+1)-uplet$, $(w, w_1, \ldots, w_{n-1}, v) \in \mathcal{W}^{n+1}$ est appelé un* TB-chemin de w à v dans $\mathcal{M}$ *si, et seulement si, $\forall k \in [|1, n|], \exists \mathcal{R}_k \in \{\mathcal{B}_i\}_{i\in\mathcal{N}} \cup \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}$: $w_{k-1}\mathcal{R}_k w_k$, où $w_0 = w$ et $w_n = v$.*

Définition 5.15 - *Modèle TB-connexe* : *Soient H un ensemble de Hintikka sur Σ, un ensemble fermé de formules, un modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$, et un monde possible $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models H$. Nous appelons* modèle TB-connexe de $\mathcal{M}$ depuis w, *un modèle $\mathcal{M}_w = (\mathcal{W}_w, \{\mathcal{B}^w_i\}_{i\in\mathcal{N}}, \{\mathcal{T}^w_{i,j}\}_{i,j\in\mathcal{N}}, V_w)$ tel que :*

- $\mathcal{W}_w = \{w\} \cup \{v \in \mathcal{W} :$ *il existe un TB-chemin de w à v dans $\mathcal{M}_w\}$*
- $\forall u, v \in \mathcal{W}_w, \forall \mathcal{R} \in \{\mathcal{B}_i\}_{i\in\mathcal{N}} \cup \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, u\mathcal{R}^w v$ *ssi* $u\mathcal{R}v$;
- $\forall p \in \mathcal{A}(\mathcal{L}), V_w(p) := \{v \in \mathcal{W}_w : v \in V(p)\}$.

Avec cette définition, nous obtenons les propositions suivantes :

Proposition 5.1 :

1. $\mathcal{M}_w, w \models H$;

2. $\forall v \in \mathcal{W}_w$, $\mathcal{M}_w, v \models \phi$ *si, et seulement si,* $\mathcal{M}, v \models \phi$;
3. *Toutes les propriétés du cadre TB sont vérifiées dans* $(\mathcal{W}_w, \{\mathcal{B}_i^w\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}^w\}_{i,j\in\mathcal{N}})$.

Démonstration. Les preuves sont triviales. □

Étape 2 : extraction d'une TB-filtration La seconde étape consiste à extraire une filtration qui préserve les propriétés sur les relations du cadre TB.

Définition 5.16 - : *Soient* Σ *un ensemble fermé de formules et* $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$ *un modèle.*

$$\forall w, v \in \mathcal{W} : w \leftrightsquigarrow_\Sigma v \text{ ssi } \forall \phi \in \Sigma : (\mathcal{M}, w \models \phi \text{ ssi } \mathcal{M}, v \models \phi)$$

Nous écrivons $|w|_\Sigma = \{v \in \mathcal{W} : w \leftrightsquigarrow_\Sigma v\}$ *et* $|w|$ *s'il n'y a aucune ambiguïté.*

Définition 5.17 - *Filtration* : *Nous appelons* $\mathcal{M}_\Sigma^f$ *une* filtration *de* $\mathcal{M}$ *à travers* Σ *si, et seulement si,* $\mathcal{M}_\Sigma^f = (\mathcal{W}_\Sigma, \{\mathcal{B}_i^f\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}^f\}_{i,j\in\mathcal{N}}, V^f)$ *tel que pour tout* $u, v \in \mathcal{W}$, *pour tout* $\mathcal{R}_\Diamond \in \{\mathcal{B}_i\}_{i\in\mathcal{N}} \cup \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}$, *toutes les propriétés suivantes sont vérifiées :*

1. *si* $u\mathcal{R}_\Diamond v$, *alors* $|u|_\Sigma \mathcal{R}_\Diamond^f |v|_\Sigma$;
2. *si* $|u|_\Sigma \mathcal{R}_\Diamond^f |v|_\Sigma$, *alors* $\forall \Diamond\phi \in \Sigma \setminus \{\emptyset\}, \mathcal{M}, v \models \phi \Longrightarrow \mathcal{M}, u \models \Diamond\phi$;
3. $\forall p \in \mathcal{A}(\mathcal{L}), V^f(p) := \{|v| : v \in \mathcal{W}_w \wedge \mathcal{M}, w \models p\}$.

Proposition 5.2 : *Soient* Σ *un ensemble fermé de formules,* $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$ *un modèle, et* $\mathcal{M}_\Sigma^f = (\mathcal{W}_\Sigma, \{\mathcal{B}_i^f\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}^f\}_{i,j\in\mathcal{N}}, V^f)$ *tel que :*

- $\forall u, v \in \mathcal{W} : |u|_\Sigma \mathcal{B}_i^f |v|_\Sigma$ *si, et seulement si,* $\forall \langle B_i\rangle\phi \in \Sigma, \mathcal{M}, v \models \phi \vee \langle B_i\rangle\phi \vee B_i\phi \Longrightarrow \mathcal{M}, u \models B_i\langle B_i\rangle\phi$;
- $\forall u, v \in \mathcal{W} : |u|_\Sigma \mathcal{T}_{i,j}^f |v|_\Sigma$ *si, et seulement si,* $\forall \langle T_{i,j}^s\rangle\phi \in \Sigma, \mathcal{M}, v \models \phi \Longrightarrow \mathcal{M}, u \models \langle T_{i,j}^s\rangle\phi$;
- $\forall p \in \mathcal{A}(\mathcal{L}), V^f(p) := \{|v| : v \in \mathcal{W}_w \wedge \mathcal{M}, w \models p\}$.

$\mathcal{M}_\Sigma^f$ *est une filtration de* $\mathcal{M}$ *à travers* Σ *qui préserve toutes les propriétés du cadre TB. Nous appelons* $\mathcal{M}_\Sigma^f$ *une* TB-filtration *de* $\mathcal{M}$ *à travers* Σ.

Démonstration. Soient Σ un ensemble fermé de formules, $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$ un modèle, et $\mathcal{M}_\Sigma^f = (\mathcal{W}_\Sigma, \{\mathcal{B}_i^f\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}^f\}_{i,j\in\mathcal{N}}, V^f)$ le modèle donné dans la proposition 5.2.
(1) Prouvons que pour tout $u, v \in \mathcal{W}$, si $u\mathcal{B}_i v$, alors $|u|_\Sigma \mathcal{B}_i^f |v|_\Sigma$.

- Supposons que $u\mathcal{B}_i v$ et prenons $\langle B_i\rangle\phi \in \Sigma$ tel que $\mathcal{M}, v \models \phi \vee \langle B_i\rangle\phi \vee B_i\phi$;
- Comme $u\mathcal{B}_i v$, nous avons $\mathcal{M}, u \models \langle B_i\rangle(\phi \vee \langle B_i\rangle\phi \vee B_i\phi)$;
- Donc $\mathcal{M}, u \models \langle B_i\rangle\phi \vee \langle B_i\rangle\langle B_i\rangle\phi \vee \langle B_i\rangle B_i\phi$;

— Cependant $\models \langle B_i\rangle\langle B_i\rangle\phi \Rightarrow \langle B_i\rangle\phi$ et $\models \langle B_i\rangle B_i\phi \Rightarrow \langle B_i\rangle\phi$;

— Ainsi $\mathcal{M}, u \models \langle B_i\rangle\phi$ et $\models \langle B_i\rangle\phi \Rightarrow B_i\langle B_i\rangle\phi$;

— Par conséquent $\mathcal{M}, u \models B_i\langle B_i\rangle\phi$;

— Donc $|u|_\Sigma \mathcal{B}_i^f |v|_\Sigma$.

(2) Prouvons que pour tout $u, v \in \mathcal{W}$, si $|u|_\Sigma \mathcal{B}_i^f |v|_\Sigma$, alors $\forall \langle B_i\rangle\phi \in \Sigma \setminus \{\emptyset\}, \mathcal{M}, v \models \phi \Longrightarrow \mathcal{M}, u \models \langle B_i\rangle\phi$.

— Supposons que $|u|_\Sigma \mathcal{B}_i^f |v|_\Sigma$ et $\langle B_i\rangle\phi \in \Sigma$ tel que $\mathcal{M}, v \models \phi$;

— Comme $\models \phi \Rightarrow \phi \vee \langle B_i\rangle\phi \vee B_i\phi$, alors $\mathcal{M}, v \models \phi \vee \langle B_i\rangle\phi \vee B_i\phi$;

— Comme $|u|_\Sigma \mathcal{B}_i^f |v|_\Sigma$, nous avons $\mathcal{M}, u \models B_i\langle B_i\rangle\phi$;

— Cependant $\models B_i\langle B_i\rangle\phi \Rightarrow \langle B_i\rangle\phi$;

— Par conséquent $\mathcal{M}, u \models \langle B_i\rangle\phi$.

Concernant $\mathcal{T}_{i,j}^f$, la preuve est standard car il s'agit de la *plus large filtration* (Blackburn et al., 2002). Nous venons donc de montrer que $\mathcal{M}_\Sigma^f$ est une filtration de $\mathcal{M}$ à travers Σ.

(3) Prouvons que pour tout $\mathcal{R}_\Diamond \in \{\mathcal{B}_i\}_{i\in\mathcal{N}} \cup \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, \mathcal{R}_\Diamond^f$ est *sérielle* (i.e $\forall u \in \mathcal{W}, \exists v \in \mathcal{W}, w\mathcal{R}_\Diamond v$). La sérialité est immédiate puisque $\mathcal{R}_\Diamond^f$ est une filtration et donc avec la propriété (i) de la filtration, nous avons que $\mathcal{R}_\Diamond^f$ est aussi sérielle.

(4) Prouvons que $\mathcal{B}_i^f$ est transitive.

— Pour tout $u, v, w \in \mathcal{W}$, supposons que $|w|_\Sigma \mathcal{B}_i^f |u|_\Sigma$, et $|u|_\Sigma \mathcal{B}_i^f |v|_\Sigma$, et $\mathcal{M}, v \models \phi \vee \langle B_i\rangle\phi \vee B_i\phi$ avec $\langle B_i\rangle\phi \in \Sigma \setminus \{\emptyset\}$;

— Ainsi $\mathcal{M}, u \models \langle B_i\rangle B_i\phi$;

— Comme $\models \langle B_i\rangle B_i\phi \Rightarrow \langle B_i\rangle\phi$ et $\models \langle B_i\rangle\phi \Rightarrow (\phi \vee \langle B_i\rangle\phi \vee B_i\phi)$;

— Donc $\mathcal{M}, w \models B_i\langle B_i\rangle\phi$;

— Par conséquent $|w|_\Sigma \mathcal{B}_i^f |v|_\Sigma$.

(5) Prouvons que $\mathcal{B}_i^f$ est Euclidienne.

— Pour tout $u, v, w \in \mathcal{W}$, supposons que $|w|_\Sigma \mathcal{B}_i^f |u|_\Sigma$, et $|w|_\Sigma \mathcal{B}_i^f |v|_\Sigma$, et $\mathcal{M}, v \models \phi \vee \langle B_i\rangle\phi \vee B_i\phi$ avec $\langle B_i\rangle\phi \in \Sigma$;

— Ainsi $\mathcal{M}, w \models B_i\langle B_i\rangle\phi$.

Remarquons que pour chaque filtration, nous avons :

— si $|w|_\Sigma \mathcal{B}_i^f |u|_\Sigma$, alors $\forall \langle B_i\rangle\phi \in \Sigma : \mathcal{M}, u \models \phi \Longrightarrow \mathcal{M}, w \models \langle B_i\rangle\phi$;

— Ainsi, par contraposition, $\forall \langle B_i\rangle\phi \in \Sigma : \mathcal{M}, w \not\models \langle B_i\rangle\phi \Longrightarrow \mathcal{M}, u \not\models \phi$;

— C'est-à-dire, $\forall \langle B_i\rangle\phi \in \Sigma : \mathcal{M}, w \models \neg\langle B_i\rangle\phi \Longrightarrow \mathcal{M}, u \models \neg\phi$;

— Alors $\forall \langle B_i\rangle\phi \in \Sigma$, par substitution $\psi = \neg\phi$, $\mathcal{M}, w \models B_i\psi \Longrightarrow \mathcal{M}, u \models \psi$.

Si nous substituons $\psi = \langle B_i\rangle\phi$ alors :

— Comme $\mathcal{M}, w \models B_i\langle B_i\rangle\phi$ et $|w|_\Sigma \mathcal{B}_i^f |u|_\Sigma$, nous avons : $\mathcal{M}, u \models \langle B_i\rangle\phi$;

— De plus, $\models \langle B_i\rangle\phi \Rightarrow B_i\langle B_i\rangle\phi$. Donc $\mathcal{M}, u \models B_i\langle B_i\rangle\phi$;

— Par conséquent $|u|_\Sigma \mathcal{B}_i^f |v|_\Sigma$.

(6) Prouvons que le lien relationnel entre $\mathcal{B}_i^f$ et $\mathcal{T}_{i,j}^f$ préserve la transitivité pour $\mathcal{T}_{i,j}^f$.

— Pour tout $u, v, w \in \mathcal{W}$, supposons que $|w|_\Sigma \mathcal{B}_i^f |u|_\Sigma$, et $|u|_\Sigma \mathcal{T}_{i,j}^f |v|_\Sigma$, et $\mathcal{M}, v \models \phi$ avec $\langle B_i\rangle\langle T_{i,j}^s\rangle\phi \in \Sigma \setminus \{\emptyset\}$;

— Ainsi $\mathcal{M}, u \models \langle T_{i,j}^s\rangle\phi$, en substituant $\psi = \langle T_{i,j}^s\rangle\phi$;

— Donc $\mathcal{M}, u \models \psi \vee \langle B_i\rangle\psi \vee B_i\psi$;

— Comme $|w|_\Sigma \mathcal{B}_i^f |u|_\Sigma$, nous avons $\mathcal{M}, w \models B_i\langle B_i\rangle\psi$;

— De plus $\models B_i\langle B_i\rangle\psi \Rightarrow \langle B_i\rangle\psi$. Ainsi $\mathcal{M}, w \models \langle B_i\rangle\langle T_{i,j}^s\rangle\phi$;

— Et $\models \langle B_i\rangle\langle T_{i,j}^s\rangle\phi \Rightarrow \langle T_{i,j}^s\rangle\phi$;

— Par conséquent, $\mathcal{M}, w \models \langle T_{i,j}^s\rangle\phi$ et donc $|w|_\Sigma \mathcal{T}_{i,j}^f |v|_\Sigma$.

(7) Prouvons que le lien relationnel entre $\mathcal{B}_i^f$ et $\mathcal{T}_{i,j}^f$ est Euclidien pour $\mathcal{T}_{i,j}^f$.

— Pour tout $u, v, w \in \mathcal{W}$, supposons que $|w|_\Sigma \mathcal{B}_i^f |u|_\Sigma$, $|w|_\Sigma \mathcal{T}_{i,j}^f |v|_\Sigma$ et $\mathcal{M}, v \models \phi$ avec $\langle B_i\rangle\langle T_{i,j}^s\rangle\phi \in \Sigma$;

— Ainsi, puisque $|w|_\Sigma \mathcal{T}_{i,j}^f |v|_\Sigma$, nous avons $\mathcal{M}, w \models \langle T_{i,j}^s\rangle\phi$;

— De plus, $\models \langle T_{i,j}^s\rangle\phi \Rightarrow B_i\langle T_{i,j}^s\rangle\phi$;

— Donc $\mathcal{M}, w \models B_i\psi$ en substituant $\psi = \langle T_{i,j}^s\rangle\phi$;

— Ainsi puisque $|w|_\Sigma \mathcal{B}_i^f |u|_\Sigma$, nous avons $\mathcal{M}, u \models \psi$, i.e. $\mathcal{M}, u \models \langle T_{i,j}^s\rangle\phi$;

— Par conséquent $|u|_\Sigma \mathcal{T}_{i,j}^f |v|_\Sigma$.

(8) Prouvons que le lien relationnel entre $\mathcal{B}_i^f$ et $\mathcal{B}_j^f$ est transitif par rapport à $\mathcal{T}_{i,j}$.

— Pour tout $u, v, w \in \mathcal{W}$, supposons que $|w|_\Sigma \mathcal{B}_i^f |u|_\Sigma$, $|u|_\Sigma \mathcal{B}_j^f |v|_\Sigma$ et $\mathcal{M}, v \models \phi$ avec $\langle B_i\rangle\langle B_j\rangle\phi \in \Sigma$ et $\langle T_{i,j}^s\rangle\phi \in \Sigma$;

— Puisque $\models \phi \Rightarrow (\phi \vee \langle B_j\rangle\phi \vee B_j\phi)$, nous avons $\mathcal{M}, v \models \phi \vee \langle B_j\rangle\phi \vee B_j\phi$;

— Puisque $|u|_\Sigma \mathcal{B}_j^f |v|_\Sigma$, nous avons $\mathcal{M}, v \models B_j\langle B_j\rangle\phi$;

— Puisque $\models B_j\langle B_j\rangle\phi \Rightarrow \langle B_j\rangle\phi$, nous avons $\mathcal{M}, v \models \langle B_j\rangle\phi$

— En substituant $\psi = \langle B_j\rangle\phi$, nous déduisons immédiatement par des arguments analogues que $\mathcal{M}, u \models \psi \vee \langle B_i\rangle\psi \vee B_i\psi$;

— Ensuite, puisque $|w|_\Sigma \mathcal{B}_i^f |u|_\Sigma$, nous avons $\mathcal{M}, w \models B_i\langle B_i\rangle\psi$;

— Comme $\models B_i\langle B_i\rangle\psi \Rightarrow \langle B_i\rangle\psi$, nous déduisons $\mathcal{M}, w \models \langle B_i\rangle\langle B_j\rangle\phi$;

— Cependant $\models \langle B_i\rangle\langle B_j\rangle\phi \Rightarrow \langle T_{i,j}^s\rangle\phi$, ainsi $\mathcal{M}, w \models \langle T_{i,j}^s\rangle\phi$;

— Par conséquent, nous avons prouvé que $|w|_\Sigma \mathcal{T}_{i,j}^f |v|_\Sigma$.

□

S'il existe un modèle $\mathcal{M}$ satisfaisant un ensemble de Hintikka H sur un ensemble fermé de formules Σ, alors il existe une TB-filtration $\mathcal{M}_\Sigma^f$ de ce modèle à travers Σ satisfaisant H, et réciproquement. Ce résultat est une conséquence directe de l'application du théorème 5.3.

Théorème 5.3 - *Théorème de filtration* : *Soient Σ un ensemble fermé de formules, un modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$, et $\mathcal{M}^f_\Sigma = (\mathcal{W}_\Sigma, \{\mathcal{B}^f_i\}_{i\in\mathcal{N}}, \{\mathcal{T}^f_{i,j}\}_{i,j\in\mathcal{N}}, V^f)$ une* filtration *de $\mathcal{M}$ à travers Σ.*

$$\forall\phi \in \Sigma, \forall u \in \mathcal{W} : \mathcal{M}, u \models \phi \text{ si, et seulement si, } \mathcal{M}^f_\Sigma, |u|_\Sigma \models \phi$$

Démonstration. Il s'agit d'une preuve standard par induction sur le degré d'une formule (Blackburn et al., 2002). □

En résumé, jusqu'à présent pour démontrer la complétude de la méthode, nous avons :

— Si H est un TB-atome sur Σ, alors il existe $\mathcal{M}, w \models H$ et donc $\mathcal{M}_w, w \models H$;

— Par application du théorème de filtration, il existe une TB-filtration $\mathcal{M}^f_{w,\Sigma}$ de $\mathcal{M}_w$ à travers Σ qui préserve toutes les propriétés du cadre TB et telle que :

$$\forall\phi \in \Sigma, \forall u \in \mathcal{W}_w : \mathcal{M}_w, u \models \phi \text{ ssi } \mathcal{M}^f_{w,\Sigma}, |u|_\Sigma \models \phi$$

— Ainsi, à ce stade de la preuve, nous avons : $\mathcal{M}^f_{w,\Sigma}, |w|_\Sigma \models H$.

Étape 3 : Construction d'un générateur d'ensembles témoins à partir de la filtration Nous montrons maintenant qu'à chaque classe d'équivalence de $\mathcal{W}^f_\Sigma$ dans la TB-filtration $\mathcal{M}^f_{w,\Sigma}$, correspond un élément d'un ensemble TB-témoin généré par l'ensemble de Hintikka H sur Σ. Ainsi, un TB-atome implique l'existence d'un ensemble TB-témoin.

Théorème 5.4 - *Complétude* : *Si H est un TB-atome sur Σ où Σ est un ensemble fermé de formules, alors il existe un ensemble TB-témoin généré par H sur Σ.*

Démonstration. Supposons que H est un TB-atome sur Σ avec Σ un ensemble fermé de formules. Il existe un modèle $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, V)$ et $w \in \mathcal{W}$ tel que $\mathcal{M}, w \models H$. Si $\mathcal{M}_w$ est un modèle TB-connexe autour de w, alors $\mathcal{M}^f_{w,\Sigma} = (\mathcal{W}_{w,\Sigma}, \{\mathcal{B}^f_i\}_{i\in\mathcal{N}}, \{\mathcal{T}^f_{i,j}\}_{i,j\in\mathcal{N}}, V^f)$ est une TB-filtration de $\mathcal{M}_w$ à travers Σ.

Nous définissons la fonction $g\colon \mathcal{W}_{w,\Sigma} \to 2^\Sigma$ telle que :

$$\begin{array}{rccl} g\colon & \mathcal{W}_{w,\Sigma} & \to & 2^\Sigma \\ & |u|_\Sigma & \mapsto & \{\phi \in \Sigma : \mathcal{M}^f_{w,\Sigma}, |u|_\Sigma \models \phi\} \end{array}$$

Nous avons immédiatement par le théorème de filtration que $H = g(|w|_\Sigma)$ (cf. théorème 5.3).

Dans la suite, nous posons[3] l'ensemble $\mathcal{G}_H^{TB}$ tel que :

$$\mathcal{G}_H^{TB} = \{\{H\}\} \bigcup_{\substack{|v|_\Sigma \in \mathcal{W}_\Sigma \\ \mathcal{R} \in \{\mathcal{B}_i^f, \mathcal{T}_{i,j}^f\}_{i,j \in \mathcal{N}}}} \{\{g(|u|_\Sigma) : |u|_\Sigma \in \mathcal{W}_\Sigma, |v|_\Sigma (\mathcal{R})^* |u|_\Sigma \text{ ou } |v|_\Sigma (\mathcal{R}^{-1})^* |u|_\Sigma\}\}$$

Montrons que l'ensemble $\mathcal{G}_H^{TB}$ est un générateur d'ensembles TB-témoins à partir de H sur Σ et donc que $\mathcal{H}_H = \{g(|u|_\Sigma) : |u|_\Sigma \in \mathcal{W}_\Sigma\}$ est un ensemble TB-témoin généré par H sur Σ.

Pour tout $|v|_\Sigma \in \mathcal{W}_{w,\Sigma}$ et pour tout $\mathcal{R} \in \{\mathcal{B}_i^f, \mathcal{T}_{i,j}^f\}_{i,j \in \mathcal{N}}$, nous notons :

$$\mathcal{C}^{\mathcal{R}}_{|v|_\Sigma} = \{|u|_\Sigma \in \mathcal{W}_\Sigma : |v|_\Sigma (\mathcal{R})^* |u|_\Sigma \text{ ou } |v|_\Sigma (\mathcal{R}^{-1})^* |u|_\Sigma\}$$

Remarquons que comme $|v|_\Sigma \in \mathcal{C}^{\mathcal{R}}_{|v|_\Sigma}$, nous avons $\mathcal{C}^{\mathcal{R}}_{|v|_\Sigma} \neq \emptyset$ et donc pour tout $|v|'_\Sigma \in \mathcal{W}_{w,\Sigma}$ et $\mathcal{R}' \in \{\mathcal{B}_i^f, \mathcal{T}_{i,j}^f\}_{i,j \in \mathcal{N}}$, il existe toujours un ensemble $\mathcal{C}^{\mathcal{R}'}_{|v|'_\Sigma}$. De plus, puisque $\mathcal{M}^f_{w,\Sigma}$ est TB-connexe par la propriété (i) de la filtration, pour tout $|u|_\Sigma \in \mathcal{C}^{\mathcal{R}}_{|v|_\Sigma}$, il existe un TB-chemin de $|w|_\Sigma$ à $|u|_\Sigma$ dans $\mathcal{M}^f_{w,\Sigma}$. Ainsi, si nous notons $\mathcal{P}(\{|w|_\Sigma\}, \mathcal{C}^{\mathcal{R}}_{|v|_\Sigma}) = (\mathcal{C}_0, \ldots, \mathcal{C}_n)$ le $(n+1) - uplet$ tel que $\forall k \in [|1, n|], \exists |x|_\Sigma, |y|_\Sigma \in \mathcal{W}_{w,\Sigma}, \exists \mathcal{R}', \mathcal{R}'' \in \{\mathcal{B}_i^f, \mathcal{T}_{i,j}^f\}_{i,j \in \mathcal{N}}, \mathcal{C}_{k-1} = \mathcal{C}^{\mathcal{R}'}_{|x|_\Sigma}$, $\mathcal{C}_k = \mathcal{C}^{\mathcal{R}''}_{|y|_\Sigma}$ et $\exists |x'|_\Sigma \in \mathcal{C}_{k-1}, |y'|_\Sigma \in \mathcal{C}_k$ tel que $|x'|_\Sigma \mathcal{R}'' |y'|_\Sigma$ où $\mathcal{C}_0 = \{|w|_\Sigma\}$ et $\mathcal{C}_n = \mathcal{C}^{\mathcal{R}}_{|v|_\Sigma}$, par les arguments qui précèdent, il existe nécessairement $\mathcal{P}(\{|w|_\Sigma\}, \mathcal{C}^{\mathcal{R}}_{|v|_\Sigma})$, i.e un chemin de $\{|w|_\Sigma\}$ à $\mathcal{C}^{\mathcal{R}}_{|v|_\Sigma}$.

Nous posons $\mathcal{H}^{\Diamond}_{H',|v|_\Sigma} := \{g(|u|_\Sigma) : u \in \mathcal{C}^{\mathcal{R}_\Diamond}_{|v|_\Sigma}\}$ avec $\mathcal{R}_\Diamond$ la relation associée à $\Diamond$ et $H' = g(|h|_\Sigma)$, où $|h|_\Sigma \in \mathcal{C}^{\mathcal{R}_\Diamond}_{|v|_\Sigma}$ est tel que : $\forall |x|_\Sigma \in \mathcal{C}^{\mathcal{R}_\Diamond}_{|v|_\Sigma} : |h|_\Sigma \mathcal{R}^*_\Diamond |x|_\Sigma$.

De plus, nous définissons $\mathcal{Q}(\{H\}, \mathcal{H}^{\mathcal{R}_\Diamond}_{H',|v|_\Sigma}) = (\mathcal{H}_0, \ldots, \mathcal{H}_n)$ le $(n+1) - uplet$, $n \in \mathbb{N}^*$, tel que $\mathcal{H}_0 = \{\{H\}\}$, $\mathcal{H}_n = \mathcal{H}^{\mathcal{R}_\Diamond}_{H',|v|_\Sigma}$ et $\forall k \in [|1, n|], \exists K \in \mathcal{H}_{k-1}, \exists \Gamma \subseteq \Sigma : \mathcal{H}_{k-1} \diamond' \mathcal{H}^{\Diamond'}_{K,\Gamma}$, $\mathcal{H}_k = \mathcal{H}^{\Diamond'}_{K,\Gamma}$ et $\mathcal{H}^{\Diamond'}_{K,\Gamma}$ est un $\diamond'$-témoin généré par K et contraint par Γ.

Remarquons que nous avons par ces définitions que pour tout $\mathcal{C}^{\mathcal{R}_\Diamond}_{|v|_\Sigma} \neq \emptyset$, $\mathcal{H}^{\mathcal{R}_\Diamond}_{H',|v|_\Sigma} \in \mathcal{G}_H^{TB}$ où il existe $H' = g(|h|_\Sigma)$ tel que $|h|_\Sigma \in \mathcal{C}^{\mathcal{R}_\Diamond}_{|v|_\Sigma}$, $\forall |x|_\Sigma \in \mathcal{C}^{\mathcal{R}_\Diamond}_{|v|_\Sigma}, |h|_\Sigma \mathcal{R}^*_\Diamond |x|_\Sigma$.

Puisque le modèle $\mathcal{M}^f_{w,\Sigma}$ contient au maximum $2^{|\Sigma|}$ mondes (preuve triviale cf. (Blackburn et al., 2002)), l'ensemble $\mathcal{G}_H^{TB}$ est fini. Ainsi, en construisant par récurrence sur la longueur des chemins $\mathcal{Q}(\{H\}, \mathcal{H}^{\mathcal{R}_\Diamond}_{H',|v|_\Sigma})$, nous montrons que : $\forall \mathcal{H}^{\mathcal{R}_\Diamond}_{H',|v|_\Sigma} \in \mathcal{G}_H^{TB}, \exists \mathcal{Q}(\{H\}, \mathcal{H}^{\mathcal{R}_\Diamond}_{H',|v|_\Sigma}) \neq \emptyset$.

Il est alors immédiat que $\mathcal{G}_H^{TB}$ est un générateur d'ensemble TB-témoin généré à partir de H et donc $\mathcal{H}_H = \{g(|u|_\Sigma) : |u|_\Sigma \in \mathcal{W}_\Sigma\}$ est un ensemble TB-témoin généré par H sur Σ. □

5.2.3 Réciproque du théorème

Prouvons maintenant la réciproque de ce théorème qui est la correction de la méthode. S'il existe un ensemble témoin généré à partir d'un ensemble de Hintikka, alors cet ensemble de Hintikka est satisfiable. Pour démontrer cette partie du théorème, la méthode consiste à construire un modèle de façon récursive en parcourant l'arbre formé par générateur d'ensembles témoins et

3. Nous rappelons que pour une relation binaire $\mathcal{R}$, la relation binaire inverse correspond à $\mathcal{R}^{-1} = \{(x, y) : (y, x) \in \mathcal{R}\}$ et la clôture transitive réflexive correspond à $\mathcal{R}^* = \{(x, z) : \exists n \in \mathbb{N}^*, y_0, \ldots, y_n \text{ tq } y_0 = x, y_n = z, (y_0, y_1) \in \mathcal{R}, \ldots, (y_{n-1}, y_n) \in \mathcal{R} \text{ ou } x = z\}$.

en appliquant les contraintes du cadre TB sur ce modèle. Le modèle ainsi obtenu satisfait cet ensemble de Hintikka.

Théorème 5.5 - *Correction :* *S'il existe un ensemble $\mathcal{H}_H$ TB-témoin généré par H sur Σ où Σ est un ensemble fermé de formules et H un ensemble de Hintikka sur Σ, alors H est un TB-atome sur Σ.*

Démonstration. Soient Σ un ensemble fermé de formules, H un ensemble de Hintikka sur Σ et un ensemble $\mathcal{H}_H$ TB-témoin généré par H sur Σ. Donc il existe un générateur d'ensemble TB-témoin $\mathcal{G}_H^{TB}$ généré par H. Considérons un ensemble fini et dénombrable $\mathcal{W} = \{w_0, w_1, w_2, \ldots\} \cup \{w_\top\}$ et posons le cadre $\mathcal{F} = (\mathcal{F}_n)_{n\in\mathbb{N}} = (\mathcal{W}_n, \{\mathcal{B}_i^n\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}^n\}_{i,j\in\mathcal{N}}, f_n)_{n\in\mathbb{N}}$ où $\mathcal{W}_0 = \{w_0\}$, pour tout $i, j \in \mathcal{N}$, $\mathcal{B}_i^0 = \emptyset$, $\mathcal{T}_{i,j}^0 = \emptyset$ et $f_0(w_0) = H$.

Soit $n \in \mathbb{N}$, $\mathcal{F}$ est construit tel que :

1. s'il existe $w \in \mathcal{W}_n$ et $\langle B_j\rangle\phi \in f_n(w)$ ou $B_j\phi \in f_n(w)$, $f_n(w) \in \mathcal{H}_{H',\Gamma}^{B_j}$ avec $\mathcal{H}_{H',\Gamma}^{B_j} \in \mathcal{G}_H^{TB}$, $H' = f_n(r'), r' \in \mathcal{W}_n$ et **il n'existe pas** $w' \in \mathcal{W}_n$ tel que $w\mathcal{B}_j^n w'$, $f_n(w') \in \mathcal{H}_{H',\Gamma}^{B_j}$, et $\phi \in f_n(w')$, alors :
   - $\mathcal{W}_{n+1} = \mathcal{W}_n \cup \{w_{n+1}\}$;
   - $\mathcal{B}_j^{n+1} = \mathcal{B}_j^n \cup \{(r', w_{n+1})\} \cup \{(w_{n+1}, w_{n+1}\} \bigcup\limits_{\substack{v'\in\mathcal{W}_n:\\ r'\mathcal{B}_j^n v'}} \{(v', w_{n+1}), (w_{n+1}, v')\}$;
   - $\mathcal{T}_{j,j}^{n+1} = \mathcal{T}_{j,j}^n \cup \{(r', w_{n+1})\} \cup \{(w_{n+1}, w_{n+1}\} \bigcup\limits_{\substack{v'\in\mathcal{W}_n:\\ r'\mathcal{B}_j^n v'}} \{(v', w_{n+1}), (w_{n+1}, v')\}$;
   - s'il existe $i \neq j$ et $\mathcal{H}_{H'',\Gamma'}^{B_i} \in \mathcal{G}_H^{TB}$, avec $H'' = f_n(r''), r'' \in \mathcal{W}_n$ tel que $\mathcal{H}_{H'',\Gamma'}^{B_i}\langle B_j\rangle\mathcal{H}_{H',\Gamma}^{B_j}$ alors :

$$\mathcal{T}_{i,j}^{n+1} = \mathcal{T}_{i,j}^n \cup \{(r'', w_{n+1})\} \bigcup_{\substack{v''\in\mathcal{W}_n:\\ r''\mathcal{B}_i^n v''}} \{(v'', w_{n+1})\}$$

$$\forall k \in \mathcal{N} \setminus i, \forall l \in \mathcal{N}, \mathcal{T}_{k,l}^{n+1} = \mathcal{T}_{k,l}^n$$

     sinon $\forall k, l \in \mathcal{N}, \mathcal{T}_{k,l}^{n+1} = \mathcal{T}_{k,l}^n$;
   - $\forall k \in \mathcal{N} \setminus j, \mathcal{B}_k^{n+1} = \mathcal{B}_k^n$;
   - $f_{n+1} = f_n \cup \{(w_{n+1}, I)\}$ où $I \in \mathcal{H}_{H',\Gamma}^{B_j}$ est tel que $f_n(w)\langle B_j\rangle_\phi I$.
2. sinon s'il existe $w \in \mathcal{W}_n$ et $\langle T_{i,j}^s\rangle\phi \in f_n(w)$ ou $T_{i,j}^s\phi \in f_n(w)$, $f_n(w) \in \mathcal{H}_{H',\Gamma}^{T_{i,j}^s}$ avec $\mathcal{H}_{H',\Gamma}^{T_{i,j}^s} \in \mathcal{G}_H^{TB}$, $H' = f_n(r'), r' \in \mathcal{W}_n$ et **il n'existe pas** $w' \in \mathcal{W}_n$ tel que $w\mathcal{T}_{i,j}^n w'$, $f_n(w') \in \mathcal{H}_{H',\Gamma}^{T_{i,j}^s}$, $\phi \in f_n(w')$ alors :
   - $\mathcal{W}_{n+1} = \mathcal{W}_n \cup \{w_{n+1}\}$;
   - $\mathcal{T}_{i,j}^{n+1} = \mathcal{T}_{i,j}^n \cup \{(w, w_{n+1})\} \bigcup\limits_{\substack{v'\in\mathcal{W}_n:\\ w\mathcal{B}_i^n v'}} \{(v', w_{n+1})\} \bigcup\limits_{\substack{v'\in\mathcal{W}_n:\\ v'\mathcal{B}_i^n w}} \{(v', w_{n+1})\}$;

— $\forall k \in \mathcal{N}, \mathcal{B}_k^{n+1} = \mathcal{B}_k^n$;
— $\forall (k,l) \in \mathcal{N}^2 \setminus (i,j), \mathcal{T}_{k,l}^{n+1} = \mathcal{T}_{k,l}^n$;
— $f_{n+1} = f_n \cup \{(w_{n+1}, I)\}$ où $I \in \mathcal{H}_{H',\Gamma}^{T_{i,j}^s}$ tel que $f_n(w)\langle T_{i,j}^s\rangle_\phi$.

3. s'il existe $w \in \mathcal{W}_n$ et **il n'existe ni** $\langle B_j\rangle\phi \in f_n(w)$, **ni** $B_j\phi \in f_n(w)$, $f_n(w) \in \mathcal{H}_{H',\Gamma}^{B_j}$: $f_n(w) = H'$ avec $\mathcal{H}_{H',\Gamma}^{B_j} \in \mathcal{G}_H^{TB}$, $H' = f_n(r'), r' \in \mathcal{W}_n$ et **il n'existe pas** $w_\Gamma \in \mathcal{W}_n$ tel que $w\mathcal{B}_j^n w_\Gamma$, $f_n(w_\Gamma) \in \mathcal{H}_{H',\Gamma}^{B_j}$, $\Gamma \subseteq f_n(w_\Gamma)$ alors :
— $\mathcal{W}_{n+1} = \mathcal{W}_n \cup \{w_{n+1}\}$;
— $\mathcal{B}_j^{n+1} = \mathcal{B}_j^n \cup \{(r', w_{n+1})\} \cup \{(w_{n+1}, w_{n+1}\} \bigcup_{\substack{v' \in \mathcal{W}_n : \\ r'\mathcal{B}_j^n v'}} \{(v', w_{n+1}), (w_{n+1}, v')\}$;
— $\mathcal{T}_{j,j}^{n+1} = \mathcal{T}_{j,j}^n \cup \{(r', w_{n+1})\} \cup \{(w_{n+1}, w_{n+1}\} \bigcup_{\substack{v' \in \mathcal{W}_n : \\ r'\mathcal{B}_j^n v'}} \{(v', w_{n+1}), (w_{n+1}, v')\}$;
— s'il existe $i \neq j$ et $\mathcal{H}_{H'',\Gamma'}^{B_i} \in \mathcal{G}_H^{TB}$, avec $H'' = f_n(r''), r'' \in \mathcal{W}_n$ tel que $\mathcal{H}_{H'',\Gamma'}^{B_i}\langle B_j\rangle\mathcal{H}_{H',\Gamma}^{B_j}$ alors :

$$\mathcal{T}_{i,j}^{n+1} = \mathcal{T}_{i,j}^n \cup \{(r'', w_{n+1})\} \bigcup_{\substack{v'' \in \mathcal{W}_n : \\ r''\mathcal{B}_i^n v''}} \{(v'', w_{n+1})\}$$

$$\forall k \in \mathcal{N} \setminus i, \forall l \in \mathcal{N}, \mathcal{T}_{k,l}^{n+1} = \mathcal{T}_{k,l}^n$$

sinon $\forall k, l \in \mathcal{N}, \mathcal{T}_{k,l}^{n+1} = \mathcal{T}_{k,l}^n$;
— $\forall k \in \mathcal{N} \setminus j, \mathcal{B}_k^{n+1} = \mathcal{B}_k^n$;
— $f_{n+1} = f_n \cup \{(w_{n+1}, I)\}$ où $I \in \mathcal{H}_{H',\Gamma}^{B_j}$ est tel que $\Gamma \subseteq I$.

4. s'il existe $w \in \mathcal{W}_n$ et **il n'existe ni** $\langle T_{i,j}^s\rangle\phi \in f_n(w)$, **ni** $T_{i,j}^s\phi \in f_n(w)$, $f_n(w) \in \mathcal{H}_{H',\Gamma}^{T_{i,j}^s}$: $f_n(w) = H'$ avec $\mathcal{H}_{H',\Gamma}^{T_{i,j}^s} \in \mathcal{G}_H^{TB}$, $H' = f_n(r'), r' \in \mathcal{W}_n$ et **il n'existe pas** $w_\Gamma \in \mathcal{W}_n$ tel que $w\mathcal{T}_{i,j}^n w_\Gamma$, $f_n(w_\Gamma) \in \mathcal{H}_{H',\Gamma}^{T_{i,j}^s}$, $\Gamma \subseteq f_n(w_\Gamma)$ alors :
— $\mathcal{W}_{n+1} = \mathcal{W}_n \cup \{w_{n+1}\}$;
— $\mathcal{T}_{i,j}^{n+1} = \mathcal{T}_{i,j}^n \cup \{(w, w_{n+1})\} \bigcup_{\substack{v' \in \mathcal{W}_n : \\ w\mathcal{B}_i^n v'}} \{(v', w_{n+1})\} \bigcup_{\substack{v' \in \mathcal{W}_n : \\ v'\mathcal{B}_i^n w}} \{(v', w_{n+1})\}$;
— $\forall k \in \mathcal{N}, \mathcal{B}_k^{n+1} = \mathcal{B}_k^n$;
— $\forall (k,l) \in \mathcal{N}^2 \setminus (i,j), \mathcal{T}_{k,l}^{n+1} = \mathcal{T}_{k,l}^n$;
— $f_{n+1} = f_n \cup \{(w_{n+1}, I)\}$ où $I \in \mathcal{H}_{H',\Gamma}^{T_{i,j}^s}$ est tel que $\Gamma \subseteq I$.

5. sinon le processus de construction se termine,
— $\mathcal{W}_{fin} = \mathcal{W}_n \cup \{w_\top\}$;
— $\forall i \in \mathcal{N}, \mathcal{B}_i^{fin} = \mathcal{B}_i^n \cup \{(w_\top, w_\top)\} \bigcup_{\substack{w \in \mathcal{W}_n : \\ \nexists v \in \mathcal{W}_n, w\mathcal{B}_i^n v}} \{(w, w_\top)\}$;

— $\forall i,j \in \mathcal{N}, \mathcal{T}_{i,j}^{fin} = \mathcal{T}_{i,j}^{n} \bigcup_{\substack{w\in\mathcal{W}_n:\\ \nexists v\in\mathcal{W}_n, w\mathcal{T}_{i,j}^{n} v}} (\{(w_\top, w_\top), (w, w_\top)\} \bigcup_{\substack{v\in\mathcal{W}_n:\\ (w\mathcal{B}_i^n v \vee v\mathcal{B}_i^n w)\\ \wedge \nexists u\in\mathcal{W}_n, w\mathcal{T}_{i,j}^{n} u\\ \text{ou } (v\mathcal{B}_i^n w)\\ \wedge \nexists u\in\mathcal{W}_n, w\mathcal{B}_j^n u}} \{(v, w_\top)\})$;

— $f_{fin} = f_n$.

Les relations $\Diamond'$ entre $\mathcal{H}^{\Diamond}_{H',\Gamma} \Diamond' \mathcal{H}^{\Diamond'}_{H'',\Gamma'}$ de $\mathcal{G}_H^{TB}$ forment un arbre de profondeur au plus $deg(\Sigma)$ et nous avons $\mathcal{H}^{\Diamond'}_{H'',\Gamma'} \setminus H'' < deg(\mathcal{H}^{\Diamond}_{H',\Gamma})$. Puisque tous les $\mathcal{H}^{\Diamond}_{H',\Gamma} \in \mathcal{G}_H^{TB}$ sont des ensembles finis d'ensembles de formules eux mêmes finis, la procédure pour construire $\mathcal{F}$ termine après un nombre fini d'étapes. Soit $m \in \mathbb{N}^*$ l'étape à partir de laquelle la procédure se termine, et considérons $\mathcal{C} = (\mathcal{W}_m, \{\mathcal{B}_i^m\}_{i\in\mathcal{N}}, \{\mathcal{T}_{i,j}^m\}_{i,j\in\mathcal{N}})$ et le modèle $\mathcal{M} = (\mathcal{C}, V)$ tel que $w \in V(p)$ si, et seulement si $p \in f_m(w)$. Il suffit alors de montrer que $\mathcal{M}, w_0 \models H$, ce qui est trivial à démontrer en appliquant un raisonnement par récurrence sur le degré de Σ et en appliquant la définition du cadre $\mathcal{F}_n$. Par conséquent, nous venons de montrer que H est un TB-atome. □

5.3 Un algorithme pour résoudre le problème SAT dans TB

La section 5.2 a présenté une nouvelle méthode des tableaux pour résoudre le problème de satisfiabilité dans le cadre de TB. Cette méthode nous permet de définir un algorithme prenant en entrée un ensemble de formules et construit un ensemble TB-témoin afin de vérifier la satisfiabilité de cet ensemble. Cette section s'organise en trois parties : la section 5.3.1 présente l'algorithme, puis la section 5.3.2 présente un ensemble de bibliothèques et logiciels pour les logiques modales et la section 5.3.3 présente une instanciation de la méthode.

5.3.1 Description de l'algorithme

L'algorithme de la méthode est décomposé en quatre algorithmes différents :

1. l'algorithme 1 est le point d'entrée de la méthode et prend en paramètre un ensemble de formules Γ dont on cherche à vérifier sa satisfiabilité. Cet algorithme initialise l'ensemble de Hintikka H sur $Cl(\Gamma)$ et contenant l'ensemble de formules Γ. Cet algorithme génère l'ensemble TB-témoin et appelle l'algorithme 2 permettant de construire les ensembles $\langle B_i\rangle$-témoins $\mathcal{H}_H^{\langle B_i\rangle}$ pour tout agent $i \in \mathcal{N}$;
2. l'algorithme 2 construit les ensembles $\langle B_i\rangle$-témoins $\mathcal{H}_{H'}^{\langle B_i\rangle}$ et appelle l'algorithme 3 permettant de générer les autres ensembles $\Diamond$-témoins qui doivent être générés depuis les éléments $I \in \mathcal{H}_{H'}^{\langle B_i\rangle} \setminus H'$;
3. l'algorithme 3 construit les autres ensembles $\langle T_{i,k}^s\rangle$-témoins à partir des ensembles $H'' \in \mathcal{H}_{H'}^{\langle B_i\rangle}$ en appelant l'algorithme 4, puis appelle l'algorithme 2 pour construire les autres ensembles $\langle B_k\rangle$-témoins $\mathcal{H}_J^{\langle B_k\rangle}$ avec $J \in \mathcal{H}_{H'}^{\langle B_i\rangle} \setminus H'$ pour tout $k \in \mathcal{N}, k \neq i$;
4. l'algorithme 4 construit les ensembles $\langle T_{i,j}^s\rangle$-témoins $\mathcal{H}_{H''}^{\langle T_{i,j}^s\rangle}$ et appelle l'algorithme 2 pour construire les autres ensembles témoins générés par des éléments de $K \in \mathcal{H}_{H''}^{\langle T_{i,j}^s\rangle} \setminus \{H''\}$.

Cet algorithme termine lorsqu'il ne peut plus générer de structures $\mathcal{H}^{\Diamond}_{H'} \neq \{H'\}$. La suite de la section présente en détails ces différents algorithmes.

Soit Γ un ensemble de formules de TB. L'algorithme 1 prend en entrée cet ensemble de formules Γ et renvoie la valeur Vrai si l'algorithme a pu calculer un ensemble TB-témoin et renvoie Faux dans le cas contraire. La première étape de l'algorithme consiste à calculer la fermeture Σ de l'ensemble Γ, puis de calculer un ensemble de Hintikka H sur Σ. En premier lieu, nous construisons, pour tout agent $i \in \mathcal{N}$, les structures $\mathcal{H}^{\langle B_i \rangle}_H$ avec la fonction récursive $WitnessB_i()$ [4]. Au moment de la construction de cette structure, nous mémorisons l'ensemble des modalités, $T^s_{i,j}$ présentes dans cette structure, pour tout agent $j \in \mathcal{N}$. Une fois cette structure calculée, nous construisons les structures $\mathcal{H}^{\langle T^s_{i,j} \rangle}_H$, en appelant la fonction $WitnessT^s_{i,j}()$ et en passant en paramètre la liste des $T^s_{i,j}$ mémorisés à l'étape de construction de la structure $\mathcal{H}^{\langle B_i \rangle}_H$. Ensuite, pour tout ensemble $I \in \mathcal{H}^{\langle B_i \rangle}_H \setminus \{H\}$, nous construisons les autres structures $\mathcal{H}^{\langle B_k \rangle}_I$ avec $k \neq i$, en appelant la fonction $WitnessB_k()$.

Algorithme 1 Witness(Γ)

1: $\Sigma_H \leftarrow Cl(\Gamma)$
2: **si** il existe un ensemble de Hintikka H sur Σ_H contenant Γ et tel que :
3: $\forall i \in \mathcal{N}$, Witness$B_i$(H, Σ_H, $\emptyset$, $\emptyset$, $\emptyset$, $\emptyset$, $\emptyset$, $\emptyset$) **alors :**
4: **retourner** $\top$
5: **sinon**
6: **retourner** $\bot$
7: **fin si**

L'algorithme 2 a pour rôle de construire les ensembles $\mathcal{H}^{B_j}_H$ en appelant la fonction récursive Witness$B_j()$ [5], pour tout agent $j \in \mathcal{N}$. Cette fonction prend différents paramètres en entrée :

H : l'ensemble de Hintikka sur la fermeture Σ_H qui est la racine de la structure $\mathcal{H}^{B_j}_H$ que nous cherchons à calculer ;

Σ_H : la fermeture de l'ensemble de Hintikka H ;

$\mathcal{H}^{B_i}_G$: une structure qui précède celle que nous sommes en train de construire ;

$\mathcal{L}_{B_j}$: l'ensemble des formules qui ont été déjà calculées dans la structure ;

$\overline{\mathcal{L}_{B_j}}$: l'ensemble des formules restant à calculer ;

$\Box^{B_j}_H$: l'ensemble des modalités B_j à appliquer sur les ensembles de la structure ;

S : la structure temporaire calculée comme un ensemble de couples $(I, \Box)$ avec I l'ensemble de Hintikka calculé actuellement avec l'ensemble des formules $\Box = \{\theta : B_j\theta \in \mathcal{L}_{B_j} \cup \{H\}\}$;

$\mathcal{H}^{B_j}_H$: représente le résultat de la nouvelle structure calculée ;

4. Pour des raisons de clarté de l'algorithme, et pour faciliter la lecture de celui-ci, nous sortons le paramètre agent i sur le nom de la fonction. Cette manière d'écrire est bien évidemment équivalente à noter $WitnessB(i, \ldots)$.

5. Ici, l'indice de l'ensemble témoin est bien j et non i. En effet, dans l'algorithme 2, l'indice i est réservé pour l'ensemble $\langle B_i \rangle$-témoin qui précède l'ensemble $\langle B_j \rangle$-témoin $\mathcal{H}^{B_j}_H$ que nous calculons dans la fonction $WitnessB_j()$.

La première étape de la fonction Witness$B_j()$ consiste à vérifier si la structure S n'est pas vide. Si tel est le cas, nous initialisons alors la structure $\mathcal{H}_H^{B_j}$. Nous calculons tout d'abord l'ensemble $\Box_H^{B_j}$ qui représente l'ensemble des modalités $\Box$ pour calculer les extensions de Hintikka. À cet ensemble, nous ajoutons les formules ϕ telles que $B_j\phi \in H$ et les formules ψ telles que $T^s_{j,j}\psi \in H$ ainsi que les formules ϕ telles que $T^s_{i,j}\phi \in I$, où $I \in \mathcal{H}_G^{B_i}$. Ces dernières formules correspondent aux $T^s_{i,j}$ contenues dans la structure $\mathcal{H}_G^{B_i}$ qui précède la structure $\mathcal{H}_H^{B_j}$ que nous sommes en train de calculer. Ensuite, nous ajoutons la liste des formules $\phi : \langle B_j\rangle\phi \in H$ à $\overline{\mathcal{L}_{B_j}}$ qui devront être calculées aux étapes suivantes. Puisque les B_j sont des modalités sérielles, nous ajoutons aussi les formules $\phi : B_j\phi \in H$. Pour initialiser cette structure, nous prenons alors une formule ϕ de cet ensemble $\overline{\mathcal{L}_{B_j}}$ et calculons un premier ensemble de Hintikka contenant la demande suscitée par $\langle B_j\rangle\phi$ dans H. Nous passons alors à la seconde étape de la construction de la structure.

Nous venons donc d'initialiser une structure temporaire S. La seconde étape consiste à prendre une [6] formule ϕ de $\overline{\mathcal{L}_{B_j}}$ est de l'intégrer dans la structure S : soit en l'intégrant dans un ensemble H_J déjà existant dans la structure et tel que $(H_J, \Box_J) \in S$ et H_J est compatible avec cette formule ϕ (i.e. nous pouvons calculer un ensemble de Hintikka H'_J contenant $\{\phi\}\cup H_J\cup\Box_H^{B_j}$ sur la fermeture de $\{\phi\} \cup H_J \cup \Box_H^{B_j}$) ; soit il n'est pas possible d'intégrer la formule ϕ dans un élément de la structure temporaire S calculée, alors nous calculons un nouvel élément $(H_J, \Box_H^{B_j})$, avec H_J l'ensemble de Hintikka contenant $\{\phi\} \cup \Box_H^{B_j}$, à la structure S . Sinon cela signifie que nous n'avons pas pu ajouter à un ensemble déjà existant la formule ϕ et nous n'avons pas pu construire un nouvel ensemble contenant ϕ, nous renvoyons alors Faux.

La troisième étape de la construction de la structure $\mathcal{H}_H^{B_j}$ consiste à mettre à jour tous les éléments de S avec la dernière liste de $\Box_H^{B_j}$ calculée. Nous prenons un élément $I = (H_I, \Box_I) \in S$, puis nous calculons le nouvel ensemble de ensemble de Hintikka H'_I contenant $H_I \cup \Box_I^{B_j}$. Nous répétons le processus tant qu'il y a des éléments de S à mettre à jour et, si besoin, nous recommençons les calculs de l'étape précédente s'il existe une nouvelle formule ϕ qui devrait être ajoutée à la structure et n'a pas encore été ajoutée dans la structure jusqu'à présent suite au calcul de H'_I.

Les variables booléennes suivantes représentent chacune des étapes décrites précédemment et sont utilisées dans l'algorithme 2 :

FINI := « $\exists\mathcal{H}_H$ ensemble TB-temoin généré par H » est traduit dans l'algorithme par l'appel récursif à la fonction Witness$B_j()$, ou Witness$T^s_{i,j}()$, ou encore Witness() ;

nous pouvons initialiser une structure $\mathcal{H}_H^{B_j}$ tq FINI := « $\exists\ \phi \in \overline{\mathcal{L}_{B_j}}$ tq : $\exists\ H_E$ Hintikka sur $Cl(\{\phi\} \cup \Box_H^{B_j})$ contenant $\{\phi\} \cup \Box_H^{B_j}$ et WitnessB_j(H, Σ_H, $\mathcal{H}_G^{B_i}$, $\mathcal{L}_{B_j}\cup H_E$, $\overline{\mathcal{L}_{B_j}} \setminus(\mathcal{L}_{B_j}\cup H_E)$, $\Box_H^{B_j} \cup \{\phi : B_j\phi \in H_E\} \cup \{\phi : T^s_{j,j}\phi \in H_E\}$, $\{(H_E, \Box_H^{B_j})\}$, $\{H_E\} \cup \{H\})$ » ;

nous pouvons ajouter à un élément de S la formule ϕ tq FINI := « $\exists$ J $= (H_J, \Box_J^{B_j}) \in S$ tq $\exists\ H'_J$ Hintikka sur $Cl(H_J \cup \{\phi\} \cup \Box_H^{B_j})$ contenant $H_J \cup \{\phi\} \cup \Box_H^{B_j}$ et tq WitnessB_j(H, Σ_H, $\mathcal{H}_G^{B_i}$, $\mathcal{L}_{B_j} \cup H'_J$, $\overline{\mathcal{L}_{B_j}} \cup \{\phi : \langle B_j\rangle\phi \in H'_J\} \setminus (\mathcal{L}_{B_j} \cup H'_J)$, $\Box_H^{B_j} \cup \{\phi : B_j\phi \in H'_J\} \cup$

6. Le contrôle **prendre un** est équivalent à considérer une boucle **pour tout** dans laquelle, au premier élément récupéré de la liste, nous quittons la boucle. Si la liste est vide, alors nous passons à la suite du code.

Algorithme 2 WitnessB_j(H, Σ_H, $\mathcal{H}_G^{B_i}$, $\mathcal{L}_{B_j}$, $\overline{\mathcal{L}_{B_j}}$, $\Box_H^{B_j}$, S, $\mathcal{H}_H^{B_j}$) avec $i, j \in \mathcal{N} : i \neq j$

si S $= \emptyset$ **alors :**
  $\Box_H^{B_j} \leftarrow \{\phi : B_j\phi \in H\} \cup \{\phi : T_{j,j}^s\phi \in H\} \cup \{\phi : T_{i,j}^s\phi \in I, I \in \mathcal{H}_G^{B_i}\}$
  $\overline{\mathcal{L}_{B_j}} \leftarrow \{\phi : \langle B_j\rangle\phi \in H\} \cup \Box_H^{B_j}$
  S $\leftarrow \{\emptyset\}$
  $\mathcal{H}_H^{B_j} \leftarrow \{H\}$
  si $\overline{\mathcal{L}_{B_j}} \neq \emptyset$ **alors :**
    si nous pouvons initialiser une structure $\mathcal{H}_H^{B_j}$ tq FINI **alors :**
      retourner $\top$
    sinon
      retourner $\bot$
    fin si
  sinon
    retourner End-Of-Witness$B_j(\mathcal{H}_H^{B_j})$
  fin si
sinon
  prendre un $\phi \in \overline{\mathcal{L}_{B_j}}$ **faire :**
    si nous pouvons ajouter à un élément de S la formule ϕ tq FINI **alors :**
      retourner $\top$
    sinon si nous pouvons construire un nouvel élément de S contenant ϕ tq FINI **alors :**
      retourner $\top$
    sinon
      retourner $\bot$
    fin si
  fin prendre
  prendre un I $= (H_I, \Box_I^{B_j}) \in S$ tq $\Box_I^{B_j} \neq \Box_H^{B_j}$ **faire :**
    S $\leftarrow S \setminus I$
    si nous pouvons mettre à jour $\Box_I^{B_j}$ de l'élément $I \in S$ à $\Box_H^{B_j}$ tq FINI **alors :**
      retourner $\top$
    sinon
      retourner $\bot$
    fin si
  fin prendre
  retourner End-Of-Witness$B_j(\mathcal{H}_H^{B_j})$
fin si

$\{\phi : T^s_{j,j}\phi \in H'_J\}$, $S \setminus J \cup \{(H'_J, \Box^{B_j}_H)\}$, $(\mathcal{H}^{B_j}_H \setminus \{H_J\}) \cup \{H'_J\})$ » ;

nous pouvons construire un nouvel élément de S contenant ϕ tq FINI := « $\exists\, H'_J$ Hintikka sur Cl($\{\phi\} \cup \Box^{B_j}_H$) contenant $\{\phi\} \cup \Box^{B_j}_H$ et tq WitnessB_j(H, Σ_H, $\mathcal{H}^{B_i}_G$, $\mathcal{L}_{B_j} \cup H'_J$, $(\overline{\mathcal{L}_{B_j}} \cup \{\phi : \langle B_j\rangle\phi \in H'_J\})\backslash(\mathcal{L}_{B_j} \cup H'_J)$, $\Box^{B_j}_H \cup \{\phi : B_j\phi \in H'_J\} \cup \{\phi : T^s_{j,j}\phi \in H'_J\}$, S $\cup$ $\{(H'_J, \Box^{B_j}_H)\}$, $(\mathcal{H}^{B_j}_H \cup \{H'_J\})$ » ;

nous pouvons mettre à jour $\Box^{B_j}_I$ de l'élément $I \in S$ à $\Box^{B_j}_H$ tq FINI := « $\exists\ H'_I$ Hintikka sur $Cl(H_I \cup \Box^{B_j}_H)$ contenant $H_I \cup \Box^{B_j}_H$ et tq WitnessB_j(H, Σ_H, $\mathcal{H}^{B_i}_G$, $\mathcal{L}_{B_j} \cup H'_I$, $(\overline{\mathcal{L}_{B_j}} \cup \{\phi : \langle B_j\rangle\phi \in H'_I\})\backslash(\mathcal{L}_{B_j} \cup H'_I)$, $\Box^{B_j}_H \cup \{\phi : B_j\phi \in H'_I\} \cup \{\phi : T^s_{j,j}\phi \in H'_I\}$, S $\cup$ $\{(H'_I, \Box^{B_j}_H)\}$, $(\mathcal{H}^{B_j}_H \setminus \{H'_I\}) \cup \{H'_I\})$ » ;

Enfin, s'il n'existe plus d'éléments de S à mettre à jour, la structure $\mathcal{H}^{B_j}_H$ a terminé d'être calculée. Nous passons alors à la fin de la méthode donnée par l'algorithme 3, qui consiste pour tous les $I \in \mathcal{H}^{B_j}_H \setminus \{H\}$, d'une part à calculer les structures $\mathcal{H}^{T^s_{j,k}}_I$ qui suivent la structure $\mathcal{H}^{B_j}_H$, et d'autre part à calculer les autres structures $\mathcal{H}^{B_k}_I$ avec $k \neq j$, qui sont suscitées dans les ensembles de $\mathcal{H}^{B_j}_H \setminus \{H\}$. Par ailleurs, nous remarquons qu'en raison des contraintes sémantiques $\forall w, u, v \in \mathcal{W} : w\mathcal{B}_i u \wedge u\mathcal{T}_{i,j}v \Rightarrow w\mathcal{T}_{i,j}v$ et $w\mathcal{B}_i u \wedge w\mathcal{T}_{i,j}v \Rightarrow u\mathcal{T}_{i,j}v$, il revient de manière équivalente et plus efficacement de construire une seule structure $\mathcal{H}^{T^s_{j,k}}_H$ contenant les demandes suscitées par les $\langle T^s_{i,j}\rangle\phi$ contenus dans la structure $\mathcal{H}^{B_j}_H$.

Algorithme 3 End-Of-Witness$B_j(\mathcal{H}^{B_j}_H)$ avec $j \in \mathcal{N}$

pour tout $k \in \mathcal{N}$ **faire :**
  si $WitnessT^s_{j,k}(\{\phi : \langle T^s_{j,k}\rangle\phi \in I, I \in \mathcal{H}^{B_j}_H\}, \{\phi : T^s_{j,k}\phi \in I, I \in \mathcal{H}^{B_j}_H\})$ est faux **alors :**
    retourner $\bot$
  fin si
fin pour
si $\mathcal{H}^{B_j}_H \setminus \{H\} \neq \emptyset$ **alors :**
  pour tout $H_I \in \mathcal{H}^{B_j}_H \setminus \{H\}$ **faire :**
    si $\exists k \in \mathcal{N}, k \neq j$ tq Witness$B_k(H_I, Cl(H_I), \mathcal{H}^{B_j}_H, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset)$ est faux **alors :**
      retourner $\bot$
    fin si
  fin pour
fin si
retourner $\top$

Pour construire les structures $\mathcal{H}^{T^s_{i,j}}_H$, nous définissons la fonction Witness$T^s_{i,j}$() présentée dans l'algorithme 4. Cette fonction possède deux paramètres :

$\Diamond^{T^s_{i,j}}$: la liste des formules ϕ telles que $\langle T^s_{i,j}\rangle\phi$ est contenue dans la structure $\mathcal{H}^{B_i}_H$ qui précède la structure $\mathcal{H}^{T^s_{i,j}}_H$;

$\Box^{T^s_{i,j}}$: la liste des formules ϕ telles que $T^s_{i,j}\phi$ est contenue dans les éléments de la structure $\mathcal{H}^{B_i}_H$ qui précède la structure $\mathcal{H}^{T^s_{i,j}}_H$.

Cette fonction calcule pour chaque formule ϕ contenue dans $\Diamond^{T^s_{i,j}}$ un ensemble de Hintikka H sur $Cl(\{\phi\} \cup \Box^{T^s_{i,j}})$ tel que H contient $\{\phi\} \cup \Box^{T^s_{i,j}}$ et tel que l'algorithme termine (i.e. pour chaque H calculé, nous pouvons calculer pour tout $k \in \mathcal{N}$ une nouvelle structure $\mathcal{H}_H^{B_k}$ par l'appel à la fonction WitnessB_k()).

Algorithme 4 Witness$T^s_{i,j}$($\Diamond^{T^s_{i,j}}$, $\Box^{T^s_{i,j}}$) avec $i, j \in \mathcal{N}$

si $\Diamond^{T^s_{i,j}} \neq \emptyset$ **alors :**
  pour tout $\phi \in \Diamond^{T^s_{i,j}}$ **faire :**
    si il n'existe pas d'ensemble de Hintikka H sur $\Sigma_H = Cl(\{\phi\} \cup \Box^{T^s_{i,j}})$ contenant $\{\phi\} \cup \Box^{T^s_{i,j}}$ tel que : $\forall i \in \mathcal{N}$, Witness$B_i$(H, Σ_H, $\emptyset$, $\emptyset$, $\emptyset$, $\emptyset$, $\emptyset$, $\emptyset$) **alors :**
      retourner $\bot$
    fin si
  fin pour
sinon
  si $\Box^{T^s_{i,j}} \neq \emptyset$ **alors :**
    prendre un $\phi \in \Box^{T^s_{i,j}}$ **faire :**
      si il n'existe pas d'ensemble de Hintikka H sur $\Sigma_H = Cl(\{\phi\} \cup \Box^{T^s_{i,j}})$ contenant $\{\phi\} \cup \Box^{T^s_{i,j}}$ tel que : $\forall i \in \mathcal{N}$, Witness$B_i$(H, Σ_H, $\emptyset$, $\emptyset$, $\emptyset$, $\emptyset$, $\emptyset$, $\emptyset$) **alors :**
        retourner $\bot$
      fin si
    fin prendre
  fin si
fin si
retourner $\top$

Il est facile de voir que cet algorithme termine. En effet, pour chaque structure $\mathcal{H}_H^{\Diamond}$ calculée à partir d'un ensemble de Hintikka H, nous avons $deg(\mathcal{H}_H^{\Diamond} \setminus \{H\}) < deg(H)$. Ainsi, l'algorithme termine lorsqu'il n'y a plus aucun $\langle B_j \rangle$, B_j, $\langle T^s_{j,k} \rangle$ ou $T^s_{j,k}$ à calculer. De plus, l'union de toutes les structures $\mathcal{H}_H^{\Diamond}$ générées forment un ensemble TB-témoin. Ainsi, par application du théorème de correction/complétude de la méthode, pour un ensemble de formules Γ, si Γ est satisfiable, alors la fonction $Witness(\Gamma)$ renvoie Vraie et réciproquement. Par conséquent, cet algorithme est aussi correct. Enfin, nous pourrions prouver que la classe de complexité du problème TB-SAT est dans PSPACE. En effet, nous pouvons nous rendre compte que comme le nombre d'appels récursifs est polynomial par rapport au degré de l'ensemble H, nous pourrions définir une machine de Turing dont l'espace mémoire occupé dans le pire cas est polynomial.

5.3.2 Bibliothèques et logiciels pour logiques modales

Les méthodes de calcul présentées en section 5.1 ont été appliquées pour développer des bibliothèques et logiciels permettant de résoudre les problèmes de satisfiabilité et de validité dans les logiques modales. Dans cette section, nous présentons quelques unes de ces librairies et logiciels, puis nous proposons à la section suivante une implémentation en JAVA de la méthode des tableaux dédiée au cadre TB.

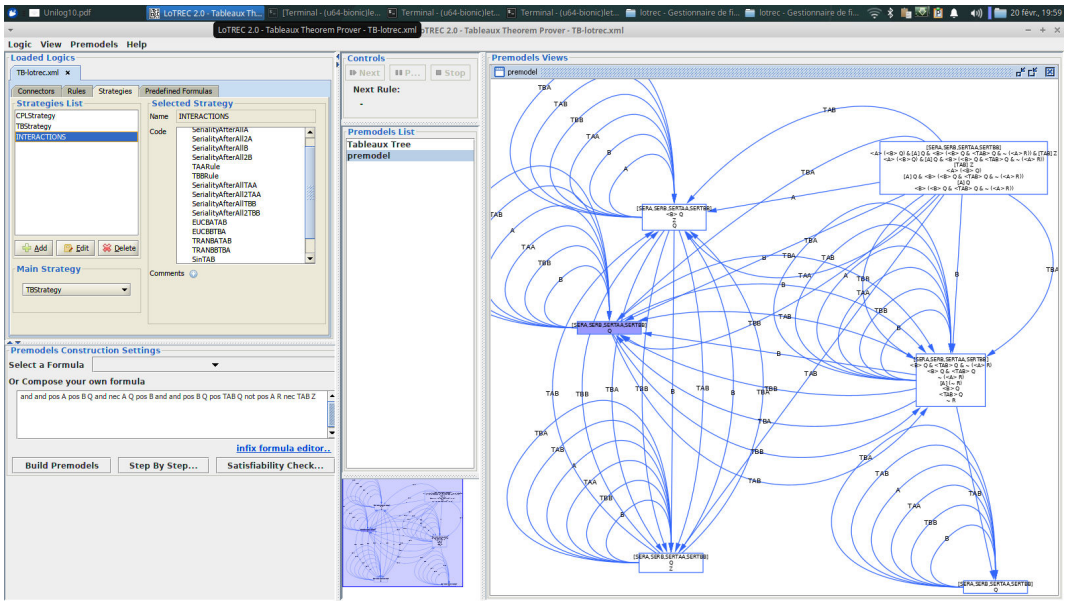

FIGURE 5.4 – Exemple d'application de la méthode des tableaux avec LoTreC

Il existe de nombreux solveurs SAT pour les logiques modales et ils s'organisent en deux types de solveurs : les solveurs génériques comme LoTreC (del Cerro et al., 2001) qui implémentent une méthode des tableaux abstraite destinée à résoudre le problème SAT dans n'importe quel cadre de Kripke multi-modal ; et les solveurs spécialisés et optimisés qui résolvent le problème SAT dans une logique spécifique. Même si l'objectif de ce chapitre est d'implémenter une méthode pour résoudre le problème SAT dans TB, parmi les solveurs spécialisés et optimisés pour des logiques modales, nous pouvons mentionner FaCT (Horrocks, 1997), RACER (Haarslev and Möller, 2001), DLP (Patel-Schneider, 1998), LWB (Heuerding et al., 1996) ou encore MLTP (Li, 2008) qui sont spécialisés pour la logique K ou encore KT. Par exemple, MSPASS (Hustadt et al., 1999) est un solveur spécialisé dans les logiques multi-modales mais ne gère pas les systèmes avec des dépendances logiques entre les modalités. Pour une revue détaillée des méthodes utilisées dans ces solveurs SAT optimisés, le lecteur intéressé peut se référer à la thèse de (Li, 2008).

LoTreC est un solveur générique permettant de vérifier la satisfiabilité d'une formule dans un cadre quelconque. L'utilisateur précise les opérateurs, les contraintes et une stratégie à suivre pour construire un modèle du cadre. Cette stratégie représente l'ordre dans lequel les contraintes du cadre doivent être appliquées. La figure 5.4 montre un exemple de résolution du problème de satisfiabilité dans le cadre TB appliqué à la formule : $\phi = \langle B_A\rangle\langle B_B\rangle Q \wedge B_A Q \wedge \langle B_B\rangle(\langle B_B\rangle Q \wedge T^s_{A,B} Q \wedge Q \wedge \neg(\langle B_A\rangle R)) \wedge T^s_{A,B} Z$ avec A, B deux agents et Q, R, Z trois variables propositionnelles. Pour construire un modèle, LoTreC calcule tous les modèles possibles en appliquant les contraintes du cadre spécifiées par l'utilisateur. La méthode des tableaux appliquée par LoTreC étant générale à tout cadre, elle n'exploite pas les propriétés spécifiques du cadre pour aboutir plus rapidement à un résultat. L'inconvénient d'appliquer une telle méthode générique est donc son manque d'efficacité comme le souligne (Li, 2008).

5.3.3 Implémentation du solveur SAT en JAVA pour TB

Dans la suite de ce chapitre, nous présentons une implémentation en JAVA de la méthode des tableaux présentée en section 5.2 pour résoudre TB-SAT. Cette instanciation de la méthode a été réalisée à partir de la bibliothèque TweetyProject, développée par Matthias Thimm (Thimm, 2014). Dans un premier temps, nous montrons des exemples de résolution de la satisfiabilité de formules avec notre méthode, puis dans un second temps nous comparons notre méthode avec une méthode des tableaux naïve, i.e. qui va construire tous les modèles possibles et vérifier si l'un d'entre eux décide de la satisfiabilité d'une formule.

Exemples de résolution

Reprenons la formule de l'exemple 5.9, $\phi = B_i\langle B_j\rangle q \wedge T^s_{i,j} q \wedge p$ et appliquons l'algorithme pour construire un modèle satisfaisant cette formule. Dans l'exemple 5.9, nous avons montré que $\mathcal{G}^{TB}_H = \{\{H\}, \mathcal{H}^{B_i}_H, \mathcal{H}^{B_j}_H, \mathcal{H}^{T^s_{i,j}}_H, \mathcal{H}^{B_j}_{I_0}\}$ était un générateur d'ensemble TB-témoin où $\mathcal{H}^{B_i}_H = \{H, I_0 = \{\langle B_j\rangle q, q\}\}$, $\mathcal{H}^{B_j}_H = \{H, J_0 = \{q\}\}$, $\mathcal{H}^{T^s_{i,j}}_H = \{H, K_0 = \{q\}\}$ et $\mathcal{H}^{B_j}_{I_0} = \{I_0, J_{01} = \{q\}\}$. En suivant le déroulement de l'algorithme présenté en section 5.3 et en appliquant les contraintes de TB à chaque étape de calcul des $\mathcal{H}^{\Diamond}_{H'}$, nous construisons un modèle satisfaisant la formule $\phi = B_i\langle B_j\rangle q \wedge T^s_{i,j} q \wedge p$. La figure 5.5 représente un modèle obtenu lors de l'application de la méthode. Chaque nœud du graphe représente un monde et le label associé (1,2,...) représente la profondeur dans l'arbre du générateur $\mathcal{G}^{TB}_H$. Dans ce modèle, nous avons donc un ensemble de mondes possibles $\mathcal{W} = \{w0 = \{B_i\langle B_j\rangle q \wedge T^s_{i,j} q \wedge p, \langle B_j\rangle q, q, p\}, w1 = \{q, \neg p\}, w2 = \{\langle B_j\rangle q, q, \neg p\}, w3 = \{q, \neg p\}, w4 = \{q, \neg p\}\} \cup \{w_\top\}$. Nous rappelons que l'ensemble TB-témoin calculé était $\mathcal{H}_H = \{\{B_i\langle B_j\rangle q \wedge T^s_{i,j} q \wedge p, \langle B_j\rangle q, q, p\}, \{\langle B_j\rangle q, q, \neg p\}, \{q, \neg p\}\}$, ce qui correspond bien à la valuation donnée par chaque monde. Lorsque la méthode a terminé, il existe des mondes et des arcs qui n'ont pas de successeurs. Ainsi, pour appliquer la sérialité pour tout couple $(u \in \mathcal{W}, \mathcal{R}_\Diamond)$ tel que u n'a pas de successeur pour $\mathcal{R}_\Diamond$. Il suffit alors de construire un nouvel arc dans $\mathcal{R}_\Diamond$, vers un monde quelconque puisque les contraintes Γ associées à la modalité $\Diamond$ sont alors vides. Or, l'ensemble vide est inclus dans tout ensemble. Cependant, dans cet exemple, pour des raisons de lisibilité, nous faisons converger tous les arcs vers un monde « puits ». Ce monde est désigné par la lettre T.

Par ailleurs, nous présentons en annexe B d'autres résultats de modèles calculés par la méthode des ensembles TB-témoins.

Comparaison avec une méthode des tableaux naïve

Nous proposons dans cette section de comparer notre méthode des tableaux, désignée par l'appellation **TB-M**[7], à une méthode des tableaux dite *naïve*, désignée par **NA-M**[8]. La méthode naïve construit un modèle en appliquant la méthode des tableaux présentée en section 5.1.2 puis une fois l'arbre construit, applique les contraintes du cadre sur ce modèle et vérifie si la formule

7. TB-Méthode
8. NAïve-Méthode

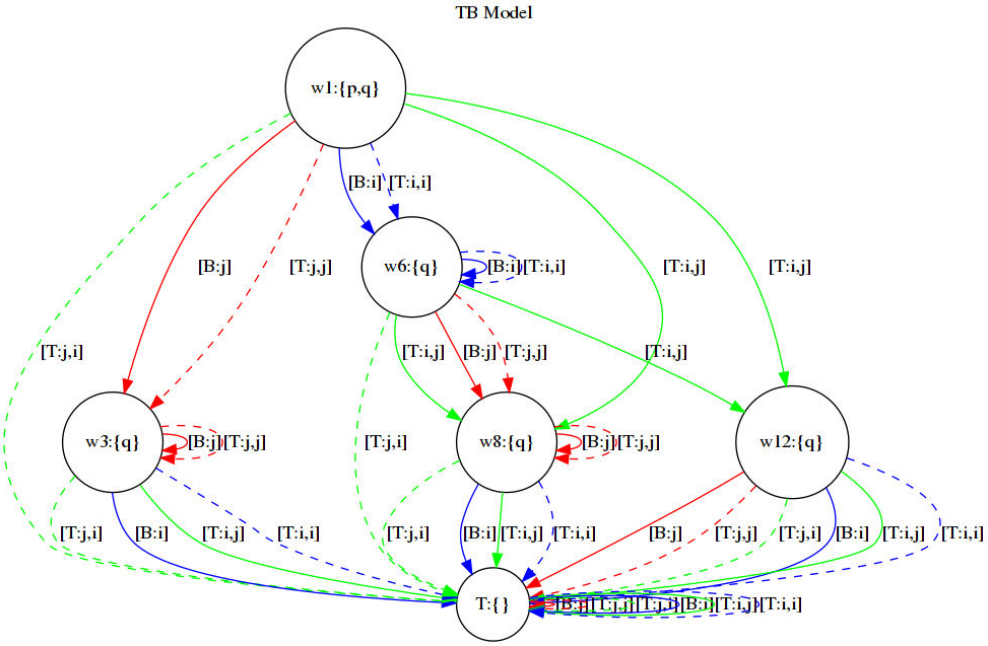

FIGURE 5.5 – Application de la méthode pour résoudre TB-SAT

est satisfaite ou non. La méthode recommence jusqu'à ce qu'un modèle satisfasse la formule ou jusqu'à ce que tous les modèles possibles aient été testés.

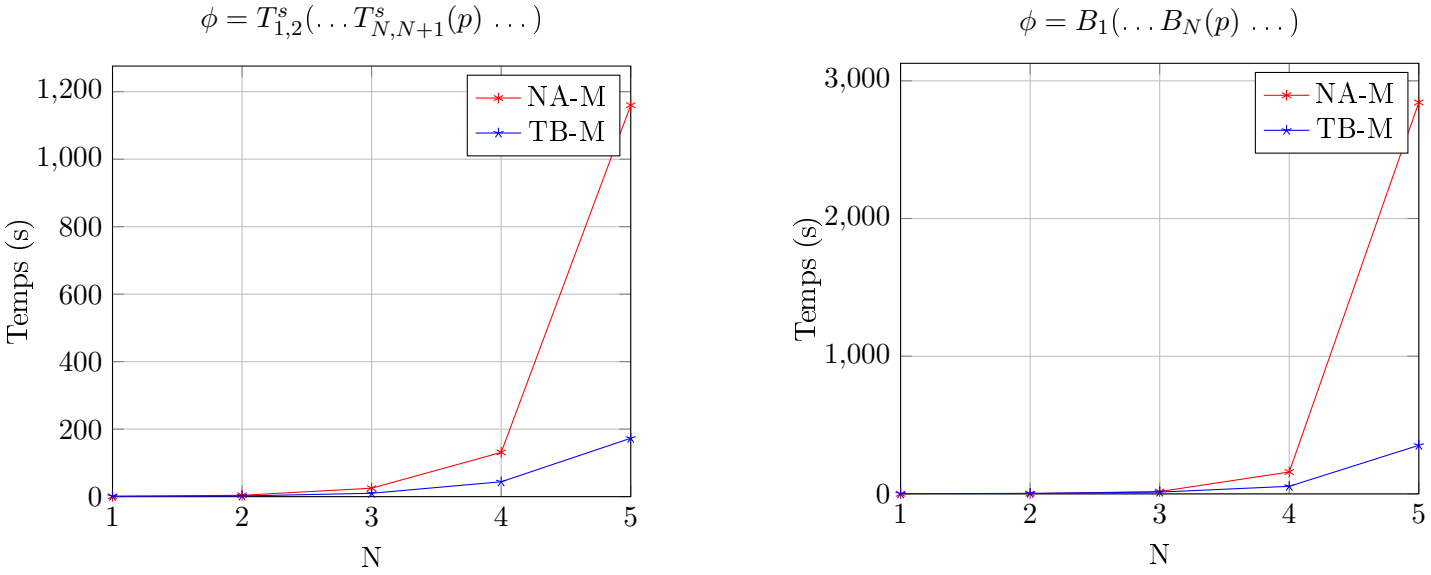

FIGURE 5.6 – Évolution du temps de calcul en fonction du degré de ϕ.

La suite de cette section présente l'évolution du temps de calcul d'une formule ϕ en fonction du

degré de celle-ci. Les caractéristiques de la machine sur laquelle l'expérimentation a été effectuée sont données ci-dessous :

Processeur : Intel Core i5-2520M CPU @ 2.50GHz x 4 ;

Mémoire vive : 3,8 Gio ;

Carte graphique : Intel Sandybridge Mobile ;

Système d'exploitation : UBUNTU 18.04.2 LTS ;

Type système d'exploitation : 64-bit.

Dans la figure 5.6 de gauche, nous présentons l'évolution du temps de calcul en fonction du degré N d'une formule de la forme $T^s_{1,2}(\dots T^s_{N,N+1}(p)\dots)$. Ainsi, la formule testée pour $N = 1$ est $T^s_{1,2}p$, la formule testée pour $N = 2$ est $T^s_{1,2}T^s_{2,3}p$, etc. Enfin, la formule testée pour $N = n$ avec $n \in \mathbb{N}^*$, est notée $T^s_{1,2}(\dots T^s_{n,n+1}(p)\dots)$. La première formule testée pour $N = 1$ est $T^s_{1,2}p$. Le temps de calcul pour ces deux méthodes est quasiment similaire. En revanche, lorsque N augmente, la différence est marquante. Pour $N = 4$, la méthode **NA-M** met 131 secondes alors que la méthode **TB-M** met 44 secondes. Enfin, pour $N = 5$, la méthode **NA-M** met 1160 secondes tandis que notre méthode des tableaux 173 secondes.

Dans la figure 5.6 de droite, nous montrons l'évolution du temps de calcul d'un modèle satisfaisant une formule ϕ de la forme $B_1(\dots B_N(p)\dots)$. La formule testée pour $N = 1$ est B_1p, la formule testée pour $N = 2$ est B_1B_2p, ..., la formule testée pour $N = n$ avec $n \in \mathbb{N}^*$ est $B_1(\dots B_n(p)\dots)$. Ainsi, lorsque $N \leq 3$, il n'existe que très peu de différences entre les deux méthodes. En revanche, à partir de $N = 5$ la méthode **NA-M** met 2843 secondes tandis que la méthode des tableaux **TB-M** met 353 secondes.

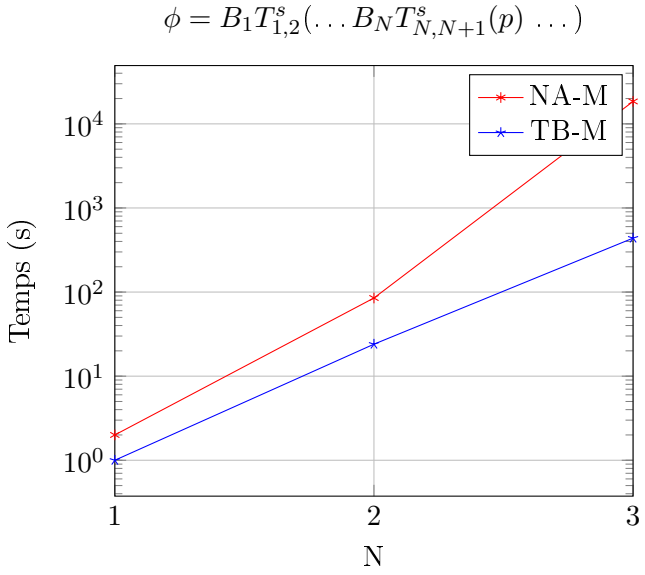

FIGURE 5.7 – Évolution du temps de calcul en fonction de N, en échelle logarithmique.

Enfin, dans la figure 5.7, nous illustrons l'évolution du temps de calcul des deux méthodes en fonction du degré d'une formule ϕ de la forme $B_1T^s_{1,2}(\dots B_NT^s_{N,N+1}(p)\dots)$. La formule testée

pour $N = 1$ est $B_1 T^s_{1,2} p$, la formule testée pour $N = 2$ est $B_1 T^s_{1,2}(B_2 T^s_{2,3} p)$, ..., la formule testée pour $N = n$ avec $n \in \mathbb{N}^*$ est $B_1 T^s_{1,2}(\ldots(B_n T^s_{n,n+1}(p))\ldots)$. Contrairement à l'évolution du temps de calcul des formules de la figure 5.6, nous travaillons ici par paire de modalités $(B_i, T^s_{i,j})$. Nous ne pouvons donc étudier que trois degrés de formule pour $N = 1$ $deg(\phi) = 2$ pour $N = 2$ $deg(\phi) = 4$ et pour $N = 3$, $deg(\phi) = 6$. De plus, la figure 5.7 présente les résultats du temps de calcul en échelle logarithmique. Ainsi, nous constatons que le temps de calcul de la méthode **NA-M** explose. En effet, pour $N = 6$, la méthode **NA-M** met 18474 secondes tandis que la méthode **TB-M** met 438 secondes.

Nous pouvons donc conclure que la méthode des tableaux présentée dans ce chapitre est plus efficace en temps de calcul qu'une méthode des tableaux naïve. Cette différence s'explique notamment du fait que dans la méthode présentée dans ce chapitre, nous vérifions étape par étape si chaque ensemble $\mathcal{H}^{\Diamond}_{H'}$ forme bien un ensemble témoin alors que dans la méthode naïve, nous construisons à chaque fois un modèle complet.

En conclusion, dans ce chapitre nous avons dans un premier temps présenté deux méthodes de preuve pour raisonner dans les logiques modales : *la méthode des tableaux* et *la méthode des arbres labellisée*. La première méthode permet de vérifier l'existence d'un modèle satisfaisant une certaine formule ϕ tandis que la seconde permet de vérifier si une certaine formule ϕ est valide. Dans un second temps, nous nous sommes intéressés à instancier une méthode des tableaux dédiée à résoudre le problème de satisfiabilité dans TB. Nous nous sommes alors appuyés sur la notion d'ensembles de Hintikka et nous avons défini une notion d'ensemble témoins s'appuyant directement des propriétés du cadre TB pour construire un modèle satisfaisant ces contraintes. Enfin, nous avons proposé un algorithme fondé sur cette méthode ainsi qu'un instanciation de celui-ci, puis nous avons étudié le temps de calcul en fonction de l'évolution du degré d'une formule ϕ. Nous avons alors pu constater que la méthode des tableaux présentée dans le cadre TB était plus efficace en temps de calcul qu'une méthode des tableaux naïve.

Chapitre 6

Conclusion et perspectives

Dans ce travail de thèse, nous nous sommes intéressés à la question de la représentation formelle de la manipulation dans les systèmes multi-agents (SMA). Ce chapitre a pour objectif de conclure ce travail mais aussi d'apporter des perspectives à celui-ci. Ainsi, en section 6.1 nous rappelons le positionnement et la problématique de cette thèse, puis en section 6.2 nous résumons l'ensemble des contributions réalisées au cours de cette thèse, enfin en section 6.3 nous présentons des perspectives de recherches futures.

6.1 Positionnement et problématique de la thèse

Pour contrecarrer les agents manipulateurs, des travaux se sont intéressés à la construction de systèmes robustes. Une première approche consiste à étudier l'axiomatisation du système pour rendre inopérable certaines stratégies de manipulation. Une seconde approche consiste à montrer que la complexité est trop importante pour rendre opérable une stratégie de manipulation. Une autre approche consiste à intégrer des mécanismes pour inciter les agents à bien se comporter. Cependant, un SMA parfaitement robuste à la manipulation n'existe pas. Certaines techniques consistent alors à détecter les stratégies de manipulation que ce soit par des méthodes statistiques, ou bien par des techniques de raisonnement automatique. Cependant, ces approches s'intéressent uniquement à la tromperie, et non à la manipulation.

Si ces deux concepts sont proches, conceptuellement parlant, ils s'en différencient. La manipulation est une intention délibérée d'influencer tout en veillant à la dissimuler, alors que la tromperie est une intention délibérée de dissimuler. Ainsi, nous avons défini la manipulation comme l'intention délibérée d'instrumentaliser un autre agent tout en veillant à lui dissimuler cette intention. Une telle définition ne peut exister que dans des SMA où les agents ont des états mentaux. Ainsi, puisque les intentions et les connaissances sont des états mentaux, nous nous sommes alors restreints aux SMA avec des agents cognitifs. La problématique de cette thèse était alors de trouver un formalisme permettant de représenter une telle définition de la manipulation et permettant de caractériser certaines stratégies de manipulation fondées sur la connaissance, les croyances des agents, mais aussi sur la confiance. Caractériser formellement la notion de manipulation ainsi que les stratégies de manipulation peut alors nous aider à mieux concevoir

des systèmes informatiques robustes à la manipulation ou, dans le cas où il est impossible de construire de tels systèmes, de les détecter.

6.2 Contributions

C'est pourquoi nous nous sommes intéressés à construire un outil logique pour donner une description formelle à la manipulation. Nous avons alors proposé dans un premier temps un système logique, nommé KBE, permettant de raisonner sur la manipulation, mais aussi d'exprimer des stratégies de manipulation fondées sur les croyances et les connaissances des agents. Cependant, la logique KBE ne permet pas d'exprimer des stratégies de manipulation fondées sur la confiance en la sincérité, ni même de raisonner sur cette notion. Ainsi, dans un second temps nous avons proposé la logique TB permettant d'exprimer cette notion de confiance en la sincérité. Enfin, dans un dernier temps nous avons proposé d'implémenter une nouvelle méthode algorithmique fondée sur les méthodes des tableaux pour raisonner dans le système TB et résoudre le problème de TB-satisfiabilité.

Les contributions de ce travail ont donc été de :

1. construire un système logique, pour raisonner sur la manipulation (KBE) ;
2. construire un système logique, pour raisonner sur la confiance en la sincérité (TB) ;
3. proposer une nouvelle méthode algorithmique pour raisonner dans les systèmes logiques multi-modaux. Cette méthode est appliquée au cadre logique de TB.

6.2.1 Une logique de la manipulation

Pour exprimer la manipulation comme une intention délibérée d'instrumentaliser une victime tout en s'assurant à lui dissimuler cette intention, nous avons proposé un cadre logique, nommé KBE. Ce cadre introduit explicitement une nouvelle modalité pour exprimer l'intention délibérée. L'intention délibérée est définie sémantiquement comme l'ensemble des stratégies ayant été adoptées et mises en œuvre par un agent. En prouvant que ce système était correct et complet, nous avons ensuite pu utiliser ce système pour déduire des théorèmes. Par exemple, un agent qui influence de manière délibérée ne peut avoir l'intention de montrer que d'autres agents, y compris lui-même, puissent détenir la vérité. Un autre exemple de théorème déductible dans KBE est le principe *qui facit per alium facit per se*, i.e. « celui qui agit à travers un autre fait acte lui-même ». Enfin, en définissant dans KBE la *coercition*, la *persuasion*, et la *tromperie*, nous avons pu constater que la manipulation était bien différente de ces trois notions, ce qui est cohérent avec les résultats en sciences sociales. Ainsi, prouver qu'un agent exerce une influence, ou qu'il trompe d'autres agents ne suffit pas pour montrer une manipulation. Il est tout d'abord nécessaire de prouver qu'un agent a eu l'intention délibérée d'influencer, puis de montrer que cet agent a eu aussi l'intention délibérée de dissimuler sa stratégie d'influencer. Par ailleurs, la dissimulation est souvent liée à la stratégie adoptée par l'agent pour manipuler. En effet, lorsqu'un agent applique une stratégie de manipulation comme la révélation d'un faux profil de préférence dans un système de vote, une partie de la stratégie de manipulation consiste évidemment à ne

pas montrer aux autres agents que son profil de préférence révélé n'est pas son véritable profil de préférence.

6.2.2 Une logique de la confiance en la sincérité

Lorsque des agents manipulent d'autres agents, il est nécessaire que les agents sélectionnent les informations à croire ou ne pas croire. Une information communiquée par un agent peut être crue par un autre dès lors que ce dernier a confiance en la fiabilité de l'agent communiquant, mais aussi lorsque l'agent a confiance en la sincérité de l'autre. Ainsi, nous avons proposé un second cadre logique, nommé TB, permettant d'exprimer la notion de confiance en la sincérité. Nous introduisons alors dans ce système une modalité $T^s_{i,j}$ signifiant qu'un agent i a confiance en la sincérité de j sur une proposition. Cette modalité permet alors de décrire des relations logiques existants lorsqu'un agent décide d'accorder sa confiance. En particulier, lorsqu'un agent i accorde sa confiance en la sincérité d'un agent j, cet agent i va nécessairement croire que l'agent j croit ce qu'il dit. Cette relation logique est appelée *axiome de sincérité*. Cependant, comme un agent a toujours le choix de ne pas accorder sa confiance et s'il croit que l'autre agent est en train de le manipuler, la réciproque de cet axiome ne peut pas être considérée. Nous avons ensuite démontré que ce système était correct, complet, fortement correct et fortement complet. Nous avons étudié certaines propriétés logiques de la confiance en la sincérité. Nous avons ensuite étendu cette notion à celle de confiance collective, nommée *confiance en la sincérité partagée*. Enfin, nous avons montré que la confiance partagée était un système KD.

6.2.3 Une méthode des tableaux pour des logiques modales

Le système KBE et le système TB sont deux logiques multi-modales non normales et normales. Si nous les avons présenté sur un plan purement théorique, dans le chapitre 5 nous avons proposé une nouvelle méthode de calcul permettant de décider de la satisfiabilité d'une formule dans TB. Cette méthode est fondée sur la notion d'ensembles de Hintikka et plus particulièrement, sur une nouvelle notion introduite dans ce chapitre, appelée *générateur d'ensembles TB-témoins à partir d'un ensemble de Hintikka*. Les ensembles TB-témoins, qui sont définis comme l'existence d'un générateur d'ensemble TB-témoins, permettent de définir un relation d'équivalence entre l'existence de ces ensembles témoins et la satisfaisabilité d'un ensemble de formules dans TB. Ce théorème d'équivalence, prouvé, nous a alors permis de construire un algorithme décidant de la satisfiabilité d'un ensemble de formules. Nous avons ensuite constaté que cette méthode de calcul fondée sur la construction de ces ensembles TB-témoins était plus efficace qu'une méthode des tableaux naïve consistant à construire tous les modèles possibles pour obtenir un modèle satisfaisant un ensemble de formules.

6.3 Perspectives

Notre travail nous amène à nous poser de nouvelles questions. Tout d'abord, certaines stratégies de manipulation peuvent se fonder sur d'autres états mentaux que les intentions et les connaissances. Il serait donc intéressant d'intégrer au système KBE de nouvelles modalités pour considérer des états mentaux comme les normes, les désirs, les émotions ou encore la confiance. Ensuite, une autre perspective consiste à intégrer les travaux sur la prise de conscience et d'ajouter un aspect dynamique au système KBE. Enfin, une dernière perspective porte sur la généralisation et l'amélioration de la méthode algorithmique pour résoudre le problème de satisfaisabilité dans les logiques multi-modales normales et non normales. Il serait alors intéressant de comparer cette méthode à d'autres méthodes génériques et de benchmarker la méthode.

6.3.1 Intégrer les normes, les désirs et la confiance au système KBE

Dans le chapitre 1, nous avons présenté différentes bases de stratégies de manipulation comme les stratégies fondées sur la rationalité d'un agent. Par exemple, influencer sur les choix d'un agent est une stratégie fondée sur la rationalité. Un autre type de stratégie de manipulation peut se fonder sur les croyances et les connaissances d'un agent comme avec la tromperie ou le mensonge. Si nous pouvons exprimer de telles stratégies de manipulation dans le système logique KBE, nous ne pouvons pas exprimer les autres formes de stratégies de manipulation lorsqu'elles reposent sur les désirs d'un agent, les normes d'un système, ou encore lorsqu'elles reposent sur la confiance en la sincérité qu'un agent accorde à un autre agent. En effet, le système logique TB a été défini de façon indépendante et n'a pas été fusionné au système KBE. Une première perspective consiste alors à fusionner ces deux systèmes ensemble. Ce nouveau système se nommerait TKBE. Nous proposons une ébauche d'un tel système en annexe C. Par exemple, un tel système permettrait de déduire que si un agent j fait confiance en la sincérité d'un agent i sur le fait que i ne l'instrumentalise pas, alors il ne peut être le cas que l'agent j croit que i puisse l'instrumentaliser, c'est-à-dire le théorème :

$$\vdash T^s_{j,i} \neg E^d_i E_j \phi \Rightarrow \neg B_j E^d_i E_j \phi$$

Nous aurions alors de nouvelles formes de manipulation dont la dissimulation serait fondée sur la confiance entretenue entre les deux agents. De la même manière nous pourrions ajouter au système KBE la notion de désirs, de normes ou encore d'obligations déontiques comme nous avons pu les présenter en section 2.1.3. Une autre perspective consiste alors à reprendre ces travaux et les intégrer dans KBE. Par exemple, nous pourrions avoir un ensemble de modalités d'obligations déontiques $\{\Box_\eta\}_{\eta \in \mathcal{O}}$ par rapport à un ensemble de normes $\mathcal{O}$ ainsi qu'un ensemble de modalités de désirs $\{D_i\}_{i \in \mathcal{N}}$ représentant les désirs de chaque agent du système. Supposons par exemple la *norme de réciprocité* que nous notons η, décrite comme le fait que si un agent i offre un service à un autre agent, alors l'autre agent j doit lui rendre la pareille. Pour décrire cette norme, nous avons besoin de considérer un ensemble supplémentaire de modalités comme une modalité temporelle $Y\phi$ pour traduire que dans le passé il s'est produit ϕ, $X\phi$ pour traduire que dans le futur il va se produire ϕ, une modalité représentant l'acquisition d'information $I_{j,i}\phi$ signifiant qu'un agent j a acquis par i l'information ϕ. Ainsi, la norme de réciprocité η serait

traduite par la formule, avec E_i^d notre modalité d'intention délibérée et E_i la modalité capturant tous les effets des actions effectuées par l'agent i :

$$rec_{i,j}(\phi,\psi) \stackrel{\triangle}{=} Y(D_j\phi \wedge E_i^d\phi) \Rightarrow \Box_\eta(I_{j,i}D_i\psi \Rightarrow XE_j\psi)$$

Ce prédicat traduit littéralement que si dans le passé, l'agent i a délibérément satisfait un désir de l'agent j, alors il est nécessaire au regard de la norme de réciprocité η que lorsque l'agent j reçoit l'information que i désire ψ, nécessairement dans le futur, j doit veiller à ce que ψ soit vérifiée. Par exemple, lorsqu'un agent entre dans une chocolaterie où des vendeurs viennent à sa rencontre pour lui faire goûter des échantillons de chocolat gratuitement, si l'agent accepte de goûter aux chocolats, alors il accepte aussi un service offert gratuitement par les vendeurs. Ainsi, si l'agent applique la norme de réciprocité, alors il doit rendre la pareille aux vendeurs en leur achetant un produit. Cette stratégie de manipulation consiste alors à activer une norme qu'un agent suit, ici la norme de réciprocité, pour manipuler cet agent.

6.3.2 Étendre KBE aux logiques de la prise de conscience

Lorsque nous avons défini la manipulation au sein du système KBE comme l'intention délibérée d'instrumentaliser un agent tout en veillant à lui dissimuler cette intention, nous avons défini la dissimulation comme la non connaissance ou la non croyance des intentions du manipulateur. Cependant, dans de nombreuses situations, un agent manipulé n'a pas conscience d'avoir été manipulé. Or, nous avons présenté en section 2.3.3 des travaux portant sur la représentation de la prise de conscience. Une perspective est d'intégrer la prise de conscience dans la manipulation. Si A_i est la modalité de conscience associée à un agent i telle que présentée dans (Schipper, 2014), Σ un ensemble fermé et fini de formules tel que $\top \in \Sigma$, $\phi \in \Sigma$ et $\mathcal{A}_i : \mathcal{W} \to 2^\Sigma$ la fonction de conscience associée à la modalité A_i, nous pourrions alors définir de nouvelles formes de manipulations avec absence de conscience. Ainsi, la *manipulation constructive douce avec absence de conscience* serait définie par le prédicat suivant :

$$MCEA_{i,j}^\Sigma\phi = \bigvee_{\psi\in\Sigma} E_i^d(E_j\phi \wedge \neg A_j E_i^d E_j\phi \wedge \psi)$$

La *manipulation constructive forte avec absence de conscience* consisterait en :

$$MCE^dA_{i,j}^\Sigma\phi = \bigvee_{\psi\in\Sigma} E_i^d(E_j^d\phi \wedge \neg A_j E_i^d E_j^d\phi \wedge \psi)$$

La *manipulation destructive douce avec absence de conscience* serait représentée par :

$$MDEA_{i,j}^\Sigma\phi = \bigvee_{\psi\in\Sigma} E_i^d(\neg E_j\phi \wedge \neg A_j E_i^d \neg E_j\phi \wedge \psi)$$

Enfin, la *manipulation destructive forte avec absence de conscience* serait définie par le prédicat :

$$MDE^d A^{\Sigma}_{i,j}\phi = \bigvee_{\psi\in\Sigma} E^d_i(\neg E^d_j\phi \wedge \neg A_j E^d_i \neg E^d_j\phi \wedge \psi)$$

De plus, remarquons que nous avons défini la fonction $\mathcal{A}_i$ telle que pour tout $w \in \mathcal{W}$, $\mathcal{A}_i(w) \subseteq \Sigma$. Cela signifie que dans un tel système, nous supposons que les agents n'ont jamais conscience de formules qui ne sont pas dans Σ.

Enfin, une autre perspective consiste à intégrer les modèles des logiques dynamiques pour considérer des mécanismes de révision des connaissances et des croyances des agents lorsque des agents sont en train de manipuler. Pour ce faire, nous pourrions reprendre un modèle doxastique dynamique tel que nous l'avons présenté en section 2.1.3 et définir un modèle événementiel de la manipulation comme celui de (Van Ditmarsch et al., 2007) où ils définissent un modèle événementiel pour exprimer les croyances des agents après l'annonce d'un mensonge.

6.3.3 Améliorer et généraliser la méthode des tableaux aux autres logiques

Nous avons proposé une nouvelle méthode des tableaux pour résoudre le problème de satisfiabilité dans les logiques modales. L'algorithme proposé ainsi que ses implémentations peuvent être améliorées. Une première amélioration de l'algorithme en espace consisterait à intégrer une structure d'arbre de décision binaire ordonnée et réduite (ROBDD) pour représenter les ensembles de Hintikka. Un ROBDD permet de représenter de façon compacte les formules du calcul propositionnel. Par exemple, la formule $\phi = (p \wedge q) \vee (r \wedge s)$ est représentée par le graphe acyclique de la figure 6.1.

La valeur de vérité de la formule est donnée par le chemin de la racine p à une feuille $\bot$, ou $\top$ et les valuations des variables correspondantes sont données par les arcs. Un arc en pointillés d'origine p signifie que dans ce chemin (interprétation), la variable est considérée à faux tandis qu'un arc plein signifie que la variable p est à vrai. Ainsi, dans cet exemple, nous pouvons remarquer que le chemin $((p,q),(q,r),(r,s),(s,\top))$ qui mène à rendre vrai la formule ϕ correspond à l'interprétation I des variables telle que $I(p) = \top$, $I(q) = \bot$, $I(r) = \top$ et $I(s) = \top$.

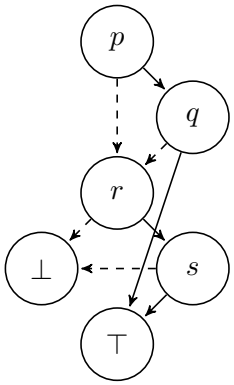

FIGURE 6.1 – ROBDD de la formule $(p \wedge q) \vee (r \wedge s)$

L'algorithme peut aussi être amélioré en temps en intégrant certaines méthodes d'optimisation comme celles présentées dans la thèse de (Li, 2008). Parmi les techniques proposées que nous pourrions adapter au cadre TB, nous retrouvons la simplification, la propagation unitaire, la normalisation lexicale, le backjumping dynamique, la mise en cache, l'utilisation d'heuristiques ou encore en simulant l'algorithme de Davis-Putnam-Logemann-Loveland (DPLL) (Davis and Putnam, 1960) où les occurrences des formules positives et négatives dans deux branches alternatives sont mémorisées afin d'explorer les modèles qu'une seule fois.

Enfin, si la méthode des tableaux proposée dans cette thèse est uniquement dédiée au cadre TB, elle pourrait être généralisée à tout cadre multi-modal normal et non normal. Nous pourrions donc adapter cette méthode au cadre de KBE. De plus, il semble pertinent d'expérimenter cette méthode généralisée par rapport aux autres méthodes génériques comme celles proposées sur les arbres labellisées de (Baldoni, 2000) ou encore en implémentant une version plus récente de la méthode des tableaux utilisée dans LoTreC et proposée par (Schwarzentruber, 2011).

En conclusion, cette thèse aura permis de proposer de nouveaux modèles théoriques permettant d'exprimer formellement les notions de manipulation et de confiance en la sincérité dans les systèmes multi-agents. Si une méthode algorithmique a été proposée pour raisonner sur la confiance en la sincérité, cette méthode reste théorique et ne permet pas encore une implémentation sur des systèmes réels. Par implémentation sur des systèmes réels, nous entendons la déduction automatique des intentions d'un agent, de la confiance d'un agent ou encore des connaissances et des croyances que peuvent avoir les agents envers les autres. Par conséquent, une étape future pourrait être de se questionner sur la manière dont nous pourrions intégrer ces méthodes de calcul sur des systèmes réels afin de détecter la manipulation mais aussi les stratégies de manipulation quand elles ont lieu ?

Annexe A

Preuves avec les systèmes à la Hilbert

Ce chapitre en annexe propose des preuves avec des systèmes à la Hilbert pour les propriétés (1) et (2) du théorème 3.7. Nous rappelons le théorème à démontrer.

Théorème 1.1 - :

1. *Un agent qui influence de manière délibérée ne peut avoir l'intention de montrer que d'autres agents, y compris lui-même, puissent détenir la vérité, c'est-à-dire :*

$$\vdash E_i^d B_j \phi \Rightarrow \neg E_i^d B_j K_k \neg \phi$$

2. *Un agent qui veille de manière délibérée qu'un agent ne croit pas une information, ne peut pas aussi veiller que ce dernier sache que des agents détiennent la vérité, c'est-à-dire :*

$$\vdash E_i^d \neg B_j \phi \Rightarrow \neg E_i^d B_j K_k \phi$$

La preuve suivante correspondant à la propriété (1) du théorème 3.7 :

Démonstration.

1. $\vdash B_j\phi \wedge B_j K_k \neg\phi \Rightarrow B_j\phi$
2. $\vdash B_j(K_k\neg\phi \Rightarrow \neg\phi)$
3. $\vdash B_j(K_k\neg\phi \Rightarrow \neg\phi) \Rightarrow (B_j K_k\neg\phi \Rightarrow B_j\neg\phi)$
4. $\vdash B_j K_k\neg\phi \Rightarrow B_j\neg\phi$
5. $\vdash B_j\neg\phi \Rightarrow \neg B_j\phi$
6. $\vdash B_j K_k\neg\phi \Rightarrow B_j\neg\phi \Rightarrow \neg B_j\phi$
7. $\vdash (B_j K_k\neg\phi \Rightarrow B_j\neg\phi \Rightarrow \neg B_j\phi) \Rightarrow ((B_j K_k\neg\phi \Rightarrow B_j\neg\phi) \Rightarrow (B_j K_k\neg\phi \Rightarrow \neg B_j\phi))$
8. $\vdash B_j K_k\neg\phi \Rightarrow \neg B_j\phi$
9. $\vdash B_j\phi \wedge B_j K_k\neg\phi \Rightarrow \neg B_j\phi$

10. $\vdash ((B_j\phi \wedge B_jK_k\neg\phi \Rightarrow B_j\phi)) \Rightarrow ((B_j\phi \wedge B_jK_k\neg\phi \Rightarrow \neg B_j\phi) \Rightarrow \neg(B_j\phi \wedge B_jK_k\neg\phi))$
11. $\vdash \neg(B_j\phi \wedge B_jK_k\neg\phi)$
12. $\vdash \neg(B_j\phi \wedge B_jK_k\neg\phi) \equiv (B_j\phi \Rightarrow \neg B_jK_k\neg\phi)$
13. $\vdash B_j\phi \Rightarrow \neg B_jK_k\neg\phi$
14. $\vdash E_i(B_j\phi \Rightarrow \neg B_jK_k\neg\phi)$
15. $\vdash E_i(B_j\phi \Rightarrow \neg B_jK_k\neg\phi) \Rightarrow (E_iB_j\phi \Rightarrow E_i\neg B_jK_k\neg\phi)$
16. $\vdash E_iB_j\phi \Rightarrow E_i\neg B_jK_k\neg\phi$
17. $\vdash E_i^dB_j\phi \Rightarrow E_iB_j\phi$
18. $\vdash (E_i^dB_j\phi \Rightarrow E_iB_j\phi \Rightarrow E_i\neg B_jK_k\neg\phi) \Rightarrow ((E_i^dB_j\phi \Rightarrow E_iB_j\phi) \Rightarrow (E_i^dB_j\phi \Rightarrow E_i\neg B_jK_k\neg\phi))$
19. $\vdash E_i^dB_j\phi \Rightarrow E_i\neg B_jK_k\neg\phi$
20. $\vdash E_i\neg B_jK_k\neg\phi \Rightarrow \neg E_iB_jK_k\neg\phi$
21. $\vdash \neg E_iB_jK_k\neg\phi \Rightarrow \neg E_i^dB_jK_k\neg\phi$
22. $\vdash E_i^dB_j\phi \Rightarrow \neg E_iB_jK_k\neg\phi \Rightarrow \neg E_i^dB_jK_k\neg\phi$
23. $\vdash (E_i^dB_j\phi \Rightarrow \neg E_iB_jK_k\neg\phi \Rightarrow \neg E_i^dB_jK_k\neg\phi) \Rightarrow (E_i^dB_j\phi \Rightarrow \neg E_iB_jK_k\neg\phi) \Rightarrow E_i^dB_j\phi \Rightarrow \neg E_i^dB_jK_k\neg\phi$
24. $\vdash E_i^dB_j\phi \Rightarrow \neg E_i^dB_jK_k\neg\phi$

□

La preuve suivante correspondant à la propriété (2) du théorème 3.7 :

Démonstration.

1. $\vdash K_k\phi \Rightarrow \phi$
2. $\vdash B_j(K_k\phi \Rightarrow \phi)$
3. $\vdash B_j(K_k\phi \Rightarrow \phi) \Rightarrow B_jK_k\phi \Rightarrow B_j\phi$
4. $\vdash B_jK_k\phi \Rightarrow B_j\phi$
5. $\vdash (B_jK_k\phi \Rightarrow B_j\phi) \Rightarrow \neg B_j\phi \Rightarrow \neg B_jK_k\phi$
6. $\vdash \neg B_j\phi \Rightarrow \neg B_jK_k\phi$
7. $\vdash E_i(\neg B_j\phi \Rightarrow \neg B_jK_k\phi)$
8. $\vdash E_i(\neg B_j\phi \Rightarrow \neg B_jK_k\phi) \Rightarrow E_i\neg B_j\phi \Rightarrow E_i\neg B_jK_k\phi$
9. $\vdash E_i\neg B_j\phi \Rightarrow E_i\neg B_jK_k\phi$
10. $\vdash E_i^d\neg B_j\phi \Rightarrow E_i\neg B_j\phi$
11. $\vdash E_i^d\neg B_j\phi \Rightarrow E_i\neg B_j\phi \Rightarrow E_i\neg B_jK_k\phi$
12. $\vdash (E_i^d\neg B_j\phi \Rightarrow E_i\neg B_j\phi \Rightarrow E_i\neg B_jK_k\phi) \Rightarrow (E_i^d\neg B_j\phi \Rightarrow E_i\neg B_j\phi) \Rightarrow E_i^d\neg B_j\phi \Rightarrow E_i\neg B_jK_k\phi$
13. $\vdash E_i^d\neg B_j\phi \Rightarrow E_i\neg B_jK_k\phi$
14. $\vdash E_i\neg B_jK_k\phi \Rightarrow \neg E_iB_jK_k\phi$

15. $\vdash \neg E_i B_j K_k \phi \Rightarrow \neg E_i^d B_j K_k \phi$
16. $\vdash E_i^d \neg B_j \phi \Rightarrow E_i \neg B_j K_k \phi \Rightarrow \neg E_i^d B_j K_k \phi$
17. $\vdash (E_i^d \neg B_j \phi \Rightarrow E_i \neg B_j K_k \phi \Rightarrow \neg E_i^d B_j K_k \phi) \Rightarrow (E_i^d \neg B_j \phi \Rightarrow E_i \neg B_j K_k \phi) \Rightarrow E_i^d \neg B_j \phi \Rightarrow \neg E_i^d B_j K_k \phi$
18. $\vdash E_i^d \neg B_j \phi \Rightarrow \neg E_i^d B_j K_k \phi$

□

La preuve à la Hilbert correspondant au théorème 3.8 :

Démonstration.

1. $\vdash (E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow E_i \phi$
2. $\vdash E_i^d E_j \phi \Rightarrow E_i E_j \phi$
3. $\vdash E_j \phi \Rightarrow \phi$
4. $\vdash E_i(E_j \phi \Rightarrow \phi)$
5. $\vdash E_i(E_j \phi \Rightarrow \phi) \Rightarrow E_i E_j \phi \Rightarrow E_i \phi$
6. $E_i E_j \phi \Rightarrow E_i \phi$
7. $E_i^d E_j \phi \Rightarrow E_i E_j \phi$
8. $E_i^d E_j \phi \Rightarrow E_i E_j \phi \Rightarrow E_i \phi$
9. $(E_i^d E_j \phi \Rightarrow E_i E_j \phi \Rightarrow E_i \phi) \Rightarrow (E_i^d E_j \phi \Rightarrow E_i E_j \phi) \Rightarrow E_i^d E_j \phi \Rightarrow E_i \phi$
10. $E_i^d E_j \phi \Rightarrow E_i \phi$
11. $E_i^d E_j^d \phi \Rightarrow E_i E_j^d \phi$
12. $E_i E_j^d \phi \Rightarrow E_i E_j \phi$
13. $E_i^d E_j^d \phi \Rightarrow E_i E_j^d \phi \Rightarrow E_i E_j \phi$
14. $(E_i^d E_j^d \phi \Rightarrow E_i E_j^d \phi \Rightarrow E_i E_j \phi) \Rightarrow (E_i^d E_j^d \phi \Rightarrow E_i E_j^d \phi) \Rightarrow E_i^d E_j^d \phi \Rightarrow E_i E_j \phi$
15. $E_i^d E_j^d \phi \Rightarrow E_i E_j \phi$
16. $E_i^d E_j^d \phi \Rightarrow E_i E_j \phi \Rightarrow E_i \phi$
17. $(E_i^d E_j^d \phi \Rightarrow E_i E_j \phi \Rightarrow E_i \phi) \Rightarrow (E_i^d E_j^d \phi \Rightarrow E_i E_j \phi) \Rightarrow E_i^d E_j^d \phi \Rightarrow E_i \phi$
18. $E_i^d E_j^d \phi \Rightarrow E_i \phi$
19. $(E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow ((E_i^d E_j \phi \Rightarrow E_i \phi) \Rightarrow (E_i^d E_j^d \phi \Rightarrow E_i \phi) \Rightarrow E_i \phi))$
20. $((E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow ((E_i^d E_j \phi \Rightarrow E_i \phi) \Rightarrow (E_i^d E_j^d \phi \Rightarrow E_i \phi) \Rightarrow E_i \phi)) \Rightarrow ((E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow (E_i^d E_j \phi \Rightarrow E_i \phi)) \Rightarrow (E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow (E_i^d E_j^d \phi \Rightarrow E_i \phi) \Rightarrow E_i \phi$
21. $(E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow (E_i^d E_j^d \phi \Rightarrow E_i \phi) \Rightarrow E_i \phi$
22. $((E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow (E_i^d E_j^d \phi \Rightarrow E_i \phi) \Rightarrow E_i \phi) \Rightarrow ((E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow (E_i^d E_j^d \phi \Rightarrow E_i \phi)) \Rightarrow (E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow E_i \phi$
23. $(E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow E_i \phi$

□

Annexe B

Exemples de modèles calculés avec la méthode TB-M

Dans ce chapitre d'annexe, nous présentons quelques exemples de modèles obtenus lorsque nous appliquons la méthode des tableaux proposée en chapitre 5 et fondée sur les ensembles TB-témoins.

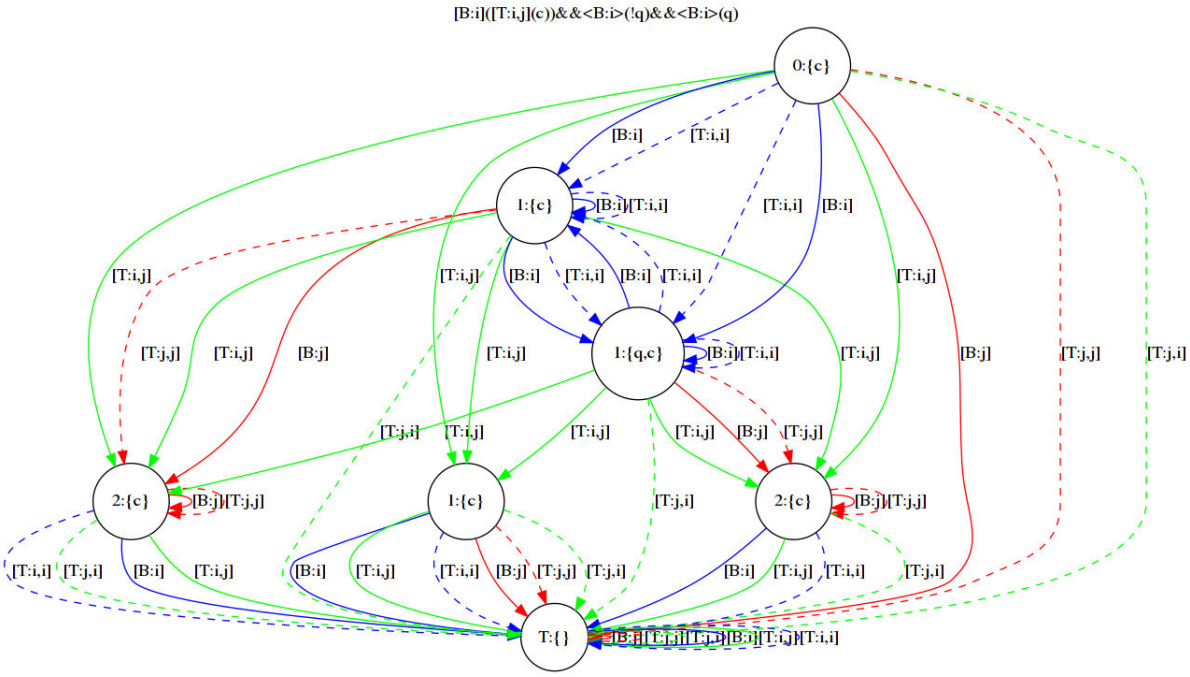

FIGURE B.1 – Modèle satisfaisant la formule $B_i T^s_{i,j} c \wedge \langle B_i \rangle \neg q \wedge \langle B_i \rangle q$.

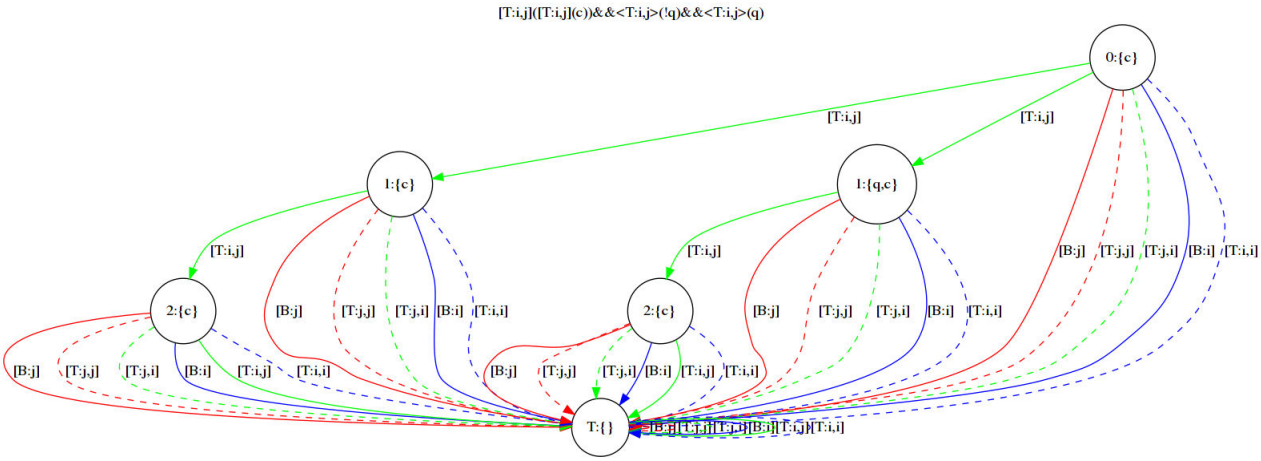

FIGURE B.2 – Modèle satisfaisant la formule $T^s_{i,j}T^s_{i,j}c \wedge \langle T^s_{i,j}\rangle\neg q \wedge \langle T^s_{i,j}\rangle q$.

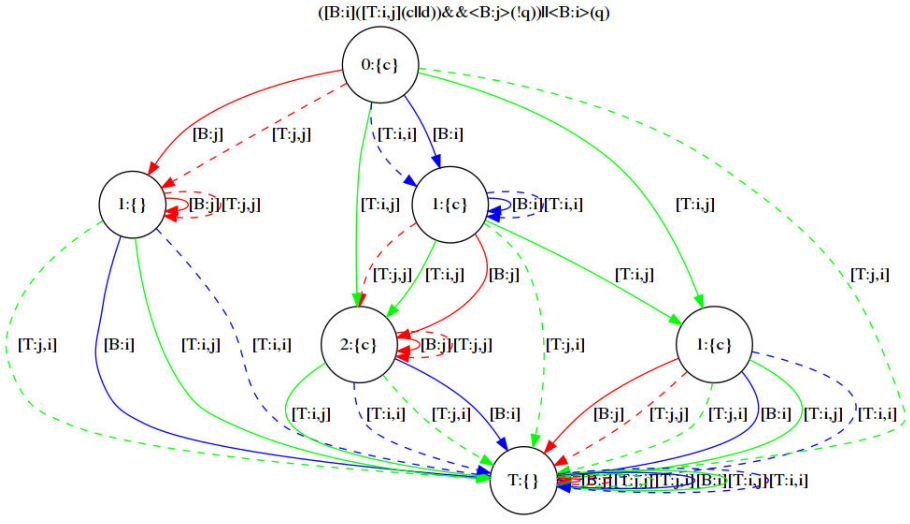

FIGURE B.3 – Modèle satisfaisant la formule $B_iT^s_{i,j}(c||d) \wedge \langle B_j\rangle\neg q \vee \langle B_i\rangle(q)$.

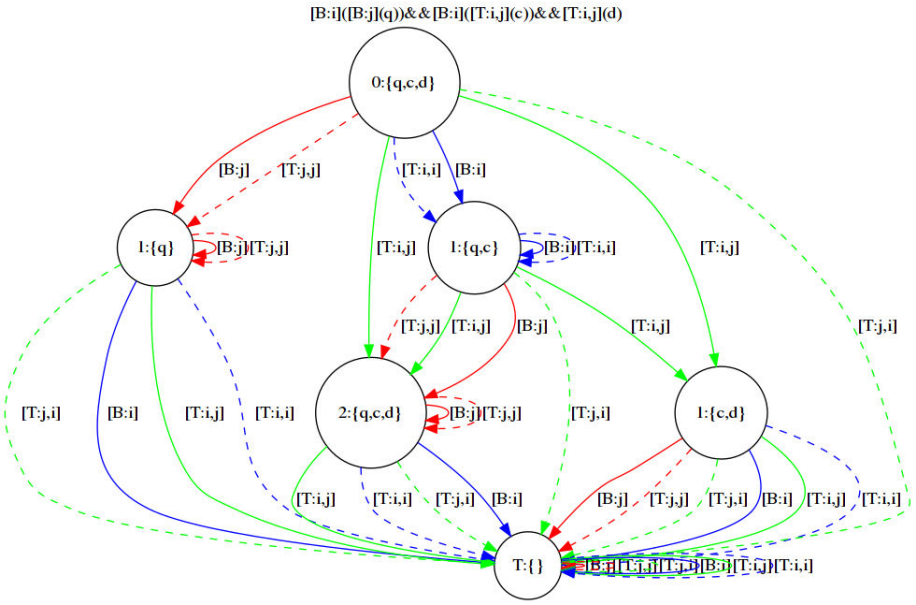

FIGURE B.4 – Modèle satisfaisant la formule $B_i B_j q \wedge B_i T^s_{i,j} c \wedge T^s_{i,j} d$.

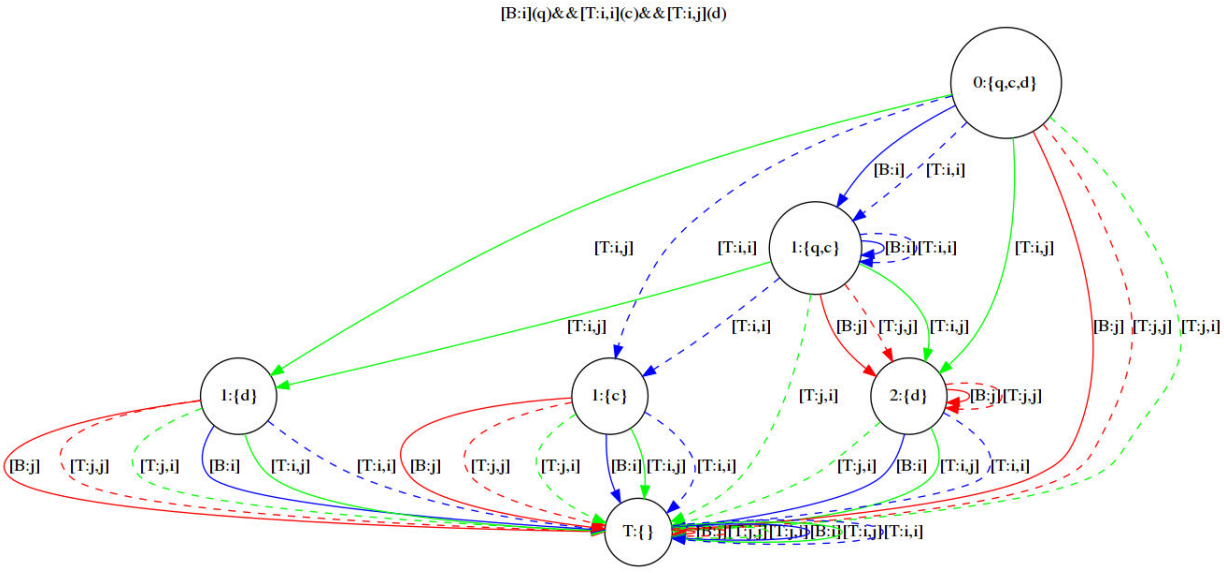

FIGURE B.5 – Modèle satisfaisant la formule $B_i q \wedge T^s_{i,i} c \wedge T^s_{i,j} d$.

Annexe C

Système TKBE

Ce chapitre en annexe propose de fusionner les systèmes KBE et TB ensemble. Le système fondé sur la fusion des deux cadres est nommé *système TKBE*.

C.1 Langage et sémantique du cadre TKBE

Soient un ensemble de lettres propositionnelles $\mathcal{P} = \{a, b, c, ...\}$, un ensemble d'agents $\mathcal{N}$ avec $i, j \in \mathcal{N}$ deux agents, et $p \in \mathcal{P}$ une variable propositionnelle. Le langage $\mathcal{L}_{KBE}$ est généré par la grammaire sous forme de Backus-Naur suivante :

$$\phi ::= p \mid \neg\phi \mid \phi \Rightarrow \phi \mid T^s_{i,j}\phi \mid K_i\phi \mid B_i\phi \mid E_i\phi \mid E^d_i\phi$$

Pour interpréter les formules du langage $\mathcal{L}_{TKBE}$, nous considérons le cadre $\mathcal{C} = (\mathcal{W}, \{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}^d_i\}_{i\in\mathcal{N}})$ tel que :

1. $\mathcal{W}$ un ensemble de mondes possibles non vide ;
2. $\{\mathcal{T}_{i,j}\}_{i,j\in\mathcal{N}}$ un ensemble de relations binaires telles que :

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{T}_{i,j}(w) := \{v \in \mathcal{W} | w\mathcal{T}_{i,j}v\}$$

3. $\{\mathcal{B}_i\}_{i\in\mathcal{N}}$ un ensemble de relations binaires telles que :

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{B}_i(w) := \{v \in \mathcal{W} | w\mathcal{B}_iv\}$$

4. $\{\mathcal{K}_i\}_{i\in\mathcal{N}}$ un ensemble de relations binaires telles que :

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{K}_i(w) := \{v \in \mathcal{W} | w\mathcal{K}_iv\}$$

5. $\{\mathcal{E}_i\}_{i\in\mathcal{N}}$ un ensemble de relations binaires telles que :

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{E}_i(w) := \{v \in \mathcal{W} | w\mathcal{E}_iv\}$$

6. $\{\mathcal{E}_i^d\}_{i\in\mathcal{N}}$ un ensemble de fonctions de voisinage :

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{E}_i^d(w) \in 2^{2^{\mathcal{W}}}$$

Nous reprenons les contraintes du cadre KBE.

C.1.1 Contraintes sur les connaissances et croyances

Pour tout agent $i \in \mathcal{N}$, $\mathcal{K}_i$ est une relation d'équivalence (transitive, réflexive et symétrique) et $\mathcal{B}_i$ est une relation sérielle, transitive et euclidienne. De plus ces deux relations possèdent les contraintes suivantes :

$$\forall w \in \mathcal{W} : \mathcal{B}_i(w) \subseteq \mathcal{K}_i(w) \quad (KB1)$$

$$\forall w, u, v \in \mathcal{W} : w\mathcal{K}_i u \wedge u\mathcal{B}_i v \Rightarrow w\mathcal{B}_i v \quad (KB2)$$

$$\forall w, u, v \in \mathcal{W} : w\mathcal{K}_i u \wedge w\mathcal{B}_i v \Rightarrow u\mathcal{B}_i v \quad (KB3)$$

C.1.2 Contraintes sur les effets des actions

Pour tout agent $i \in \mathcal{N}$, $\mathcal{E}_i$ est une relation d'équivalence.

C.1.3 Contraintes sur les intentions délibérées

Soit $i \in \mathcal{N}$ un agent. L'intention délibérée respecte les contraintes suivantes :

$$\forall w \in \mathcal{W} : W \notin \mathcal{E}_i^d(w) \quad (\overline{E_{Nec}^d})$$

$$\forall w \in \mathcal{W} : S \in \mathcal{E}_i^d(w) \Rightarrow \mathcal{E}_i(w) \subseteq S \quad (E^d E)$$

$$\forall w \in \mathcal{W} : S \in \mathcal{E}_i^d(w) \wedge T \in \mathcal{E}_i^d(w) \Longrightarrow S \cap T \in \mathcal{E}_i^d(w) \quad (E_{\Rightarrow,\wedge})$$

$$\forall w \in \mathcal{W} : \mathcal{E}_i^d(w) = \bigcap_{v\in\mathcal{W}:w\mathcal{K}_i v} \mathcal{E}_i^d(v) \quad (E_{KP}^d + E_{KN}^d)$$

C.1.4 Sémantique de la confiance dans TKBE

La sémantique de la confiance est légèrement modifiée dans TKBE. En effet, puisque nous considérons une modalité de connaissance, il est préférable de considérer l'introspection de la confiance par rapport aux connaissances plutôt qu'aux croyances. Soient $i, j \in \mathcal{N}$ deux agents, le cadre TKBE doit satisfaire les contraintes suivantes :

1. $\forall w \in \mathcal{W}, \exists v \in \mathcal{W} : w\mathcal{T}_{i,j}v$
2. $\forall w, u, v \in \mathcal{W} : w\mathcal{K}_i u \wedge u\mathcal{T}_{i,j}v \Rightarrow w\mathcal{T}_{i,j}v$
3. $\forall w, u, v \in \mathcal{W} : w\mathcal{K}_i u \wedge w\mathcal{T}_{i,j}v \Rightarrow u\mathcal{T}_{i,j}v$
4. $\forall w, u, v \in \mathcal{W} : w\mathcal{B}_i u \wedge u\mathcal{B}_j v \Rightarrow w\mathcal{T}_{i,j}v$

La propriété de la sincérité reste inchangée, seule l'introspection est adaptée au cadre.

C.2 Système axiomatique TKBE

Nous reprenons les notations habituelles $\vdash \phi$ signifie que ϕ est un théorème de TKBE. De plus, les notions de conséquence sémantique et de $TKBE$-déductibilité sont elles aussi habituelles.

(CP)	Tous les théorèmes du CP
(RE)	$\forall \Box_i \in \{T^s_{i,j}, B_i, K_i, E_i, E^d_i\}$ Si $\vdash \phi \Leftrightarrow \psi$ alors $\vdash \Box_i\phi \Leftrightarrow \Box_i\psi$
$(DUAL)$	$\forall(\Box_i, \Diamond_i) \in \{(T^s_{i,j}, \langle T^s_{i,j}\rangle), (B_i, \langle B_i\rangle), (K_i, \langle K_i\rangle), (E_i, \langle E_i\rangle), (E^d_i, \langle E^d_i\rangle)\} \vdash \Box_i\phi \Leftrightarrow \neg\Diamond_i\neg\phi$
$(S5_{K_i})$	Tous les théorèmes de S5 sont vérifiés pour K_i
$(S5_{E_i})$	Tous les théorèmes de S5 sont vérifiés pour E_i
$(KD45_{B_i})$	Tous les théorèmes de KD45 sont vérifiés pour B_i
$(E^d_i E_i)$	$\vdash E^d_i\phi \Rightarrow E_i\phi$
$(C_{E^d_i})$	$\vdash E^d_i\phi \wedge E^d_i\psi \Rightarrow E^d_i(\phi \wedge \psi)$
$(\neg N_{E^d_i})$	$\vdash \neg E^d_i\top$
$(4_{K_iE_i})$	$\vdash E^d_i\phi \Rightarrow K_iE^d_i\phi$
$(5_{K_iE_i})$	$\vdash \neg E^d_i\phi \Rightarrow K_i\neg E^d_i\phi$
(KB)	$\vdash K_i\phi \Rightarrow B_i\phi$
(4_{KB})	$\vdash B_i\phi \Rightarrow K_iB_i\phi$
(5_{KB})	$\vdash \neg B_i\phi \Rightarrow K_i\neg B_i\phi$
$(KD_{T^s_{i,j}})$	Tous les théorèmes de KD sont vérifiés pour $T^s_{i,j}$
(4_{TK})	$\vdash T^s_{i,j}\phi \Rightarrow K_iT^s_{i,j}\phi$
(5_{TK})	$\vdash \neg T^s_{i,j}\phi \Rightarrow K_i\neg T^s_{i,j}\phi$
(S)	$\vdash T^s_{i,j}\phi \Rightarrow B_iB_j\phi$

FIGURE C.1 – Système axiomatique du cadre TKBE

Les théorèmes de déduction sont vérifiés dans TKBE.

Théorème 3.1 - *Théorèmes de déduction* : *Soient* Σ *un ensemble de formules prouvées dans TKBE,* ϕ *et* ψ *deux formules prouvées.*

$$\Sigma \cup \{\psi\} \vdash \phi \textit{ si, et seulement si, } \Sigma \vdash \psi \Rightarrow \phi$$

$$\Sigma \cup \{\psi\} \models \phi \textit{ si, et seulement si, } \Sigma \models \psi \Rightarrow \phi$$

Puisque nous n'avons fait que combiner les contraintes de KBE et TB, le système TKBE possède les propriétés usuelles suivantes :

Théorème 3.2 - *Propriétés de TKBE* :

1. *Le système TKBE est correct ;*
2. *Le système TKBE est complet ;*
3. *Le système TKBE est fortement correct ;*
4. *Le système TKBE est fortement complet.*

C.3 Des théorèmes déduits dans TKBE

Théorème 3.3 - *Lien entre confiance et dissimulation* : *Soit* $\circ \in \{\emptyset, d\}$,

$$\vdash T^s_{j,i} \neg E^d_i E^\circ_j \phi \Rightarrow \neg B_j E^d_i E^\circ_j \phi$$

Démonstration. En appliquant l'axiome de sincérité, nous avons :

$$\vdash T^s_{j,i} \neg E^d_i E^\circ_j \phi \Rightarrow B_j B_i \neg E^d_i E^\circ_j \phi$$

Par application de l'axiome (D_{B_i}) puis par MP :

$$\vdash T^s_{j,i} \neg E^d_i E^\circ_j \phi \Rightarrow B_j \neg B_i E^d_i E^\circ_j \phi$$

Or par contraposition sur l'axiome KB, nous avons : $\vdash \neg B_i E^d_i E^\circ_j \phi \Rightarrow \neg K_i E^d_i E^\circ_j \phi$ $(*)$. De plus, par la relation d'équivalence $\vdash K_i E^d_i E^\circ_j \phi \equiv E^d_i E^\circ_j \phi$, nous avons de manière équivalente que $\vdash \neg K_i E^d_i E^\circ_j \phi \equiv \neg E^d_i E^\circ_j \phi$ est aussi un théorème. Puis, par nécessitation des B_j sur $(*)$, nous prouvons que $\vdash B_j \neg B_i E^d_i E^\circ_j \phi \Rightarrow B_j \neg E^d_i E^\circ_j \phi$ est un théorème. Enfin, par application de modus ponens, nous obtenons le théorème :

$$\vdash T^s_{j,i} \neg E^d_i E^\circ_j \phi \Rightarrow \neg B_j E^d_i E^\circ_j \phi$$

□

A ce théorème plusieurs corollaires immédiats sont alors déduits :

Corollaire 3.1 : *Soit* $\circ \in \{\emptyset, d\}$,

1. $\vdash E_i^d T_{j,i}^s \neg E_i^d E_j^\circ \phi \Rightarrow E_i \neg B_j E_i^d E_j^\circ \phi$
2. $\vdash E_i^d T_{j,i}^s \neg E_i^d E_j^\circ \phi \Rightarrow \neg E_i^d B_j E_i^d E_j^\circ \phi$
3. $\vdash E_i^d T_{j,i}^s \neg E_i^d \neg E_j^\circ \phi \Rightarrow E_i \neg B_j E_i^d \neg E_j^\circ \phi$
4. $\vdash E_i^d T_{j,i}^s \neg E_i^d \neg E_j^\circ \phi \Rightarrow \neg E_i^d B_j E_i^d \neg E_j^\circ \phi$

Ces corollaires nous permettent d'affirmer que de nouvelles formes de manipulation peuvent être considérées. Ces formes de manipulation seraient alors fondées sur la confiance entretenue entre deux agents.

Définition 3.1 - *Manipulations fondées sur la confiance* : *Soient* Σ *un ensemble fini et fermé de formules tel que* $\top \in \Sigma$ *et* $\phi \in \Sigma$ *une formule. Nous appelons* manipulation constructive douce fondée sur la confiance, *la formule* $MCET_{i,j}^\Sigma \phi$ *caractérisée par le prédicat :*

$$MCET_{i,j}^\Sigma \phi = \bigvee_{\psi \in \Sigma} E_i^d (E_j \phi \wedge T_{j,i}^s \neg E_i^d E_j \phi \wedge \psi)$$

Nous appelons manipulation constructive forte fondée sur la confiance, *la formule* $MCE^d T_{i,j}^\Sigma \phi$ *caractérisée par le prédicat :*

$$MCE^d T_{i,j}^\Sigma \phi = \bigvee_{\psi \in \Sigma} E_i^d (E_j^d \phi \wedge T_{j,i}^s \neg E_i^d E_j^d \phi \wedge \psi)$$

Nous appelons manipulation destructive douce fondée sur la confiance, *la formule* $MDET_{i,j}^\Sigma \phi$ *caractérisée par le prédicat :*

$$MDET_{i,j}^\Sigma \phi = \bigvee_{\psi \in \Sigma} E_i^d (\neg E_j \phi \wedge T_{j,i}^s \neg E_i^d \neg E_j \phi \wedge \psi)$$

Nous appelons manipulation constructive forte fondée sur la confiance, *la formule* $MDE^d T_{i,j}^\Sigma \phi$ *caractérisée par le prédicat :*

$$MDE^d T_{i,j}^\Sigma \phi = \bigvee_{\psi \in \Sigma} E_i^d (\neg E_j^d \phi \wedge T_{j,i}^s \neg E_i^d \neg E_j^d \phi \wedge \psi)$$

Bibliographie

Abell, P. (1977). The many faces of power and liberty : Revealed preference, autonomy, and teleological explanation. *Sociology*, 11(1) :3–24.

Ackerman, F. (1995). The concept of manipulativeness. *Philosophical Perspectives*, 9 :335–340.

Adam, C., Gaudou, B., Herzig, A., and Longin, D. (2006). OCC's emotions : a formalization in a BDI logic. In *International Conference on Artificial Intelligence : Methodology, Systems, and Applications*, pages 24–32. Springer.

Aggarwal, R. K. and Wu, G. (2006). Stock market manipulations. *The Journal of Business*, 79(4) :1915–1953.

Ågotnes, T., Van Der Hoek, W., Rodríguez-Aguilar, J. A., Sierra, C., and Wooldridge, M. (2007). On the logic of normative systems. In *International Conferences on Artificial Intelligence*, pages 1181–1186.

Akopova, A. S. (2013). Linguistic manipulation : Definition and types. *International Journal of Cognitive Research in Science, Engineering and Education*, 1(2) :78–82.

Altman, A. and Tennenholtz, M. (2005). Ranking systems : the pagerank axioms. In *Conference on Electronic commerce*, pages 1–8.

Altman, A. and Tennenholtz, M. (2007). Incentive compatible ranking systems. In *International Joint Conference on Autonomous Agents and Multiagent Systems*.

Altman, A. and Tennenholtz, M. (2010). An axiomatic approach to personalized ranking systems. *Journal of the ACM*, 57.

Alur, R., Henzinger, T. A., and Kupferman, O. (2002). Alternating-time temporal logic. *Journal of the ACM (JACM)*, 49(5) :672–713.

Arrow, K. J. (2012). *Social choice and individual values*, volume 12. Yale university press.

Aumann, R. J. and Dreze, J. H. (1974). Cooperative games with coalition structures. *International Journal of Game Theory*, 3(4) :217–237.

Ausubel, L. M., Milgrom, P., et al. (2006). The lovely but lonely Vickrey auction. *Combinatorial auctions*, 17 :22–26.

Balbiani, P., Herzig, A., and Troquard, N. (2008). Alternative axiomatics and complexity of deliberative stit theories. *Journal of Philosophical Logic*, 37(4) :387–406.

Baldoni, M. (2000). Normal multimodal logics with interaction axioms. In *Labelled Deduction*, pages 33–57. Springer.

Barbera, S. (2001). An introduction to strategy-proof social choice functions. *Social Choice and Welfare*, 18(4) :619–653.

Barnhill, A. (2014). What is manipulation. *Manipulation : Theory and practice*, pages 51–72.

Baron, M. (2003). Manipulativeness. In *Addresses of the American Philosophical Association*, volume 77, pages 37–54.

Bartholdi, J. J. and Orlin, J. B. (1991). Single transferable vote resists strategic voting. *Social Choice and Welfare*, 8(4) :341–354.

Bartholdi, J. J., Tovey, C. A., and Trick, M. A. (1989). The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3) :227–241.

Belnap, N. and Perloff, M. (1988). Seeing to it that : a canonical form for agentives. *Theoria*, 54(3) :175–199.

Bentzen, M. M. (2010). *Stit, Iit, and Deontic Logic for Action Types*. PhD thesis, Section for Philosophy and Science Studies, Roskilde University.

Bhaumik, R., Williams, C., Mobasher, B., and Burke, R. (2006). Securing collaborative filtering against malicious attacks through anomaly detection. In *4th Workshop on Intelligent Techniques for Web Personalization*.

Bienvenu, M., Lang, J., Wilson, N., et al. (2010). From preference logics to preference languages, and back. In *Twelfth International Conferences on the Principles of Knowledge Representation and Reasoning*.

Binmore, K., Castelfranchi, C., Doran, J., and Wooldridge, M. (1998). Rationality in multi-agent systems. *The Knowledge Engineering Review*, 13(3) :309–314.

Blackburn, P., De Rijke, M., and Venema, Y. (2002). *Modal Logic : Graph. Darst*, volume 53. Cambridge University Press.

Boella, G., Pigozzi, G., and Van Der Torre, L. (2009). Normative framework for normative system change. In *8th International Conference on Autonomous Agents and Multiagent Systems*, pages 169–176.

Bonnet, G. (2012). A protocol based on a game-theoretic dilemma to prevent malicious coalitions in reputation systems. In *European Conference on Artificial Intelligence*, pages 187–192.

Bottazzi, E. and Troquard, N. (2015). On help and interpersonal control. In *The Cognitive Foundations of Group Attitudes and Social Interaction*, pages 1–23.

Bowers, L. (2003). Manipulation : description, identification and ambiguity. *Journal of Psychiatric and Mental Health Nursing*, 10(3) :323–328.

Broersen, J. (2008). A complete stit logic for knowledge and action, and some of its applications. In *International Workshop on Declarative Agent Languages and Technologies*, pages 47–59.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, 2(1) :14–23.

Burrows, M., Abadi, M., and Needham, R. M. (1989). A logic of authentication. volume 426, pages 233–271. The Royal Society London.

Calo, R. (2013). Digital market manipulation. *Geo. Wash. L. Rev.*, 82 :995.

Cao, J., Wu, Z., Mao, B., and Zhang, Y. (2013). Shilling attack detection utilizing semi-supervised learning method for collaborative recommender system. *World Wide Web*, 16(5-6) :729–748.

Carmo, J. and Pacheco, O. (2000). Deontic and action logics for collective agency and roles. In *5th International Workshop on Deontic Logic in Computer Science*, pages 93–124.

Castelfranchi, C., Conte, R., Paolucci, M., et al. (1998). Normative reputation and the costs of compliance. *Journal of Artificial Societies and Social Simulation*, 1(3) :3.

Castelfranchi, C. and Falcone, R. (2010). *Trust theory : A socio-cognitive and computational model*. John Wiley & Sons.

Chellas, B. F. (1992). Time and modality in the logic of agency. *Studia Logica*, 51(3-4) :485–517.

Cheng, A. and Friedman, E. (2005). Sybilproof reputation mechanisms. In *ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems*, pages 128–132.

Christianson, B. and Harbison, W. (1997). Why isn't trust transitive ? In *Security protocols*, pages 171–176. Springer.

Cialdini, R. B. (2001). Harnessing the science of persuasion. *Harvard Business Review*, 79(9) :72–81.

Cialdini, R. B. (2012). *Influence et manipulation*. First.

Clair, H. R. S. (1966). Manipulation. *Comprehensive psychiatry*, 7(4) :248–258.

Cohen, P. R. and Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42(2-3) :213–261.

Cohen, S. (2017). Manipulation and deception. *Australasian Journal of Philosophy*, pages 1–15.

Conitzer, V. and Sandholm, T. (2002). Complexity of manipulating elections with few candidates. In *Association for the Advancement of Artificial Intelligence*, pages 314–319.

Conitzer, V. and Sandholm, T. (2004). Computing shapley values, manipulating value division schemes, and checking core membership in multi-issue domains. In *Association for the Advancement of Artificial Intelligence*, volume 4, pages 219–225.

Conitzer, V. and Sandholm, T. (2006). Nonexistence of voting rules that are usually hard to manipulate. In *Association for the Advancement of Artificial Intelligence*, volume 6, pages 627–634.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and economic behavior*, 46(2) :260–281.

Dastani, M., Herzig, A., Hulstijn, J., and Van Der Torre, L. (2004). Inferring trust. In *International Workshop on Computational Logic in Multi-Agent Systems*, pages 144–160. Springer.

Davis, M. and Putnam, H. (1960). A computing procedure for quantification theory. *Journal of the ACM*, 7(3) :201–215.

de Saussure, L. and Schulz, P. J. (2005). *Manipulation and ideologies in the twentieth century : Discourse, language, mind*. John Benjamins Publishing.

del Cerro, L. F., Fauthoux, D., Gasquet, O., Herzig, A., Longin, D., and Massacci, F. (2001). Lotrec : the generic tableau prover for modal and description logics. In *International Joint Conference on Automated Reasoning*, pages 453–458.

Demolombe, R. (2004). Reasoning about trust : A formal logical framework. In *International Conference on Trust Management*, pages 291–303.

Demolombe, R. (2009). Graded trust. *International Conference on Autonomous Agents and Multiagent Systems*, pages 1–12.

Demolombe, R. (2011). Transitivity and propagation of trust in information sources : An analysis in modal logic. *Computational logic in multi-agent systems*, pages 13–28.

Demolombe, R. and Jones, A. J. (1999). On sentences of the kind "sentence 'p'is about topic t". In *Logic, Language and Reasoning*, pages 115–133.

Dias, J., Mascarenhas, S., and Paiva, A. (2014). Fatima modular : Towards an agent architecture with a generic appraisal framework. In *Emotion modeling*, pages 44–56.

Dignum, F. (1999). Autonomous agents with norms. *Artificial intelligence and law*, 7(1) :69–79.

Douceur, J. R. (2002). The sybil attack. In *International Workshop on Peer-to-Peer Systems*, pages 251–260.

Doyle, J., Shoham, Y., and Wellman, M. P. (1991). A logic of relative desire. In *International Symposium on Methodologies for Intelligent Systems*, pages 16–31.

Drawel, N., Bentahar, J., and Shakshuki, E. (2017). Reasoning about trust and time in a system of agents. *Procedia Computer Science*, 109 :632–639.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2) :268–298.

Dundua, B. and Uridia, L. (2010). Trust and belief, interrelation. In *3rd Workshop on Agreement Technologies*, pages 161–179.

Elgesem, D. (1997). The modal logic of agency. *Journal of Philosophical Logic*, 2(2).

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4) :99–117.

Emerson, E. A. and Halpern, J. Y. (1986). "sometimes" and "not never" revisited : on branching versus linear time temporal logic. *Journal of the ACM*, 33(1) :151–178.

Erickson, J. (2008). *Hacking : the art of exploitation*. No Starch Press.

Ettinger, D. and Jehiel, P. (2010). A theory of deception. *Microeconomics*, 2(1) :1–20.

Faden, R. R. and Beauchamp, T. L. (1986). *A history and theory of informed consent*. Oxford University Press.

Fagin, R. and Halpern, J. Y. (1987). Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34(1) :39–76.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. (2004). *Reasoning about knowledge*. MIT press.

Falcone, R., Pezzulo, G., and Castelfranchi, C. (2002). A fuzzy approach to a belief-based trust computation. In *Workshop on Deception, Fraud and Trust in Agent Societies*, pages 73–86.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2) :293–315.

Fan, T.-F. and Liau, C.-J. (2016). Reasoning about justified belief based on the fusion of evidence. In *European Conference on Logics in Artificial Intelligence*, pages 240–255. Springer.

Ferber, J. and Weiss, G. (1999). *Multi-agent systems : an introduction to distributed artificial intelligence*. Addison-Wesley Reading.

Fitting, M. (1983). *Proof methods for modal and intuitionistic logics*. Science & Business Media.

Florea, A., Kayser, D., and Pentiuc, S. (2019). Architecture des agents et langages.

Fox, G. (2001). Peer-to-peer networks. *Computing in Science & Engineering*, 3(3) :75–77.

Franklin, S. and Graesser, A. (1996). Is it an agent, or just a program ? : A taxonomy for autonomous agents. In *International Workshop on Agent Theories, Architectures, and Languages*, pages 21–35.

Gärdenfors, P. (1976). Manipulation of social choice functions. *Journal of Economic Theory*, 13(2) :217–228.

Gibbard, A. (1973). Manipulation of voting schemes : a general result. *Econometrica*, pages 587–601.

Giordano, L., Martelli, A., and Schwind, C. (2000). Ramification and causality in a modal action logic. *Journal of logic and computation*, 10(5) :625–662.

Girle, R. (2014). *Modal logics and philosophy*. Routledge.

Givant, S. and Halmos, P. (2008). *Introduction to Boolean algebras*. Springer Science & Business Media.

Goodin, R. E. (1980). Manipulatory politics. *The journal of Politics*.

Goré, R. (1999). Tableau methods for modal and temporal logics. In *Handbook of tableau methods*, pages 297–396. Springer.

Grice, H. P. (1991). *Studies in the Way of Words*. Harvard University Press.

Grossi, D., Lorini, E., and Schwarzentruber, F. (2015). Ceteris paribus structure in logics of game forms. *Journal of Artificial Intelligence Research*, 53 :91–126.

Gunderson, J. G. (1984). *Borderline personality disorder*. SUNY Press.

Guo, M. and Conitzer, V. (2010). Strategy-proof allocation of multiple items between two agents without payments or priors. In *9th International Conference on Autonomous Agents and Multiagent Systems*, pages 881–888.

Haarslev, V. and Möller, R. (2001). Racer system description. In *International Joint Conference on Automated Reasoning*, pages 701–705.

Handelman, S. (2009). *Thought manipulation : the use and abuse of psychological trickery*.

Hansen, J. (2006). The paradoxes of deontic logic : Alive and kicking. *Theoria*, 72(3) :221–232.

Harel, D., Kozen, D., and Tiuryn, J. (2000). *Dynamic logic*. MIT Press.

Herzig, A. and Longin, D. (2000). Belief dynamics in cooperative dialogues. *Journal of Semantics*, 17(2) :91–115.

Herzig, A., Lorini, E., Hübner, J. F., and Vercouter, L. (2010). A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1) :214–244.

Herzig, A., Lorini, E., Moisan, F., and Troquard, N. (2011). A dynamic logic of normative systems. In *International Conferences on Artificial Intelligence*, volume 2011, pages 228–233.

Heuerding, A., Jäger, G., Schwendimann, S., and Seyfried, M. (1996). The logics workbench lwb : A snapshot. *Euromath Bulletin*, 2(1) :177–186.

Hill, B. (2010). Awareness dynamics. *Journal of Philosophical Logic*, 39(2) :113–137.

Hilpinen, R. (2012). *Deontic logic : Introductory and systematic readings*, volume 33. Springer Science & Business Media.

Hintikka, J. (1965). Knowledge and belief. an introduction to the logic of the two notions. 16 :119–122.

Hoffman, K., Zage, D., and Nita-Rotaru, C. (2009). A survey of attack and defense techniques for reputation systems. *ACM Computing Surveys*, 42(1) :1–17.

Horrocks, I. R. (1997). *Optimising tableaux decision procedures for description logics*. University of Manchester Manchester, UK.

Huang, Z., Masuch, M., and Pólos, L. (1996). Alx, an action logic for agents with bounded rationality. *Artificial Intelligence*, 82(1-2) :75–127.

Hurley, N., Cheng, Z., and Zhang, M. (2009). Statistical attack detection. In *Conference on Recommender Systems*, pages 149–156.

Hustadt, U., Schmidt, R. A., and Weidenbach, C. (1999). MSPASS : Subsumption Testing with SPASS. In *Description Logic Workshop*, pages 136–137. Linköping University.

Huth, M. and Ryan, M. (2004). *Logic in Computer Science : Modelling and reasoning about systems*. Cambridge university press.

Jennings, N. and Wooldridge, M. (1996). Software agents. *IEEE Review*, 42(1) :17–20.

Johnson, P. E., Grazioli, S., and Jamal, K. (1993). Fraud detection : Intentionality and deception in cognition. *Accounting, Organizations and Society*, 18(5) :467–488.

Jones, A. J. and Sergot, M. (1996). A formal characterisation of institutionalised power. *Logic Journal of the IGPL*, 4(3) :427–443.

Josang, A. and Ismail, R. (2002). The beta reputation system. In *15th Bled Electronic Commerce Conference*, pages 2502–2511.

Joule, R.-V., Beauvois, J.-L., and Deschamps, J. C. (2002). *Petit traité de manipulation à l'usage des honnêtes gens*. Presses universitaires de Grenoble.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kamvar, S. D., Schlosser, M. T., and Garcia-Molina, H. (2003). The EigenTrust algorithm for reputation management in P2P networks. In *12th International Conference on World Wide Web*, pages 640–651.

Kant, V. and Bharadwaj, K. K. (2013). Fuzzy computational models of trust and distrust for enhanced recommendations. *International Journal of Intelligent Systems*, 28(4) :332–365.

Kligman, M. and Culver, C. M. (1992). An analysis of interpersonal manipulation. *The Journal of Medicine and Philosophy*, 17(2) :173–197.

Knobbout, M. and Dastani, M. (2012). Reasoning under compliance assumptions in normative multiagent systems. In *11th International Conference on Autonomous Agents and Multiagent Systems*, pages 331–340.

Kripke, S. A. (1959). A completeness theorem in modal logic. *Journal of Symbolic Logic*, 24(01) :1–14.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1) :79–86.

Lang, J., Van Der Torre, L., and Weydert, E. (2002). Utilitarian desires. *Journal of Autonomous Agents and Multi-agent Systems*, 5(3) :329–363.

Lang, J., Van Der Torre, L., and Weydert, E. (2003). Hidden uncertainty in the logical representation of desires. In *International Conferences on Artificial Intelligence*, pages 685–690.

Larousse, P. (1867). *Grand dictionnaire universel du XIXe siècle*. Larousse.

Lemaître, M. and Verfaillie, G. (2007). Interaction between reactive and deliberative tasks for on-line decision-making. In *7th Workshop on Planning and Plan Execution for Real-world Systems*.

Leturc, C. and Bonnet, G. (2017). Une logique modale normale de la confiance. In *11e Journées d'Intelligence Artificielle Fondamentale*.

Leturc, C. and Bonnet, G. (2018a). A normal modal logic for trust in the sincerity. In *17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 175–183.

Leturc, C. and Bonnet, G. (2018b). Une logique modale pour la caractérisation des manipulations entre agents autonomes. In *12e Journées d'Intelligence Artificielle Fondamentale*.

Levesque, H. J. (1984). A logic of implicit and explicit belief. In *Association for the Advancement of Artificial Intelligence*, pages 198–202.

Li, Z. (2008). *Efficient and generic reasoning for modal logics*. PhD thesis, University of Manchester, UK.

Liau, C.-J. (2003). Belief, information acquisition, and trust in multi-agent systems—a modal logic formulation. *Artificial Intelligence*, 149(1) :31–60.

Liu, F. (2010). Von wright's "the logic of preference" revisited. *Synthese*, 175(1) :69–88.

Lorini, E. and Demolombe, R. (2008). From binary trust to graded trust in information sources : a logical perspective. In *International Workshop on Trust in Agent Societies*, pages 205–225. Springer.

Lorini, E., Jiang, G., and Perrussel, L. (2014). Trust-based belief change. In *21st European Conference on Artificial Intelligence*, pages 549–554.

Lorini, E. and Sartor, G. (2016). A stit logic for reasoning about social influence. *Studia Logica*, 104(4) :773–812.

Mahon, J. E. (2008). Two definitions of lying. *International Journal of Applied Philosophy*, 22(2) :211–230.

Maillat, D. and Oswald, S. (2009). Defining manipulative discourse : The pragmatics of cognitive illusions. *International Review of Pragmatics*, 1(2) :348–370.

Maoz, Z. (1990). Framing the national interest : The manipulation of foreign policy decisions in group settings. *World Politics*, 43(1) :77–110.

McNamara, P. (2006). Deontic logic. In *Handbook of the History of Logic*, volume 7, pages 197–288. Elsevier.

Meyer, J.-J. C. et al. (1988). A different approach to deontic logic : deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29(1) :109–136.

Mills, C. (1995). Politics and manipulation. *Social Theory and Practice*, 21(1) :97–112.

Modica, S. and Rustichini, A. (1994). Awareness and partitional information structures. *Theory and Decision*, 37(1) :107–124.

Muller, G. and Vercouter, L. (2004). Détection décentralisée d'agents menteurs. *Journées Francophones sur les Systèmes Multi-Agents*, pages 243–248.

Nehring, K. and Puppe, C. (2007). The structure of strategy-proof social choice—part i : General characterization and possibility results on median spaces. *Journal of Economic Theory*, 135(1) :269–305.

Noggle, R. (1996). Manipulative actions : a conceptual and moral analysis. *American Philosophical Quarterly*, 33(1) :43–55.

Pacuit, E. (2017). *Neighborhood semantics for modal logic*. Springer.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking : bringing order to the web. Technical report, Stanford University.

Panait, L. and Luke, S. (2005). Cooperative multi-agent learning : The state of the art. *Autonomous agents and multi-agent systems*, 11(3) :387–434.

Parkes, D. C. and Ungar, L. H. (2000). Preventing strategic manipulation in iterative auctions : Proxy agents and price-adjustment. In *Association for the Advancement of Artificial Intelligence*, pages 82–89.

Patel-Schneider, P. F. (1998). DLP system description. In *98 Description Logic Workshop*, pages 87–89.

Pauly, M. (2002). A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1) :149–166.

Pörn, I. (2012). *Action theory and social science : Some formal models*. Springer.

Poulin, R. (2010). Parasite manipulation of host behavior : an update and frequently asked questions. *Advances in the Study of Behavior*, 41 :151–186.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind ? *Behavioral and Brain Sciences*, 1(4) :515–526.

Primiero, G. and Taddeo, M. (2012). A modal type theory for formalizing trusted communications. *Journal of Applied Logic*, 10(1) :92–114.

Prior, A. N. (1967). *Past, present and future*, volume 154. Clarendon Press Oxford.

Rao, A. S. and Georgeff, M. P. (1991). Modeling rational agents within a bdi-architecture. *KR*, 91 :473–484.

Raz, J. (1986). *The morality of freedom*. Clarendon Press.

Reis, H. D. S. (2012). *Lie to me : Lying virtual agents*. PhD thesis, Universidade Técnica de Lisboa.

Rigotti, E. (2005). Towards a typology of manipulative processes. In de Saussure, L. and Schulz, P., editors, *Manipulation and ideologies in the twentieth century : discourse, language, mind*, pages 61–83. John Benjamins Publishing Company.

Robinson, M. S. (1985). Collusion and the choice of auction. *The RAND Journal of Economics*, pages 141–145.

Ruan, Y. and Durresi, A. (2016). A survey of trust management systems for online social communities – trust modeling, trust inference and attacks. *Know ledge-Based Systems*, 106 :150–163.

Rudinow, J. (1978). Manipulation. *Ethics*, 88(4) :338–347.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence : a modern approach*. Pearson Education Limited.

Sakama, C., Caminada, M., and Herzig, A. (2015). A formal account of dishonesty. *Logic Journal of the IGPL*, 23(2) :259–294.

Sanghvi, S. and Parkes, D. (2004). Hard-to-manipulate vcg-based auctions. *Harvard Univ., Cambridge, MA, USA, Tech. Rep.*

Santos, E. and Johnson, G. (2004). Toward detecting deception in intelligent systems. In *Enabling Technologies for Simulation Science VIII*, volume 5423, pages 130–142.

Santos, F. and Carmo, J. (1996). Indirect action, influence and responsibility. In *Deontic Logic, Agency and Normative Systems*, pages 194–215.

Satterthwaite, M. A. (1975). Strategy-proofness and arrow's conditions : Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2) :187–217.

Schipper, B. (2014). Awareness. *Handbook of Epistemic Logic*, pages 77–146.

Schmitz (2019). Cours de logique modale. https://ifac.univ-nantes.fr/IMG/pdf/Complet.pdf. Online ; accessed 25 septembre 2019.

Schwarzentruber, F. (2011). Lotrecscheme. *Electronic Notes in Theoretical Computer Science*, 278 :187–199.

Searle, J. R. (1969). *Speech acts : An essay in the philosophy of language*, volume 626. Cambridge university press.

Segerberg, K., Meyer, J.-J., and Kracht, M. (2009). The logic of action.

Shim, J. and Arkin, R. C. (2013). A taxonomy of robot deception and its benefits in HRI. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 2328–2335.

Shoham, Y. (1993). Agent-oriented programming. *Artificial Intelligence*, 60(1) :51–92.

Singh, A., Ngan, T.-W., Druschel, P., and Wallach, D. (2006). Eclipse attacks on overlay networks : Threats and defenses. In *IEEE 25th International Conference on Computer Communications*.

Singh, M. P. (2011). Trust as dependence : A logical approach. In *10th International Conference on Autonomous Agents and Multiagent Systems*, pages 863–870.

Smith, C., Ambrossio, A., Mendoza, L., and Rotolo, A. (2011). Combinations of normal and non-normal modal logics for modeling collective trust in normative MAS. In *International Workshop on AI Approaches to the Complexity of Legal Systems*, pages 189–203.

Specht, S. M. and Lee, R. B. (2004). Distributed denial of service : Taxonomies of attacks, tools, and countermeasures. In *International Conference on Parallel and Distributed Computing and Communication Systems*, pages 543–550.

Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical studies*, 128(1) :169–199.

Such, J. M. (2013). Attacks and vulnerabilities of trust and reputation models. In *Agreement Technologies*, pages 467–477.

Sunstein, C. R. (2015). Fifty shades of manipulation. *Journal of Marketing Behavior*, 213 :1–32.

Tennenholtz, M. (2004). Reputation systems : An axiomatic approach. In *20th Conference on Uncertainty in Artificial Intelligence*, pages 544–551.

Testerink, B. (2017). *Decentralized Runtime Norm Enforcement*. PhD thesis, Utrecht University.

Thimm, M. (2014). Tweety : A comprehensive collection of java libraries for logical aspects of artificial intelligence and knowledge representation. In *14th International Conference on the Principles of Knowledge Representation and Reasoning*, pages 528–537.

Todd, P. (2013). Manipulation. *The international encyclopedia of ethics*, 10 :3139–3145.

Troquard, N. (2014). Reasoning about coalitional agency and ability in the logics of "bringing-it-about". *International Conference on Autonomous Agents and Multiagent Systems*, 28(3) :381–407.

Turner, J. A., Deyo, R. A., Loeser, J. D., Von Korff, M., and Fordyce, W. E. (1994). The importance of placebo effects in pain treatment and research. *Journal of the American Medical Association*, 271(20) :1609–1614.

Vallée, T. (2015). *De la manipulation dans les systèmes multi-agents : une étude sur les jeux hédoniques et les systèmes de réputation*. PhD thesis, Université de Caen Normandie.

Vallée, T. and Bonnet, G. (2015). Using kl divergence for credibility assessment. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 1797–1798.

Vallée, T., Bonnet, G., Zanuttini, B., and Bourdon, F. (2014). A study of sybil manipulations in hedonic games. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 21–28. International Foundation for Autonomous Agents and Multiagent Systems.

Van Benthem, J., Girard, P., and Roy, O. (2009). Everything else being equal : A modal logic for ceteris paribus preferences. *Journal of philosophical logic*, 38(1) :83–125.

Van Dijk, T. A. (2006). Discourse and manipulation. *Discourse & Society*, 17(3) :359–383.

Van Ditmarsch, H., French, T., Velázquez-Quesada, F. R., and Wáng, Y. N. (2018). Implicit, explicit and speculative knowledge. *Artificial Intelligence*, 256 :35–67.

Van Ditmarsch, H., van der Hoek, W., and Kooi, B. (2005). Dynamic epistemic logic with assignment. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 141–148. ACM.

Van Ditmarsch, H., van Der Hoek, W., and Kooi, B. (2007). *Dynamic epistemic logic*, volume 337. Springer Science & Business Media.

Van Ditmarsch, H., Van Eijck, J., Sietsma, F., and Wang, Y. (2012). On the logic of lying. In *Games, Actions and Social Software*, pages 41–72.

Von Wright, G. H. (1951). Deontic logic. *Mind*, 60(237) :1–15.

Wagner, A. R. and Arkin, R. C. (2009). Robot deception : recognizing when a robot should deceive. In *International Symposium on Computational Intelligence in Robotics and Automation*, pages 46–54.

Wagner, A. R. and Arkin, R. C. (2011). Acting deceptively : Providing robots with the capacity for deception. *International Journal of Social Robotics*, 3(1) :5–26.

Walsh, T. (2009). Where are the really hard manipulation problems ? the phase transition in manipulating the veto rule. In *21th International Joint Conference on Artificial Intelligence*.

Wang, J.-L. and Huang, S.-P. (2007). Fuzzy logic based reputation system for mobile ad hoc networks. In *11th Knowledge-Based Intelligent Information and Engineering Systems*, pages 1315–1322.

Ware, A. (1981). The concept of manipulation : its relation to democracy and power. *British Journal of Political Science*, 11(2) :163–181.

Whaley, B. (1982). Toward a general theory of deception. *The Journal of Strategic Studies*, 5(1) :178–192.

Wilkinson, T. M. (2013). Nudging and manipulation. *Political Studies*, 61(2) :341–355.

Wood, A. D. and Stankovic, J. A. (2002). Denial of service in sensor networks. *Computer*, 35(10) :54–62.

Wooldridge, M. (2003). *Reasoning about rational agents*. MIT press.

Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley & Sons.

Wooldridge, M. and Jennings, N. R. (1994). Agent theories, architectures, and languages : a survey. In *International Workshop on Agent Theories, Architectures, and Languages*, pages 1–39.

Xu, M. (2010). Combinations of stit and actions. *Journal of Logic, Language and Information*, 19(4) :485–503.

Zanardo, A. (1996). Branching-time logic with quantification over branches : the point of view of modal logic. *Journal of Symbolic Logic*, 61(1) :1–39.

Zawieska, K. (2015). Deception and manipulation in social robotics. In *Workshop on the Emerging Policy and Ethic of Human-Robot Interaction*, pages 2–5.

Zhang, F. and Zhou, Q. (2012). A meta-learning-based approach for detecting profile injection attacks in collaborative recommender systems. *Journal of Computers*, 7(1) :226–234.

Abstract

In recent decades, computer development has shifted from designing individual software to designing intelligent, self-contained software called agents and interacting with others to form multi-agent systems. In such systems, malicious agents sometimes implement complex strategies to induce other agents to make decisions in their favor, without the latter noticing them. We are talking about manipulation strategies. These strategies may in some cases cause problems for the agents which are victims. Such strategies are always hidden from agents. How can we detect them ? Firstly, it is necessary to define manipulation. Thus, based on work from computer sciences and social sciences, we define manipulation as the deliberate intention of an agent to instrumentalize a victim while making sure to conceal that intent. We propose to answer this question, a logical system named KBE which expresses manipulation. We prove that KBE is correct and complete, and is able to express strategies based on knowledge and beliefs of agents, like lying or bullshiting. This system can also express notions such as coercion and persuasion. Secondly, since trust is a mechanism to regulate the interactions between agents when agents may be malicious or unreliable, we propose another logical system named TB. This system, proved to be correct and complete, expresses a notion of trust in sincerity which represents the choice of an agent to take the risk of believing another agent for its sincerity. Finally, we propose an algorithmic method to reason with such systems. This method is adapted to the TB system and decides on its satisfiability problem by directly using the frame constraints to build a model.

Résumé

Ces dernières décennies, le développement informatique est passé de la conception de logiciels individuels, à la conception de logiciels intelligents, autonomes, appelés agents et interagissant avec d'autres en formant des systèmes multi-agents. Dans de tels systèmes, il arrive que des agents malintentionnés mettent en œuvre des stratégies complexes pour inciter d'autres agents à prendre des décisions en leur faveur et ce, tout en veillant à ce que les autres ne s'en aperçoivent pas. Nous parlons alors de stratégies de manipulation. Ces stratégies peuvent dans certains cas causer des problèmes aux agents qui en sont victimes. De plus, elles sont toujours dissimulées aux agents. Comment pouvons-nous alors les détecter ? Dans un premier temps, il est nécessaire de définir ce que signifie « être une manipulation ». Ainsi, sur la base de travaux issus de l'informatique mais aussi des sciences sociales, nous définissons la manipulation comme l'intention délibérée d'un agent d'instrumentaliser une victime tout en veillant à lui dissimuler cette intention. Pour répondre à cette problématique, nous proposons un système logique nommé KBE. Ce système, prouvé correct et complet, permet d'exprimer et de raisonner sur la manipulation ainsi que sur des stratégies fondées sur les connaissances et les croyances des agents comme le mensonge ou le baratinage. Ce système permet également d'exprimer des notions connexes à la manipulation comme la coercition et la persuasion. Dans un second temps, puisque la confiance est un mécanisme permettant de réguler les interactions entre agents lorsque des agents sont malintentionnés ou non fiables, nous proposons un second système logique nommé TB. Ce système, correct et complet, exprime une notion de confiance en la sincérité qui représente le choix d'un agent de prendre le risque de croire un autre agent vis-à-vis de sa sincérité. Enfin, dans un dernier temps, nous proposons une méthode algorithmique pour raisonner avec de tels systèmes. Cette méthode est adaptée au système TB et décide de son problème de satisfiabilité en utilisant directement les contraintes du cadre pour construire un modèle.