



**HAL**  
open science

# Conception d'un Cadre Formel d'Interaction pour la Découverte Scientifique Computationnelle

Christopher Dartnell

► **To cite this version:**

Christopher Dartnell. Conception d'un Cadre Formel d'Interaction pour la Découverte Scientifique Computationnelle. Autre [cs.OH]. Université Montpellier II - Sciences et Techniques du Languedoc, 2008. Français. NNT: . tel-00351202

**HAL Id: tel-00351202**

**<https://theses.hal.science/tel-00351202>**

Submitted on 8 Jan 2009

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Numéro d'identification :

ACADÉMIE DE MONTPELLIER

U N I V E R S I T É   M O N T P E L L I E R   I I  
— SCIENCES ET TECHNIQUES DU LANGUEDOC —

T H È S E

présentée à l'Université des Sciences et Techniques du Languedoc  
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : **Informatique**  
*Formation Doctorale* : **Informatique**  
*École Doctorale* : **Information, Structures, Systèmes**

**Conception d'un Cadre Formel  
d'Interaction pour la Découverte  
Scientifique Computationnelle**

par

Christopher DARTNELL

Soutenue le 18 Juin, 2008 devant le Jury composé de :

Jean-Yves BÉZIAU, Professeur, Université de Neuchâtel, ..... Rapporteur  
Paul BOURGINE, Professeur, École Polytechnique, CREA , ..... Examineur  
Jacques FERBER, Professeur, Université Montpellier II, ..... Président  
Frédéric JEANMART, Ingénieur d'affaires, Euriware, ..... Co-Directeur de thèse  
Dominique LUZEAUX, Directeur du Centre Technique des S.I., DGA, ..... Examineur  
Engelbert MEPHU NGUIFO, Maître de conférences, Université d'Artois, ..... Rapporteur  
Jean SALLANTIN, Directeur de recherche, CNRS, ..... Directeur de thèse  
Jean-Jacques SZCZECINIARZ, Professeur, Université Paris 7, ..... Examineur



Numéro d'identification :

ACADÉMIE DE MONTPELLIER

U N I V E R S I T É    M O N T P E L L I E R    I I  
— SCIENCES ET TECHNIQUES DU LANGUEDOC —

T H È S E

présentée à l'Université des Sciences et Techniques du Languedoc  
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ                   : **Informatique**  
*Formation Doctorale*       : **Informatique**  
*École Doctorale*            : **Information, Structures, Systèmes**

**Conception d'un Cadre Formel  
d'Interaction pour la Découverte  
Scientifique Computationnelle**

par

Christopher DARTNELL

Soutenue le 18 Juin, 2008 devant le Jury composé de :

Jean-Yves BÉZIAU, Professeur, Université de Neuchâtel, ..... Rapporteur  
Paul BOURGINE, Professeur, École Polytechnique, CREA , ..... Examineur  
Jacques FERBER, Professeur, Université Montpellier II, ..... Président  
Frédéric JEANMART, Ingénieur d'affaires, Euriware, ..... Co-Directeur de thèse  
Dominique LUZEAUX, Directeur du Centre Technique des S.I., DGA, ..... Examineur  
Engelbert MEPHU NGUIFO, Maître de conférences, Université d'Artois, ..... Rapporteur  
Jean SALLANTIN, Directeur de recherche, CNRS, ..... Directeur de thèse  
Jean-Jacques SZCZECINIARZ, Professeur, Université Paris 7, ..... Examineur



*À mes grand-mères*  
*To my grandmothers*



## Remerciements

Je tiens à remercier sincèrement mon directeur de thèse, Jean Sallantin, pour avoir encadré mes travaux durant ces cinq années. Il a su me faire partager sa passion pour la science et sa vision enthousiaste des problèmes liés à la rationalité artificielle et à sa modélisation. Je tiens également à remercier Frédéric Jeanmart pour avoir supervisé mes travaux dans un contexte industriel. Il m'a fait découvrir la vie en entreprise, et m'a enseigné la rigueur ainsi que le respect des délais et des engagements. Ces deux personnages ont été les deux piliers complémentaires sans l'appui desquels je n'aurais pas pu effectuer cette thèse. Merci à eux pour leur enseignement riche et diversifié, je n'aurais pu acquérir cette expérience dans aucun manuel. Merci enfin à eux pour leur amitié, j'espère qu'ils savent que celle-ci est réciproque.

Je remercie respectueusement les membres du Jury, Jean-Yves Béziau, Paul Bourguin, Jacques Ferber, Dominique Luzeaux, Engelbert Mephu-Nguifo, et Jean-Jacques Szczeciniarz, pour l'intérêt dont ils ont fait preuve à l'égard de mon travail. Merci à eux pour leurs remarques constructives et leurs conseils avisés, et enfin merci d'avoir autorisée la soutenance de cette thèse.

Je remercie profondément toutes les personnes avec qui j'ai eu la chance de pouvoir collaborer au LIRMM, Hélène Hagège, Dominique Luzeaux, Eric Martin, Sylvain Charron, David Chavalarias, Jacques Divol, Patrice Duroux, Sébastien Krut, Mathias Paulin, ainsi que chez Euriware, Christophe Leclerc, Nicolas Debeaute, Pierre Demeestere, Christophe Volland et tous les collaborateurs du projet AKSIOM. Merci à Laetitia, ma RH préférée et merci à tout le personnel du bâtiment des Vindits pour sa gentillesse. Ces personnes ont rendu mon séjour à Cherbourg moins pluvieux, Serge et George ont su faire chanter les cigales quand la Méditerranée semblait trop loin.

Un grand merci à tous mes amis les nîmois, les dJ3uNz, les margoulins, pour m'avoir accompagné dans cette aventure et m'avoir offert d'excellents moyens de détente quand cela était nécessaire. J'ai une pensée spéciale pour mes collocataires, Belet', Krrrg, et Rami, qui m'ont offert soutien quotidien précieux, ainsi que pour les personnes avec qui j'ai fait un bout de chemin, même si ce n'était qu'un bout.

Merci à mon frère, Stephen, d'être là tout simplement et grâce à qui je ne me suis jamais senti seul. Merci enfin à mes parents, Martine et Gary pour leur amour, et merci à Michel pour son affection. Je ne serais pas devenu celui que je suis aujourd'hui sans l'éducation qu'ils m'ont donnée et sans leur soutien.



# Table des matières

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Dialectique, Didactique, et Épistémologie . . . . .	5
1.1.1	Origines de la Dialectique . . . . .	6
1.1.2	La Dialectique Moderne . . . . .	6
1.1.3	Le Père de nos Pairs . . . . .	8
1.1.4	Apprentissage et didactique . . . . .	10
1.2	Découverte Scientifique Computationnelle . . . . .	11
1.3	Interagir avec une Machine Apprenante . . . . .	14
1.3.1	Vers des Solveurs Adaptatifs . . . . .	14
1.3.2	Fonctionnalités attendues de la part d'un Assistant Rationnel . . . . .	16
1.3.3	Le jeu de construction et de révision de théories . . . . .	17
1.3.4	Le rôle de l'assistant . . . . .	17
1.3.5	Caractéristiques Logiques de la Découverte . . . . .	17
1.4	Synthèse . . . . .	20
<b>I</b>	<b>Aspects Théoriques</b>	<b>23</b>
<b>2</b>	<b>Vers un Apprentissage Interactif</b>	<b>25</b>
2.1	L'Identification à la Limite (1967) . . . . .	27
2.2	La <i>PAC</i> -apprenabilité (1984) . . . . .	30
2.3	L'Apprentissage à Partir de Requêtes (1988) . . . . .	33
2.4	De l'Apprentissage par Simulation à l'Apprentissage par Imitation (Learning from different teachers) . . . . .	36
2.4.1	Formalisation de l'Apprenti . . . . .	37
2.4.2	Interaction tuteur-apprenti . . . . .	42

2.4.3	Illustration de l'Interaction Tuteur-Apprenti . . . . .	43
2.4.4	Perspectives . . . . .	44
2.5	Synthèse . . . . .	45
<b>3</b>	<b>Une Localisation Dialectique des Contradictions</b>	<b>49</b>
3.1	Le Carré des Oppositions . . . . .	51
3.1.1	Présentation . . . . .	52
3.1.2	Introduction des Modalités . . . . .	54
3.1.3	Extensions et Prise en Compte de Contradictions . . . . .	56
3.2	Le Cube des jugements . . . . .	59
3.3	Synthèse . . . . .	63
<b>4</b>	<b>Formalisation du Protocole</b>	<b>65</b>
4.1	Langage Formel . . . . .	66
4.1.1	Langage de base . . . . .	67
4.1.2	Enrichissement du Vocabulaire . . . . .	69
4.1.3	Programes Logiques Univoques et Équivoques . . . . .	70
4.2	Apprentissage Passif . . . . .	72
4.3	Apprentissage Actif . . . . .	73
4.4	Qu'en est-il de l'apprentissage <i>PAC</i> ? . . . . .	74
4.5	Apprentissage Interactif et Validation Sociale de Découvertes	76
4.5.1	Restriction des Requêtes d'Appartenance . . . . .	76
4.5.2	Une Dynamique Sociale de Réfutation à la Limite . . . . .	77
4.6	Perspectives . . . . .	79
4.7	Synthèse . . . . .	80
<b>II</b>	<b>Aspects Pratiques</b>	<b>81</b>
<b>5</b>	<b>Présentation du Prototype <i>Éleusis+Nobel</i></b>	<b>83</b>
5.1	Eleusis . . . . .	84
5.1.1	Les origines du jeu . . . . .	84
5.1.2	Les règles du jeu . . . . .	85
5.1.2.1	La règle secrète . . . . .	85
5.1.2.2	Déroulement du jeu . . . . .	86
5.1.2.3	Fin de la partie . . . . .	90

---

5.2	Nobel . . . . .	91
5.3	Eleusis+Nobel . . . . .	92
5.3.1	Les règles cachées . . . . .	92
5.3.2	Les espaces d'expérimentation . . . . .	94
5.3.3	Le protocole de publication . . . . .	94
5.4	Synthèse . . . . .	95
<b>6</b>	<b>Implantation</b>	<b>97</b>
6.1	Vers une Modélisation <i>AGR</i> d'un Protocole d'Apprentissage Distribué . . . . .	98
6.1.1	Agents . . . . .	99
6.1.2	Groupes . . . . .	100
6.1.3	Rôles . . . . .	100
6.1.4	Messages et Protocole d'Interaction . . . . .	100
6.1.4.1	Expériences . . . . .	101
6.1.4.2	Publications et Réfutations . . . . .	101
6.2	Implantation . . . . .	102
6.2.1	La Plate-Forme + Interfaces . . . . .	102
6.2.2	Le Serveur . . . . .	103
6.2.3	Le Client . . . . .	104
6.2.3.1	L'Espace Public . . . . .	104
6.2.3.2	L'Espace d'Expérimentation (privé) . . . . .	105
6.2.3.3	L'Espace de Publication (privé) . . . . .	106
6.2.4	Les Agents <i>Problem</i> . . . . .	108
6.2.5	Le Panneau d'Informations . . . . .	109
6.3	Extensions . . . . .	109
6.3.1	Ontologie Libre . . . . .	109
6.3.2	Stratégie et Politique . . . . .	110
6.3.3	Introduction des Assistants Rationnels . . . . .	113
6.3.4	Introduction d'Heuristiques . . . . .	114
6.3.5	Adaptation à d'autres Domaines . . . . .	115
6.4	Synthèse . . . . .	115

<b>7</b>	<b>Expérimentations et Résultats</b>	<b>117</b>
7.1	Impact de la Résolution Distribuée des Requêtes d'Équivalence . . . . .	117
7.1.1	Conditions expérimentales . . . . .	118
7.1.2	Problèmes rencontrés . . . . .	118
7.1.3	Exploitation des Données . . . . .	119
7.1.4	Dynamique Sociale . . . . .	119
7.1.4.1	Traitement des données . . . . .	119
7.1.4.2	Résultats et Cognition Sociale . . . . .	121
7.1.5	Dynamique Individuelle . . . . .	123
7.2	Impact Épistémologique et Intérêt Pédagogique . . . . .	126
7.2.1	Problématique . . . . .	127
7.2.2	Cadre théorique de l'étude . . . . .	127
7.2.3	Méthodologie . . . . .	129
7.2.3.1	Procédure et Sujets de l'étude . . . . .	129
7.2.3.2	Outils de mesure . . . . .	130
7.2.3.3	Analyse des résultats . . . . .	131
7.2.4	Résultats et discussion . . . . .	131
7.2.4.1	Changements dans les réponses du post-test spécifiques à la population <i>PI</i> . . . . .	135
7.2.4.2	Changements putatifs de conception spécifiques à <i>PI</i> . . . . .	138
7.2.5	Discussion . . . . .	140
7.3	Conclusion . . . . .	141
<b>8</b>	<b>Conclusion et Perspectives</b>	<b>143</b>
	<b>Bibliographie</b>	<b>151</b>
	<b>A Publications</b>	<b>161</b>
	<b>B Deux Définitions de la Complexité : Complexité Aléatoire et Complexité Organisée</b>	<b>163</b>
B.1	Théorie de la Complexité de Kolmogorov . . . . .	163
B.2	La profondeur logique de Bennett : complexité aléatoire et complexité organisée . . . . .	165

---

C	Questionnaire Présenté lors des Expériences	167
D	Évolution des formalismes logiques	169



# Chapitre 1

## Introduction

*« Les théories sont des filets destinés à capturer ce que nous appelons “le monde” : à le rendre rationnel, à l’expliquer et le maîtriser. Nous nous efforçons de resserrer de plus en plus les mailles. »*

Karl R. Popper.

LES TRAVAUX présentés dans ce document sont le fruit d’un travail collaboratif et pluri-disciplinaire axé sur la définition d’une plate-forme simulant la découverte scientifique. Les aspects conceptuels et formels ont beaucoup évolué au fur et à mesure que leurs limites étaient mises en évidence par les aspects expérimentaux (cette évolution est visible en relisant les articles correspondant [Afshar et al, 2007], et [Luzeaux et al, 2008], voir Annexe A). La définition du prototype *Éleusis+Nobel* m’a donc permis de centraliser les efforts de logiciens et de chercheurs en sciences cognitives et sociales, et a inspiré divers travaux dans ces domaines. Je vais tâcher dans ce document de synthétiser et d’organiser les différents points de vue.

Les fondements de mes recherches se situent à la croisée de l’apprentissage humain, de l’apprentissage automatique, et de la logique. Les points de départ de cette thèse ont été les travaux menés par Germana DaNobrega, Jacques Divol, et Jean Sallantin pour la réalisation d’un outil de construction interactive d’ontologies, WebRa. Cet outil permet à un *utilisateur* de décrire le *domaine* qui le concerne (dans ce cas, des contrats juridiques) afin de générer des exemples de contrats type. Ces derniers sont utilisés par la machine pour *apprendre* les *règles* qui distinguent un bon contrat d’un mauvais (problème de classification), dans le but d’automatiser la création de contrats personnalisés et pré-remplis. Cette façon de procéder est typique de ce que l’on nomme aujourd’hui la « découverte scientifique computationnelle » [Langley, 2000, Bridewell *et al.* , 2006], c’est à dire la découverte automatisée

ou assistée par ordinateur : un expert supervise l'apprentissage d'une machine en jouant sur le jeu d'exemples qui lui sont fournis, ou en modifiant le modèle de description. Ainsi, le formatage et l'encodage de modèles faisant intervenir d'hypothétiques variables prédictives et des données expérimentales permettent à des assistants plus ou moins rationnels d'apprendre des règles liant des variables comportementales à ces variables prédictives.

Les divers partenaires industriels de l'équipe se sont appropriés l'outil, et l'ont appliqué à leurs besoins respectifs : WebRa deviendra « JPhiined2<sup>TM</sup> », un outil d'aide au diagnostic de pannes, chez Euriware (mon employeur), ou « Kem<sup>TM</sup> », un outil d'aide au criblage de molécules chez Ariana Pharma [Afshar *et al.*, 2006]. Le but premier de cette thèse était de faire évoluer JPhiined2 vers un outil plus générique d'assistance à la supervision de systèmes complexes. En prenant un peu de recul, mon intérêt pour l'interaction et les systèmes multi-agents m'a conduit à la définition logique d'un protocole d'interaction afin de donner un cadre formel à une résolution distribuée de problèmes, m'éloignant quelque peu de cet objectif. Toutefois, ce protocole peut tout à fait être appliqué à une communauté d'agents (chercheurs humains ou apprentis logiciels) cherchant à comprendre un système complexe.

En effet, l'utilisation de WebRa et de ses dérivés industriels a montré un effet inattendu de l'interaction avec une machine apprenante : l'apparition d'un jeu dialectique de localisation et de correction d'erreurs, aussi bien dans les règles apprises par la machine que dans l'ontologie écrite par l'utilisateur. Ces erreurs mettent en évidence des biais de modélisation ou de mesure, ainsi que des biais liés aux exemples fournis à la machine. C'est en réalité ce processus dialectique qui est exploité dans les différentes applications, et l'apprentissage ne sert finalement qu'à générer des contradictions entre la théorie de la machine et celle de l'utilisateur. Cette constatation m'a incité à définir un cadre formel pour cette dialectique. Dans ce contexte, j'ai décidé de positionner ce mode d'interaction avec une machine apprenante par rapports aux principaux paradigmes d'apprentissage, afin de déterminer un protocole d'interaction adapté à la découverte scientifique computationnelle, c'est à dire adapté à l'apprentissage humain comme à l'apprentissage machine, et tenant compte d'une certaine démarche scientifique.

Le protocole d'interaction vers lequel je me suis orienté (voir Chapitre 4) peut être considéré comme un échange de *requêtes* permettant de confronter des jugements scientifiques à propos d'énoncés logiques. Une fois l'intuition de ce protocole dégagée (voir Chapitre 2), sa définition logique m'a occupé un certain temps, et plusieurs formalisations se sont succédées au fur et à mesure que nous<sup>1</sup> découvrons de nouveaux problèmes :

---

1. Dans ce document, j'utiliserai le sujet « nous » chaque fois que le travail d'équipe a

- Quelle logique est adaptée à un raisonnement non-trivial en situation de connaissances incomplètes et potentiellement contradictoires ?
- Les présupposés logiques concernant les changements d’avis et les différents niveaux de pertinence des lois découvertes doivent-ils caractériser un raisonnement individuel ou une forme de cognition sociale associée à l’interaction elle-même ?
- Les différents jugements scientifiques intervenant durant le processus de découverte peuvent-ils être définis logiquement, et surtout peut-on circonscrire les éventuelles contradictions entre ces jugements, qu’ils soient émis par un seul individu ou par des individus différents ?

Avant de rentrer dans le vif du sujet et présenter la formalisation qui représente l’état actuel de nos travaux, je vais donner en Section 1.1 un bref point de vue philosophique à propos de la dialectique et du rôle de l’erreur en science. Je reviendrai ensuite en Section 1.2 sur la découverte computationnelle afin d’en retirer les aspects essentiels, puis je résumerai nos différentes approches face aux questions précédentes en Section 1.3.5. Cette introduction justifiera ainsi, je l’espère, l’intérêt que je porte à l’analyse des cadres formels d’apprentissage (Chapitre 2) et à la formalisation d’une logique adaptée pour traiter les produits de l’apprentissage de façon « scientifique » (Chapitre 3).

## 1.1 Dialectique, Didactique, et Épistémologie

La découverte scientifique est généralement considérée comme l’une des activités humaines les plus compliquées, et mérite à ce titre d’être comprise pour des raisons théoriques et pratiques. La Science ne s’est pas construite uniquement par l’accumulation de découvertes faites par hasard, elle est aussi orientée par une curiosité éternelle de l’esprit pour certains aspects du monde, qui s’exprime dans une insatisfaction face à des constatations brutes [Harthong, 1992]. Depuis le début de la civilisation, l’homme cherche des explications sur sa nature et sur la nature des phénomènes qui l’entourent. Cependant, ce n’est qu’au début du *XVII<sup>e</sup>* siècle que l’on se propose d’articuler ce que nous appelons aujourd’hui la méthode de la science moderne [Chalmers, 1993], notamment à partir des idées de Francis Bacon. Dans son - “*Novum Organum*” (1620), Bacon défend l’idée selon laquelle on accéderait à la connaissance par la collecte de faits, par une observation organisée pour élaborer des théories. Sa conception empiriste de la science allait contre une importante tradition scolastique, initiée avec Platon. Je vais tenter de résumer l’évolution de la notion de dialectique à travers un extrait de [Clément *et al.* , 1994, Lecourt, 1999].

---

primé sur le travail individuel, comme dans ce cas, ou lorsque j’ai participé aux travaux cités mais que mon rôle est resté secondaire.

Deux grandes périodes sont à distinguer, depuis l'origine grecque du terme jusqu'à la vision moderne de la dialectique. Lors de la première période, la dialectique porte sur la démarche philosophique en général, à partir de son rapprochement à la méthode mathématique de démonstration par l'absurde, en allant jusqu'à sa portée sur la spécificité du discours philosophique. La seconde période élargira sa portée à la démarche scientifique, notamment en ce qui concerne l'indissociabilité de la science et de la critique, par le biais de la contradiction.

### 1.1.1 Origines de la Dialectique

On attribue généralement à Zénon d'Élée l'invention de la dialectique. Chez cet auteur, puis chez les socratiques en général, la dialectique est une méthode d'argumentation et de réfutation, par questions et réponses. Elle cherche à mettre l'adversaire en difficulté, notamment en montrant les contradictions de son discours [Clément *et al.*, 1994].

Pour Platon, la dialectique est l'art d'apprendre à parler et à penser, appuyé sur le dialogue. La dialectique est ainsi un art du dialogue polémique voué à la recherche de la vérité. En parlant de la dialectique comme un « art », Platon signifie que le dialogue polémique doit être conduit selon des règles déterminées. Celles-ci se concrétisent sous la forme de la réfutation socratique. Le dialogue doit procéder à un type de réfutation particulier où l'on considère les thèses de l'adversaire comme des hypothèses pour en déduire des conséquences jusqu'à ce qu'une contradiction apparaisse [Lecourt, 1999]. Platon attribue à la dialectique un statut plutôt scientifique, il la définit comme la science véritable [Clément *et al.*, 1994].

De science du vrai chez Platon, la dialectique devient technique du vraisemblable chez Aristote. Pour ce dernier, tandis que la science est d'essence démonstrative et porte sur un genre déterminé, la dialectique est d'essence argumentative et permet de soumettre n'importe quelle thèse à l'épreuve du pour et du contre. Aristote définit ainsi les règles qui conduisent la recherche du vraisemblable et du probable, à défaut du vrai et du nécessaire réservés à la science. Mais parce qu'il existe des domaines qui échappent à la science, la dialectique est un art utile et même nécessaire [Clément *et al.*, 1994].

### 1.1.2 La Dialectique Moderne

Emmanuel Kant, dans sa « critique de la raison pure » [Kant, 1845], présente une analyse des possibilités et des limites de la connaissance rationnelle du monde. Reprenant les distinctions aristotéliennes, il oppose la dialectique ou « logique de l'apparence », à l'analytique ou « logique de la vérité ». Plus

précisément, la « dialectique transcendantale » étudie les conditions de possibilité des contradictions dans lesquelles l'esprit s'empêtre nécessairement lorsqu'il fait un usage illégitime de ses facultés [Harthong, 1992]. L'idée centrale est que nous ne pouvons connaître le monde que par l'observation, donc par nos sens ; la *Raison* nous permet de comprendre ce que nous observons, mais dès que nous cherchons une connaissance sur le monde qui soit indépendante de l'observation, et qui serait donc en quelque sorte une connaissance absolue ou ontologique sur la chose en soi, la Raison ne permet plus de trancher. La recherche d'une vérité absolue ou ontologique sur le monde par la seule Raison s'opposait alors à la connaissance révélée (par les autorités religieuses) d'une telle vérité [Harthong, 1990]. Kant distingue donc les hypothèses *transcendantales*, qui ne peuvent être tranchées par l'expérience, des hypothèses *phénoménologiques* qui le sont de part leur caractère observable. Si la dialectique produit une apparence de savoir et engendre alors des antinomies, cette apparence est inévitable et ces contradictions insolubles, au moins sur un plan théorique [Clément *et al.* , 1994]. La gestion de ces contradictions ne fait alors plus appel au rationnel mais à l'émotionnel et des dogmes sont créés, on passe d'une logique de la preuve à une logique épistémique.

*Lorsque nous appliquons notre raison non plus simplement pour l'usage des principes de l'entendement à des objets de l'expérience, mais que nous essayons d'étendre ces principes au delà des limites de cette dernière, il se produit alors des propositions sophistiques qui n'ont ni confirmation à espérer, ni contradiction à craindre dans l'expérience et dont chacune non seulement est sans contradiction avec elle-même, mais trouve même dans la nature de la raison des conditions de sa nécessité, et malheureusement l'assertion du contraire est de son côté fondée sur des raisons tout aussi valables et tout aussi nécessaires.*

Kant, Critique de la Raison Pure.

Selon Hegel, la dialectique kantienne montre « l'objectivité de l'apparence et la nécessité de la contradiction ». Si Hegel considère cela comme positif, c'est qu'il voit le faux comme fécond, comme moment nécessaire du vrai. Pour lui, la pensée et l'être se développent dialectiquement selon un rythme ternaire : affirmation (ou thèse), négation (ou antithèse), négation de la négation (ou synthèse), dans laquelle les deux moments précédents sont dépassés et conservés (l'interprétation de la double négation, la pertinence du raisonnement par l'absurde ou de principe du tiers exclus sont des sujets qui nécessiteraient une discussion logico-philosophique approfondie qui dépasserait le cadre de ce document). La dialectique est donc ce « travail du négatif »

qu'elle porte sur la nature, ou sur la philosophie même. Chez Hegel, la dialectique n'a pas un statut de méthode, mais de processus d'auto-production du vrai et du savoir absolu, à partir de contradictions surmontées. Pour lui, tout savoir enveloppe la contradiction, et le savoir est toujours traversé par la contradiction. La formulation hégélienne de la dialectique provoque ainsi un déplacement du centre de gravité sur la contradiction et non plus sur le dialogue [Clément *et al.* , 1994, Lecourt, 1999].

En ce qui concerne la philosophie des sciences, l'intérêt porté à la dialectique hégélienne est dû à plusieurs aspects, parmi lesquels, le rôle constructif attribué à la négation dans la formation des concepts et la prise en compte de l'erreur comme élément moteur de la connaissance [Lecourt, 1999].

### 1.1.3 Le Père de nos Pairs

[Popper, 1963] propose de rompre avec l'épistémologie traditionnelle en plaçant la preuve hors de portée de la science. Il admet comme *vérité* un phénomène observable, et comme théorie une synthèse de ces phénomènes. L'intérêt de l'épistémologie Popperienne est de déplacer l'objet de la dialectique sur la réfutation des choses fausses plutôt que sur l'explication des choses vraies.

*Une théorie qui n'est réfutable par aucun événement qui se puisse concevoir est dépourvue de caractère scientifique. Pour les théories, l'irréfutabilité n'est pas (comme on l'imagine souvent) vertu mais défaut.*

Popper, Conjectures et réfutation p.64.

Cette façon d'appréhender la science entraîne la prise en compte de l'aspect incomplet et potentiellement inconsistant des connaissances lors du processus de découverte. En effet, les conjectures ou les théories, peuvent être trop spécifiques ou trop générales, et doivent pouvoir être remises en question tout au long de ce processus, lorsque des contradictions apparaissent. Ceci est applicable au processus de découverte en général, à l'apprentissage humain comme à l'apprentissage automatique. Les contradictions logiques font l'objet d'une incessante discussion qui dépasse le cadre non philosophique de ce document, mais nous reviendrons sur ce sujet au Chapitre 3. Notamment la logique paraconsistante, qui a connu un regain d'intérêt dans les années 1980 principalement grâce à J-Y Béziau, représente pour certains l'unique espoir de pouvoir formuler un raisonnement logique non trivial dans un tel contexte (voir Section 1.3.5).

Popper insiste donc sur le caractère nécessairement réfutable d'une théorie scientifique. Pourtant, [Stevenson, 1999] donne des arguments contre cette nécessité et contre le pouvoir de prédiction requis de la part d'une théorie. En

s'appuyant sur des propos d'Einstein qui croyait que les prédictions faites par la théorie de la relativité générale sont moins importantes que la « simplicité de ses fondements et [...] sa consistance logique », et de J.S. Mill, il met en avant certains exemples pour montrer que les prédictions réalisées à partir de théories ne les rendent pas plus scientifiques et ne leur donnent que peu de crédit supplémentaire. Je ne suis pas d'accord avec ces arguments, je vais donc m'expliquer.

*[...]it seems to be thought that an hypothesis[...] is entitled to a more favorable reception if, besides accounting for all the facts previously known, it has led[...] to the prediction of others which experience afterwards verified[...]. Such predictions and their fulfilment are, indeed, well calculated to strike the ignorant vulgar[...]. But it is strange that any particular stress should be laid upon such a coincidence by scientific thinkers.*

J. S. Mill, *A System of Logic, Ratiocinative and Inductive* (p. 296).

Selon [Stevenson, 1999], la prédiction du passage de la comète de Halley en 1758 ou la prédiction de l'existence de Neptune ne font que « montrer la confiance de leurs auteurs en la théorie de la gravitation universelle de Newton, et ne représente que des données supplémentaires lui donnant un peu plus de crédit ». Le fait que Mendeleïev ait laissé quatre cases vides dans son tableau périodique des éléments, prédisant par là-même les propriétés de ces quatre éléments qui furent découverts durant les vingt ans suivantes, n'a « ajouté que peu de preuves à celles qui existaient déjà ». Même si je suis d'accord sur le fait qu'une prédiction qui se réalise ne peut être vue comme une preuve de la théorie, je suis contre l'idée de négliger le pouvoir de prédiction d'une théorie scientifique. Je ne pense pas que les propos d'Einstein aient eu pour but de dénigrer la capacité de prédiction d'une théorie, mais plutôt de signaler que le fait que cette théorie soit également logiquement consistante et qu'elle puisse se formuler simplement la rend d'autant plus importante.

En effet, cette capacité de prédiction est selon moi le meilleur exemple de réfutabilité qu'une théorie scientifique puisse proposer. Si la prédiction se réalise la théorie ne fera que gagner un peu plus de crédit, comme le met en avant [Stevenson, 1999], et la prédiction est moins une preuve de la théorie qu'une nouvelle observation dont il faut tenir compte. Cependant si la prédiction ne se réalise pas, alors la théorie peut être réfutée, au moins en partie. Là encore, Stevenson cite l'exemple des cygnes noirs d'Australie dont l'observation ne réfute pas, selon lui, la théorie selon laquelle tous les cygnes sont blancs. L'argument qu'il met en avant est que la théorie reste vraie « en général ». Même si cet argument est justifié, ne pas remettre en question la théorie serait

une erreur. Une telle observation doit réfuter la théorie et entraîner un ensemble d'expériences pour circonscrire la contradiction qu'elle représente. Ceci peut mener soit à un effondrement de la théorie initiale qui sera alors éventuellement remplacée par une nouvelle, soit comme c'est le cas ici, restreindre son champ d'application et la rendre d'autant plus précise et plus difficilement réfutable. Le pouvoir prédictif d'une théorie est donc à mettre en relation avec une prise de risque que l'auteur est prêt à assumer.

Outre le fait de placer la l'objet de la démarche scientifique sur la réfutation expérimentale plutôt que sur la preuve, Popper plonge cette démarche dans un jeu social d'échange de conjectures. Je me suis inspiré de ce jeu social pour rendre opérationnel les paradigmes classiques d'apprentissage en les plongeant un système multi-agents où les apprentis soumettent leurs conjectures au jugement de leurs pairs.

### 1.1.4 Apprentissage et didactique

Comme nous venons de le voir, la compréhension et la formalisation du processus de découverte sont donc au coeur des travaux effectués en logique au siècle dernier. Qu'est-ce qui entraîne l'effet « Haha », l'« Euréka » d'Archimède, traduisant la découverte récente d'un nouveau concept, d'une nouvelle règle permettant d'expliquer un phénomène surprenant ? La modélisation de cette capacité étonnante du cerveau humain aurait des retombées importantes sur les travaux en Intelligence Artificielle, mais également sur la pédagogie mise en place dans nos écoles.

Il est intéressant de remarquer que si la dialectique occupe aujourd'hui une place prépondérante dans les conceptions modernes de l'apprentissage, elle n'est présente que de façon informelle dans les principaux paradigmes d'apprentissage automatique. Selon [Chan & Elliott, 2004], les conceptions traditionnelles et constructivistes de l'enseignement et de l'apprentissage peuvent toutes deux être opposées. Dans le premier cas, l'enseignement est considéré comme un transfert non problématique d'un savoir inchangé, d'un expert vers un novice. L'apprentissage correspond alors à l'absorption directe d'un tel savoir. Dans le second cas, l'apprentissage est considéré comme une création et une acquisition de connaissances à travers des raisonnements et des justifications. L'enseignement permet et facilite alors cet apprentissage, et ne peut pas être simplement réduit à une transmission de savoir. Comme le souligne [Lemberger *et al.*, 1999], beaucoup de futurs enseignants détiennent une conception positiviste<sup>2</sup> de la science et de la connaissance. Étant don-

---

2. Positivisme : « Toute conception prétendant s'en tenir aux faits du monde extérieur perceptibles par les sens et tendant à rejeter comme sans valeur toute connaissance ayant une autre origine. » (Dictionnaire de la langue philosophique)

née cette conception, les notions et les lois scientifiques existent en elles-mêmes, indépendamment de tout sujet, et sont une image exacte du monde [Fourez *et al.* , 1997]. Les sciences ne seraient pas produites pour les humains ; elles refléteraient plutôt le monde en lui-même. Cette vision peut-être qualifiée de réaliste - en opposition avec un point de vue idéaliste selon lequel le monde réel ne serait pas entièrement connaissable. La vision positiviste est cohérente avec les épistémologies *naïves* et *traditionnelles*, dans le sens où un savoir serait composé d'unités d'information progressivement ajoutées au gré du hasard et permettant par là même le progrès du savoir. En effet, une majorité d'enseignants du secondaire définissent l'enseignement comme un « transfert d'information maximal », et l'apprentissage comme une « absorption de toute information » ([Boulton-Lewis *et al.* , 2001, Waeytens *et al.* , 2002]). On peut de la même façon opposer les premiers paradigmes d'apprentissage automatique [Gold, 1967, Valiant, 1984], qui formalisent une conception traditionnelle de l'apprentissage, aux paradigmes plus récents qui sont plus proches de la constructiviste moderne (voir Chapitre 2).

## 1.2 Découverte Scientifique Computationnelle

En parallèle des améliorations physiques des ordinateurs, les recherches en informatique théorique n'ont cessé de générer de nouvelles techniques et algorithmes, suffisamment élaborés pour trouver des applications pratiques dans de nombreux domaines. Parmi ces thèmes, nous pouvons nommer ici :

- les techniques de *fouille de données* (data-mining), de *fouille de graphes* (graph-mining), ou encore de *pattern-matching*, qui ont rendu possible le séquençage de l'ADN ou le criblage et la synthèse de nouvelles molécules,
- les *systèmes multi-agents* qui permettent de simuler des environnements complexes pour calculer, par exemple, des mouvements astronomiques, ou des comportements sociaux,
- l'apparition plus récente de la *programmation par contraintes* qui a proposé une modélisation efficace de nombreux problèmes pratiques d'optimisation qui trouvent la plupart du temps des applications directes dans le domaine industriel,
- la formalisation de nouvelles logiques *défaisables* ou *paraconsistantes* qui ont rendu possible des formes de raisonnements plus proches de la démarche intellectuelle des scientifiques et des systèmes normatifs et institutionnels humains.

Peut-on pousser encore plus loin le rôle de l'ordinateur dans le processus de découverte ? Peut-on créer une machine dotée de rationalité et capable d'adopter une démarche scientifique pour assister un chercheur dans ses recherches ?

[Langley, 2002a] souligne la métaphore qui traite la découverte scientifique comme un calcul, et le fait que les travaux qui s'en sont inspirés pour l'automatiser ont finalement abouti à des découvertes qui ont fait l'objet de publications dans la littérature concernée (voir [Langley, 2000]). L'étude computationnelle de la découverte de connaissances scientifiques existe depuis [Simon, 1966], mais ce domaine n'est devenu vraiment actif que vers la fin des années 70, époque à laquelle les progrès en apprentissage automatisé ont motivé les tentatives pour reproduire de façon automatisée des découvertes scientifiques historiques. Après les succès dans ce domaine [Lenat, 1977, Langley *et al.*, 1981, Zytkow & Simon, 1986], l'intérêt s'est tout naturellement porté sur la découverte automatisée de nouvelles connaissances scientifiques, et un critère de succès qui commence à être atteint depuis la fin du siècle dernier est la publication de ces nouvelles théories : [Cheeseman *et al.*, 1988] appliquent leur système de classification Bayésien, AUTOCLASS, à la construction d'une taxonomie des étoiles fondée sur des données infra-rouges, [Lee *et al.*, 1998] proposent d'appliquer leur système d'induction de règles, *RL*, à la découverte de facteurs qualitatifs de la Cancérogénèse, et [King & Srinivasan, 1996] appliquent leur système PROGOL à la prédiction du caractère mutagène de composés chimiques.

S'étant tout d'abord focalisée sur la modélisation de phénomènes isolés et dans des contextes restreints, ce qui a été profitable à ce domaine autant qu'aux sciences naturelles elles-mêmes, la découverte scientifique computationnelle doit désormais s'adapter, comme la plupart des autres sciences, à l'intérêt grandissant pour la *science des systèmes complexes*. Ce domaine en plein essor a pour objet des systèmes de systèmes, beaucoup plus complexes en termes de nombre de composants en interaction et en termes d'organisation de ces composants. Nous renvoyons le lecteur intéressé à l'annexe B, et à [Beurier, 2007] pour un état de l'art fourni concernant la science des systèmes complexes et leur étude. La prise de conscience que les constituants d'un système peuvent s'organiser et faire émerger un comportement collectif, totalement différent de celui qui aurait pu être prédit en tenant compte des comportements individuels des constituants, remet en cause beaucoup de raisonnements réductionnistes en vertu desquels la décomposition d'un problème en sous-problèmes, permettait de trouver des solutions locales qui pourraient former, en les sommant en quelque sorte, une solution au problème initial. Le tout devient plus que la somme des parties, et [Laughlin, 2005] l'illustre de façon poétique :

*Des réalités banales comme la cohésion de l'eau ou la rigidité de l'acier sont des exemples simples, mais il y en a bien d'autres. La nature regorge de choses tout à fait fiabiles qui sont des versions primitives de la nature impressionniste. Un champ de fleurs rendu par Renoir ou Monet nous*

*intéresse vivement parce que c'est un tout parfait, alors que les touches dont il est fait ont une forme aléatoire et imparfaite*

Robert B. Laughlin, Un Univers Différent, 2005.

Mais les lois *non-fondamentales* ne sont pas pour autant bonnes à jeter, bien au contraire. La mécanique newtonienne émerge ainsi de la mécanique quantique, mais chacune d'entre elles n'est valable qu'à une échelle d'observation donnée, et devient inutile après un changement d'échelle. C'est ce facteur d'échelle qu'il me semble intéressant d'introduire dans le raisonnement d'un assistant rationnel afin de le rendre capable d'associer un ensemble de lois à un contexte d'application, et d'« oublier » celles-ci lorsque le contexte change.

Des récentes évolutions en découverte scientifique computationnelle, [Langley, 2002b] tire cinq leçons :

1. Les notations traditionnellement utilisées en apprentissage automatique ne sont pas facilement communicables aux chercheurs.
2. Les chercheurs ont généralement des modèles initiaux qui devraient influencer le processus de découverte.
3. Les données expérimentales<sup>3</sup> sont rarement abondantes, mais plutôt rares et difficiles à obtenir.
4. Les chercheurs veulent des modèles qui vont au delà de la description et fournissent des explications des données.
5. Les chercheurs veulent une assistance computationnelle plutôt que des systèmes de découvertes totalement automatisés.

1) et 5) montrent le besoin d'un langage logique proche de celui du chercheur pour manipuler les produits de l'apprentissage automatisé et mettre en place un processus dialectique de découverte. 2) justifie la démarche de [Nobrega *et al.*, 2003] : il faut permettre l'exploitation d'un modèle initial fourni par le chercheur, qui servira de formatage des données pour l'assistant, et orienter la dialectique vers la révision de ce modèle. 4) suggère qu'en plus de la capacité d'apprentissage, nous devrions doter l'assistant de capacités cognitives lui permettant d'exploiter un modèle pour expliquer des faits observés, prédire des faits à partir d'observations, et effectuer des simulations.

[Nobrega *et al.*, 2003] place sa contribution dans ce domaine à l'intersection entre l'acquisition de connaissances et l'apprentissage. Ce faisant, elle définit des *types de connaissances* afin de structurer les éléments de connaissance de l'agent, et les associe à des *éléments de communication* avec l'utilisateur : des actes de langages. Ces types de connaissances sont au coeur de son système

---

3. scientific data

de Phi-calcul qui les *transforme* en exploitant des contradictions. Ces notions et ces mécanismes sont en fait de nature profondément logique, et nous tâchons au Chapitre 3 de les exploiter dans un contexte logique plus rigoureux. Les travaux récents proposent une nouvelle vision de l'apprentissage supervisé en introduisant l'interaction au coeur du processus, favorisant ainsi la correction dialectique d'erreurs [Sallantin, 2004].

## 1.3 Interagir avec une Machine Apprenante

[Witten *et al.* , 1996] identifie plusieurs caractéristiques des techniques d'apprentissage automatique qui font obstacle à une interaction efficace avec un utilisateur.

En général, les modèles d'apprentissage automatique :

- traitent tous les exemple en même temps, alors que les systèmes interactifs fournissent les exemples de façon incrémentale, et nécessitent des prédictions instantanées,
- ne fournissent des résultats efficaces que sur un grand nombre d'exemples, alors qu'on peut difficilement demander à un utilisateur d'en fournir plus de deux ou trois avant de constater un progrès.
- sont souvent incapables d'exploiter des connaissances *à-priori* et des *avis* de l'utilisateur, alors que ce sont ces informations qui permettent d'apprendre à partir de peu d'exemples,

De plus, l'agent doit adapter la cible de son apprentissage en fonction des données :

- ces modèles supposent que les solutions soient consistantes avec l'ensemble des exemples, quel que soit le contexte, alors que toute nouvelle donnée peut potentiellement remettre en cause une théorie,
- opèrent en isolement, alors que les systèmes interactifs sont faits pour exploiter les *feedbacks*, les jugements de l'utilisateur (ou d'autres apprentis utilisant d'autre techniques, ou d'autres biais),

De plus, [Witten *et al.* , 1996] souligne le fait que l'apprentissage doit être soutenu, s'inscrire dans la durée, et être réversible.

### 1.3.1 Vers des Solveurs Adaptatifs

Les définitions usuelles d'un *Problem Solver* tiennent compte du type de problème qui le caractérise, comme dans le cas d'un solveur d'équations différentielles, ou un solveur d'équations non-linéaires. Les solveurs sont en général conçus pour résoudre des problèmes connus qui ont déjà été résolus et modélisés de façon à ce que, quelle que soit l'instance d'un problème spécifique, le

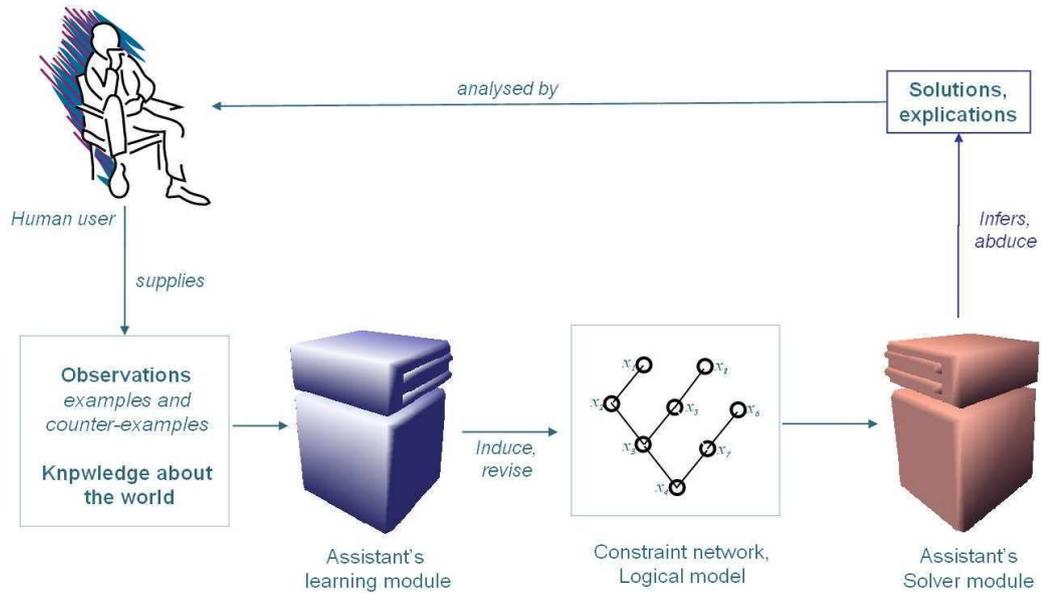


FIGURE 1.1 – Le Cycle d'Interaction Homme-Machine

solveur correspondant est capable de la résoudre et de produire les solutions. Un solveur adaptatif devrait ainsi être capable d'acquérir de nouvelles compétences en apprenant comment résoudre de nouveaux problèmes, et utiliser ses connaissances et son expérience pour s'adapter et en trouver les solutions.

Afin de résoudre un problème, il faut observer une situation problématique, puis l'analyser pour construire un langage permettant de la décrire en mettant en évidence les dimensions pertinentes pour la résolution du problème. Ces dimensions déterminent le domaine de définition des variables caractérisant le problème et entrant en jeu lors de sa résolution.

Ce langage est utilisé pour formuler des assumptions et des hypothèses qu'il faut ensuite valider par des expériences, en comparant les résultats théoriques et expérimentaux. L'expérimentation peut donc révéler des contradictions entre une théorie et la réalité, et donc conduire à la révision du langage de description et à la formulation de nouvelles hypothèses.

Les techniques d'apprentissage automatique ou de data-mining peuvent être utilisées ici pour mettre en évidence des régularités et générer de telles hypothèses. Par exemple, les techniques d'apprentissage fondées sur la théorie des graphes et treillis de Galois [Liquière, 1998], peuvent être utilisées pour trouver des implications logiques et des règles d'équivalence entre les dimensions du problème. Ces règles sont aisément interprétables par un humain puisqu'elles utilisent sa propre modélisation du problème.

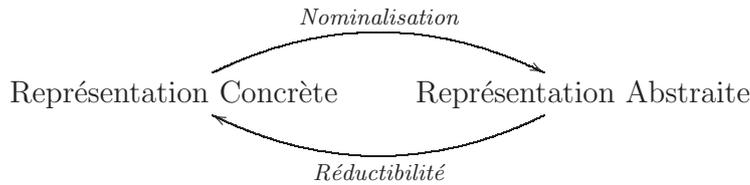


FIGURE 1.2 – Complémentarité des principes de Nominalisation et de Réductibilité (Cavaillès97)

---

### 1.3.2 Fonctionnalités attendues de la part d'un Assistant Rationnel

Dans ce contexte de découverte, dans lequel ni le langage ni la théorie ne sont connus *à-priori*, nous définissons maintenant les fonctionnalités attendues d'un solveur adaptatif.

- la capacité de construire et maintenir une *ontologie* du domaine. Par *ontologie*, nous entendons un langage logique décrivant, de façon cohérente avec les observations, les variables impliquées dans la résolution du problème. Cette ontologie est le produit du processus d'apprentissage ;
- la capacité d'analyser et de corréler l'information recueillie afin d'apprendre des énoncés ontologiques contraignant les relations entre les valeurs des différentes dimensions du problème, *i.e.*, la capacité d'inférer des règles logiques ;
- la capacité de découvrir, nommer, et utiliser de façon symbolique des régularités dans les données analysées, et réviser l'ontologie en introduisant de nouvelles dimensions dans la formulation du problème. Cette capacité à transformer une propriété observée en un objet symbolique est appelée *principe de nominalisation* ;
- la capacité de formuler et d'exprimer une *théorie* synthétisant les connaissances, afin d'expliquer le problème et de prédire des résultats futurs. Par *théorie*, nous entendons les règles de calcul des solutions du problème ;
- la capacité à mettre au point des expériences pour tester et (in)valider ces théories. Cette capacité est appelée le *principe de réductibilité*.

Les principes de nominalisation et de réductibilité (voir Figure 1.2) sont la clé de l'adaptabilité d'un solveur, en ce sens qu'ils permettent de construire une abstraction des phénomènes étudiés, mais aussi d'instancier ces concepts symboliques de façon concrète pour mettre au point les expériences permettant de valider la pertinence de ces concepts. La formalisation logique que nous proposons au Chapitre 4.1 tient compte de l'introduction de nouveaux concepts dans le langage de description.

L'interaction entre le solveur et son environnement est donc la condition *sine qua non* à son évolution permettant de localiser les contradictions entre une réalité et les théories censées l'expliquer. Nous nommons *assistant rationnel* un tel solveur adaptatif, qui interagit avec un chercheur afin de l'assister durant le processus de découverte menant à la résolution d'un problème. La Figure 1.1 synthétise cette interaction telle que nous venons de la présenter.

### 1.3.3 Le jeu de construction et de révision de théories

Les premières interactions vont ainsi permettre de construire l'ontologie servant à représenter l'objet de l'étude et de l'apprentissage et à formuler des théories sur cet(ces) objet(s). L'utilisateur construit donc un savoir avec l'aide de son assistant. Le système assiste la découverte humaine en mettant en évidence les contradictions de son instructeur. La seconde interaction intervient donc dans la résolution des paradoxes : l'agent rationnel, qui fait évoluer un programme, sera capable de relever des paradoxes dans la théorie de l'utilisateur provenant par exemple de faits contradictoires entre eux ou avec la théorie. Mais l'utilisateur doit aussi être capable de corriger son assistant s'il se trompe. Le dialogue entre l'humain et son assistant joue donc une place prépondérante dans les mécanismes de construction et de révision de l'ontologie.

### 1.3.4 Le rôle de l'assistant

L'apprentissage joue plusieurs rôles. Tout d'abord, par déduction à partir des exemples fournis par l'utilisateur, l'agent peut relever des paradoxes. Mais il peut aussi, par induction sur les exemples, proposer de nouvelles règles. Bien évidemment, ces règles peuvent être non fondées et il appartient donc à l'utilisateur de corriger son assistant en fournissant un nouvel exemple judicieusement choisi ou en définissant un nouveau terme dans son ontologie. Enfin, grâce aux exemples et aux règles, l'assistant peut formuler de nouvelles théories et entraîner ainsi l'utilisateur sur de nouvelles pistes.

### 1.3.5 Caractéristiques Logiques de la Découverte

Les axiomes du calcul propositionnel sont ceux de la logique classique. Ils permettent en particulier de prouver les formules  $P \vee \neg P$  (principe du tiers exclu) et  $\neg\neg P \rightarrow P$  (élimination de la double négation). Le principe du tiers-exclu stipule que toute chose ne peut-être caractérisée que par un attribut ou par son contraire. Par exemple, un animal est capable de voler ou ne l'est pas.

Une contradiction logique apparaît lorsqu'une proposition est à la fois vraie

et fausse :

$$p \wedge \neg p$$

Si la méthode de la preuve par l'absurde encourt parfois des reproches et si la question de sa suffisance se pose, c'est qu'elle met en évidence non pas la vérité de sa conclusion mais la fausseté de la négation de celle-ci. C'est en fondant le raisonnement sur le principe du tiers-exclu, et en éliminant la double négation que l'on peut conclure de ceci à la vérité de la proposition à démontrer. Ainsi, si l'on souhaite prouver qu'un animal donné est un oiseau, et qu'une règle stipule que les oiseaux sont les seuls animaux capables de voler ( $\forall x, Vole(x) \rightarrow Oiseau(x)$ ), on peut commencer par supposer que l'animal n'est pas un oiseau.

$$\textit{Hypothese} : \neg Oiseau(x)$$

En observant ce dernier s'élever dans les airs et en constatant qu'il vole, on se trouve face à une contradiction, puisque seuls les oiseaux volent.

$$\textit{Observation} : Vole(x)$$

$$Vole(x) \wedge \neg Oiseau(x) \wedge (Vole(x) \rightarrow Oiseau(x)) \rightarrow \perp$$

L'hypothèse de départ  $\neg Oiseau(x)$  peut ainsi être niée et l'on a  $\neg\neg Oiseau(x)$ . Par élimination de la double négation, on vient de prouver que l'animal est bien un oiseau.

$$\neg\neg Oiseau(x) \rightarrow Oiseau(x)$$

Cette démarche peut paraître abstraite en comparaison avec, par exemple, un simple enchaînement de déductions, telles que chaque proposition enchaînée « contienne la raison de celle qui suit, et soit elle-même démontrée par celle qui précède » [Piaget, 1924].

La logique intuitionniste est née de la remise en cause de l'élimination de la double négation. Ainsi, pour les intuitionnistes, un animal qui n'est pas un « *NON* oiseau » n'est pas forcément un oiseau. De plus, le principe du tiers-exclu sous entend que les connaissances représentées sont complètes et n'évolueront pas. Dans le contexte qui nous intéresse, les connaissances sont construites peu à peu, et sont constamment révisées et remises en question. Même si le raisonnement par l'absurde peut s'avérer utile, il sera donc nécessaire de le remettre en question si le raisonnement s'avère incohérent ou faux. Les logiques paraconsistantes remettent justement en question ce principe.

L'utilisation des contradictions en tant que moteur dialectique de la résolution de problème a donc pour effet de caractériser ce processus de la façon suivante :

- les modalités *déontiques* telles que l'*obligation*, l'*interdiction*, et la *permission*, permettent d'organiser des connaissances passées sous forme de normes ;
- la *paraconsistance* caractérise un raisonnement qui reste non trivial en présence de contradictions afin de maintenir les obligations ;
- la *défaisabilité* du système de normes permet d'organiser ces dernières de façon hiérarchique afin d'en faciliter la révision lorsqu'un fait contradictoire apparaît.

Tâchons maintenant de préciser les notions de paracomplétude et de paraconsistance. Un système déductif  $S$  est dit inconsistant s'il existe une formule  $\alpha$  de  $S$  telle qu'à la fois  $\alpha$  et sa négation  $\neg\alpha$  soient des théorèmes de  $S$ . Si toutes les formules de  $S$  sont des théorèmes, alors  $S$  est dit *trivial*, sinon  $S$  est dit *non-trivial*. Une logique paraconsistante est une logique dont le système déductif est inconsistant et *non-trivial*, *i.e.*, dans lequel une proposition et sa négation peuvent être toutes les deux vraies sans entraîner la vérité de toutes les autres propositions. Une logique paraconsistante peut donc être vue comme une logique dans laquelle le raisonnement par l'absurde ne s'applique pas, et qui reste non-triviale en présence contradictions logiques. De même, une logique est dite paracomplète si une proposition et sa négation peuvent être fausses toutes les deux sans entraîner la trivialité du système de déduction [DaCosta & Beziau, 1997]. Une logique paracomplète peut donc être vue comme une logique dans laquelle le principe du tiers-exclu ne s'applique pas. La paraconsistance et la paracomplétude semblent ainsi être des caractéristiques logiques nécessaires pour gérer logiquement les deux cas extrêmes que sont le surplus d'informations contradictoires et le manque d'information (respectivement).

[Nakamatsu *et al.* , 2003], par exemple, propose un tel exemple de paraconsistance fondée sur une logique déontique et défaisable. Les modalités de la logique déontique sont utilisées pour localiser les contradictions et provoquer une réorganisation de l'ensemble des normes défaisables. La paraconsistance permet au solveur d'adapter son ontologie aux nouvelles observations et de mettre en oeuvre un apprentissage incrémental.

Toutefois, [Béziau, 2007] montre que la logique classique est elle même paraconsistante. Si l'exploration de ces pistes de travail m'a permis de construire une vision logique de la notion de dialectique, j'ai délaissé ces formalismes incomplets pour les reformuler dans un cadre plus général. Je proposerai ainsi au Chapitre 3 d'utiliser des modalités pour définir de façon logique l'ensemble des jugements échangés par le chercheur et son assistant rationnel durant le processus de découverte. Ces jugements codifient l'interaction, permettant de typer des énoncés logiques en tant que conjectures, faits, postulats, etc. En

nous inspirant du carré des oppositions et de certaines de ses extensions, nous avons imaginé une structure cubique opposant chacun des jugements et garantissant que l'ensemble qu'elle définit est clos par négation, *i.e.*, que la négation de l'un d'entre eux ne nécessite jamais la définition d'une nouvelle modalité. De plus, [Luzeaux et al, 2008] met en évidence le fait que la logique sous-jacente à cette structure n'est autre que la logique  $S_5$ .

## 1.4 Synthèse

L'apprentissage, dans le cadre de l'assistance à la découverte scientifique, a pour but d'aider un humain à résoudre le problème et non de le faire résoudre par la machine, ce qui nécessite une interaction étroite entre l'agent rationnel et l'utilisateur. Les systèmes d'aide à la découverte, auxquels nous nous intéressons ici sont capables d'assister les scientifiques au long de leurs recherches ou, dans un contexte de résolution de problèmes, de résoudre des problèmes en coopération avec un expert du domaine. D'après le modèle proposé par [Nobrega *et al.*, 2003], l'expert guide le système dans la construction de son savoir en créant sa représentation des exemples et en répondant à ses questions. Le modèle duquel cet apprentissage se rapproche le plus est donc l'apprentissage à partir de requêtes (voir Chapitre 2), le rôle de l'oracle étant tenu par l'utilisateur. L'assistant va permettre à l'utilisateur de relever les paradoxes apparaissant dans sa théorie afin de reformuler le problème de façon plus juste pour le rendre plus facile à résoudre : un problème bien formulé est à moitié résolu.

Outre le fait que, dans le contexte expérimental dans lequel évolue un chercheur, les données sont sujettes à l'interprétation, la formalisation de la découverte doit tenir compte d'un contrôle logique et heuristique sur les produits de l'apprentissage (concepts, règles) afin de superviser l'exploration de l'espace des solutions et de localiser les contradictions permettant de remettre en cause ces produits.

L'originalité de ce travail est de rendre les paradigmes d'apprentissage opérationnels dans le contexte de la découverte scientifique, en définissant une structure logique permettant de tenir compte du caractère incomplet et inconsistant des produits de la découverte scientifique afin de permettre à une machine de tenir un raisonnement cohérent face aux faits, et d'adopter les comportements appropriés en cas de surprise, d'erreur, etc. Mes contributions résident dans l'étude comparée de ces paradigmes d'apprentissage, la définition logique des jugements scientifiques, et enfin la modélisation et l'implantation d'un prototype qui a été utilisé lors de différentes expérimentations.

Avant d'aborder la formalisation logique et afin de positionner mon ap-

proche, je vais présenter dans le prochain chapitre les principaux paradigmes d'apprentissage et mettre en évidence leurs limites dans ce contexte.

Ce document est donc structuré en deux parties : la première regroupe les aspects théoriques de mes travaux concernant l'étude des paradigmes d'apprentissage (Chapitre 2), la définition logique d'un ensemble de jugements scientifiques (Chapitre 3) et la synthèse de ces travaux en une formalisation générale du contexte de découverte (Chapitre 4). La seconde partie de ce document regroupe les aspects pratiques concernant le prototype *Éleusis+Nobel* : les inspirations et la définition informelle du projet (Chapitre 5), sa modélisation et son implantation (Chapitre 6), et enfin la présentation de résultats encourageants obtenus après deux campagnes d'expérimentation (Chapitre 7).

Chacun des chapitres présentés dans ce document a été publié sous la forme d'acte de conférence ou d'article de revue. La liste de ces publications est disponible en Annexe A, et certaines d'entre elles sont rapportées en Annexe D afin de montrer l'évolution des formalisations.



Première partie  
Aspects Théoriques



## Chapitre 2

# Vers un Apprentissage Interactif

*A force de sacrifier l'essentiel à l'urgent, on oublie l'urgence de l'essentiel*

Edgar Morin.

LA DÉCOUVERTE des processus impliqués dans l'apprentissage, qui font aujourd'hui encore l'objet d'études intensives en psychologie cognitive, ont des répercussions importantes dans le domaine de l'Intelligence Artificielle. En effet, devant la difficulté croissante des problèmes abordés par ce domaine de l'informatique (planification, contrôle normatif de commandes, fouille de données...), de nombreux travaux ont été menés dans le but de doter une machine de capacités d'apprentissage afin de lui permettre de construire une vue abstraite de son *environnement* au travers des données qu'elle rencontre et s'adapter aux variations de celui-ci, et permettre enfin de lui « enseigner » une tâche au lieu de la programmer.

Pour la psychologie inspirée du behaviorisme, l'apprentissage est vu comme la mise en relation entre un événement provoqué par l'extérieur (stimulus) et une réaction adéquate du sujet causant un changement de comportement qui est persistant, mesurable et spécifique, et permet à l'individu de formuler une nouvelle construction mentale ou de réviser une construction mentale préalable. Les approches connectionistes, par exemple, modélisent des réseaux de neurones artificiels et concrétisent directement l'impact des stimulus sur les comportements de l'apprenti en permettant au réseau de s'auto-organiser pour favoriser les comportements les mieux adaptés. Le processus d'apprentissage est donc ici un processus de modification des connections formant le réseau de neurones artificiel, et résulte en un système de type boîte noire dont seules les entrées et les sorties sont contrôlables. Il est donc très difficile de superviser de tels systèmes autrement qu'en les mettant dans des situations susceptibles de les faire évoluer.

Même si des travaux récents offrent des perspectives intéressantes pour la

supervision de tels systèmes, comme nous le verrons en section 2.4, l'apprentissage symbolique semble mieux convenir à l'interaction nécessaire entre un chercheur et son assistant.

En effet la vision constructiviste, dont l'apprentissage symbolique s'inspire, suppose que les connaissances de chaque sujet ne sont pas une simple « copie » de la réalité, mais une « (re)construction » de celle-ci, et s'attache à étudier les mécanismes et processus permettant la construction de la réalité chez les sujets à partir d'éléments déjà intégrés. L'apprentissage symbolique modélise des structures plus ou moins élaborées permettant à l'apprenti de construire des représentations abstraites de son environnement, d'extraire des régularités d'une base de connaissances, ou de représenter ces connaissances de façon plus concise. De telles représentations abstraites peuvent être assujetties à une certaine logique afin de construire des normes.

Parmi les nombreuses notions définies par ce domaine de l'intelligence artificielle, il me semble important de distinguer les cadres formels d'apprentissage des méthodes d'apprentissage elles-mêmes. Les cadres formels servent à définir de façon conceptuelle le processus d'apprentissage et à déterminer le critère de réussite attendu pour celui-ci afin d'obtenir des résultats théoriques concernant l'apprentissage de différentes classes de « langages ». Les diverses méthodes d'apprentissage automatique sont autant d'alternatives algorithmiques permettant de mettre en oeuvre ce processus dans un contexte donné. Ces méthodes<sup>4</sup> peuvent ainsi être plus ou moins adaptées au type de données d'apprentissage considéré (données numériques ou symboliques, données bruitées ou exactes, structurées ou non, ...), cependant elles ont toutes un point commun : elles sont biaisées par le modèle de description utilisé pour représenter ces données. Dans la lignée des travaux de [Nobrega *et al.*, 2003], nous souhaitons concevoir des machines capables d'assister des chercheurs dans leur démarche scientifique. Cette aide peut être fournie soit via l'apprentissage de règles intéressantes qui auraient échappées au chercheur, soit par l'apprentissage des biais de ce dernier et leur restitution à ce dernier, tel un *miroir rationnel*. La découverte de ces biais de modélisation pourra aider le chercheur à mieux décrire son problème lorsque ces biais s'avèrent involontaires et inadaptés. Le cadre formel est donc primordial pour nous, puisqu'il détermine le protocole grâce auquel l'assistant acquiert les données d'apprentissage. Ce protocole doit donc être adapté à la fois à l'apprentissage automatique et à l'apprentissage humain, et doit respecter une certaine démarche scientifique : permettre une attitude active de l'apprenti en laissant la possibilité de mettre au point des expériences afin de tester les hypothèses de travail.

---

4. Le lecteur intéressé peut se référer à [Mitchell, 1997] et [Nilsson, 1996] pour une présentation détaillée de ces méthodes.

Ce chapitre décrit ainsi l'évolution des paradigmes (ou cadres formels) d'apprentissage automatique du point de vue des hypothèses faites sur l'apprenti et sur son rôle durant le processus d'apprentissage, tout en gardant un regard critique vis-à-vis du contexte de découverte. Comment la cible de l'apprentissage est-elle définie ? Quelles sont les ressources de l'apprenti ? L'apprentissage doit-il être exact ou simplement approximatif ?

Nous pouvons constater que ces hypothèses, outre leur adaptation au contexte historique, ont évolué en parallèle des changements didactiques et pédagogiques cités en Section 1.1.4. Ainsi, dans la vision de Gold (Section 2.1), l'apprenti joue un rôle passif et l'apprentissage peut être considéré comme non-supervisé, c'est à dire que l'apprenti est seul face à son problème. Dans la vision de Vailant (Section 2.2), l'apprenti reste passif mais reçoit des informations positives et négatives selon une certaine probabilité. La supervision de l'apprentissage par un expert est ainsi rendue possible par le biais de ces exemples. Enfin, avec l'apprentissage à partir de requêtes développé par Angluin (Section 2.3), l'apprenti devient actif et sélectionne lui-même l'information qu'il désire vérifier. Le paradigme d'apprentissage par imitation qu'elle introduit (Section 2.4) nous rapproche encore davantage d'un modèle cognitif humain, et nous semble donc particulièrement adapté à l'interaction homme-machine que nous souhaitons formaliser.

## 2.1 L'Identification à la Limite (1967)

Motivé par le problème de l'acquisition du langage humain, [Gold, 1967] a introduit le premier modèle formel d'apprentissage. Ce modèle, connu sous le nom d'*identification à la limite*, décrit les conditions de convergence d'une procédure infinie d'identification d'une grammaire permettant d'engendrer un langage à apprendre. Cette section décrit ces conditions.

On suppose qu'un enfant n'entend lors de son apprentissage que des phrases bien construites, puisque les adultes qui composent son entourage sont supposés parler correctement. Il n'a donc à sa disposition que des exemples positifs des phrases pouvant être formées grâce à la grammaire à apprendre. Chaque phrase est susceptible de le conforter dans son hypothèse ou de le faire changer d'avis ; de plus, l'apprenti peut entendre plusieurs fois les même phrases au cours de son apprentissage, et pourrait théoriquement être amené à les entendre toutes s'il vivait suffisamment longtemps.

Gold a donc considéré que toute énumération exhaustive des phrases correctes d'une langue ou d'un langage définit un environnement admissible pour l'apprentissage de la grammaire correspondante, et il donne à cette énumération le nom de *Texte*. Un *Texte* est donc une énumération infinie des objets

pouvant être générés à partir de la grammaire cible. Cette énumération peut être redondante mais doit être exhaustive. L'Apprenti découvre une à une les phrases du Texte, et doit émettre une hypothèse à chaque fois. La suite d'hypothèses formulée par l'apprenti est ainsi infinie, et la procédure d'apprentissage peut être considérée comme un algorithme effectuant une boucle infinie sur un flux d'entrée -les phrases du Texte-.

La première condition d'une identification efficace est qu'il existe un temps fini au bout duquel l'apprenti ne change d'hypothèse qu'un nombre fini de fois, même si le texte énuméré est infini. En d'autres termes, la suite infinie d'hypothèses formulées par l'apprenti doit converger afin qu'on puisse juger du résultat, l'apprenti ne doit pas changer d'hypothèse indéfiniment. L'identification est dite correcte si l'hypothèse vers laquelle la suite converge est équivalente au langage cible.

Cette notion de Texte est centrale dans le paradigme d'identification à la limite. Cependant, deux enfants placés dans des environnements différents n'entendent pas les mêmes phrases dans le même ordre, ce qui ne perturbe pas pour autant l'apprentissage. Gold impose donc une contrainte supplémentaire à l'apprentissage qui doit être exact quel que soit le *Texte* (ou plus précisément, quel que soit l'ordre de présentation du *Texte*). Par exemple, si la grammaire à identifier est celle permettant de générer l'ensemble (infini) des nombres premiers, n'importe quelle énumération de l'ensemble des nombres premiers constitue un Texte valide, et l'identification doit être correcte quel que soit l'ordre des nombres présentés à l'apprenti. La tâche peut toutefois sembler plus simple lorsque les nombres sont donnés dans l'ordre croissant (1, 3, 5, 7, 11, 13, ...), puisque seuls les quelques premiers sont connus et identifiables par la plupart d'entre nous.

Il est également supposé que l'apprenti connaît un ensemble de grammaires parmi lesquelles il choisit ses hypothèses. Deux hypothèses différentes peuvent représenter le même langage, mais être exprimées de façons différentes. Ainsi, la complexité du langage est fixée par celle des grammaires qui le représentent (ce qui n'est pas sans rappeler la profondeur logique de Bennett [Bennett, 1988] (voir [Beurier, 2007] et Annexe B pour une synthèse des différentes mesures possibles de la complexité).

Enfin, l'identification à la limite correspond à un apprentissage exact : un apprenti identifiant un langage à la limite retournera une hypothèse exactement équivalente au langage cible, mais seulement après avoir pris connaissance d'un nombre fini, et potentiellement très grand, de phrases du texte. Avant d'avoir identifié ce langage, son hypothèse peut en être très éloignée et varie généralement à chaque nouvelle phrase démontrant sa fausseté.

**Définition 2.1 (Identification à la limite)** *Une classe de langages  $\mathcal{L}$  est*

identifiable à la limite *s'il existe un algorithme  $\sigma$  tel que quel que soit le langage  $L \in \mathcal{L}$ , et quel que soit le Texte  $t$  présentant  $L$ ,  $\sigma$  propose une grammaire  $G$  qui engendre précisément  $L$  en réponse à presque tous les segments initiaux de  $t$  (i.e.  $\sigma(t(i)) = G$  pour tout  $i \geq n$ , pour un certain entier  $n$ ).*

Le critère d'apprentissage est donc la convergence de la suite d'hypothèses vers une hypothèse correcte : à partir d'un nombre fini d'étapes, la grammaire proposée doit être identique et équivalente à celle du langage à identifier, et ne doit plus varier. Cette représentation du processus d'apprentissage est tout à fait pertinente, et ce paradigme, illustré au Chapitre 4 sur un jeu de cartes, a influencé de nombreux travaux. Il est caractérisé principalement par le fait que :

- l'apprentissage doit être exact, alors que l'ensemble des phrases du texte peut être infini. En particulier, les *phrases* qui permettraient à l'apprenti de changer d'hypothèse peuvent ne lui être présentées qu'au bout d'un temps très long.
- même si l'apprenti identifie correctement la classe de langages cible, il ne « saura » jamais que son hypothèse est correcte. Il n'y a donc pas de critère d'arrêt permettant de garantir que le processus se terminera.

Ce paradigme semble donc adapté au contexte de la découverte scientifique dans lequel nous nous devons d'admettre que même si une théorie semble correcte, elle peut être remise en cause par une observation inédite ou la prise en compte de nouveaux paramètres (voir Chapitre 1). La découverte scientifique serait donc un processus d'identification à la limite des lois de la nature. Cependant, le processus par lequel le Texte est présenté à l'apprenti définit un rôle passif pour celui-ci, ce qui n'est pas représentatif du comportement scientifique. En effet selon ce mode d'interaction, on peut se représenter l'apprenti essayant de comprendre le monde qui l'entoure en observant les phénomènes auxquels il assiste passivement, un peu comme s'il se tenait sur un tapis roulant le guidant à travers le monde et l'emmenant pas à pas vers des situations plus ou moins représentatives de ce qu'il doit apprendre. Ainsi, même lorsqu'il découvre une situation particulièrement intéressante, il ne peut pas descendre de son tapis et doit se contenter d'espérer qu'un jour peut-être, il sera confronté à une situation similaire.

De plus, quel que soit l'ordre du Texte, celui-ci doit contenir l'ensemble des phrases bien formées du langage, ce qui pose une forte contrainte sur les problèmes pouvant être modélisés dans ce cadre formel. En particulier, cette contrainte est trop forte dans le contexte de la découverte scientifique qui nous intéresse : il est peu probable qu'un chercheur soit capable de fournir à son assistant rationnel la totalité des énoncés valides dans sa théorie. Cependant, l'assistant pourrait toutefois apprendre en utilisant les énoncés déjà rencontrés

par son utilisateur.

Le paradigme présenté dans la section suivante introduit la possibilité de superviser un processus d'apprentissage. Ce paradigme fournit des outils probabilistes pour évaluer l'apprentissage, et a eu à ce titre un fort impact sur la communauté *ILP*<sup>5</sup>. Cependant le rôle de l'apprenti reste passif et il faudra attendre l'introduction d'un *oracle* et l'utilisation de requêtes (voir Section 2.3) pour définir un apprentissage actif représentatif de la démarche scientifique.

## 2.2 La *PAC*-apprenabilité (1984)

Leslie Valiant, motivé par les besoins de la technologie des systèmes experts, s'est intéressé à l'apprentissage de concepts à partir d'exemples étiquetés par des machines interagissant avec des experts. Contrairement au modèle proposé par Gold, l'apprenti proposé par [Valiant, 1984] a donc désormais accès à de l'information positive et négative, c'est à dire que l'appartenance ou la non-appartenance de chaque exemple au concept cible est signalée à l'apprenti par une *étiquette* '+' ou '-'. Cette forme de présentation d'exemples positifs et négatifs a été largement utilisée en apprentissage, notamment pour construire des *espaces de versions*<sup>6</sup> ou pour superviser une forme d'apprentissage dite *par renforcement*.

L'apprenti examine ces exemples, qui sont les éléments d'un domaine, et doit deviner un concept cible qui est une partie de ce domaine. Plus précisément, il doit deviner un programme qui calcule la fonction caractéristique du concept cible. Ce programme est donc une représentation du concept tout comme la cible de l'apprentissage à la limite était une grammaire engendrant le langage et non le langage lui-même. Lorsque les domaines considérés sont dénombrables, munir les domaines d'une structure d'espace probabilisé ne pose pas de problème. Valiant emprunte donc au calcul des probabilités une conception particulière de la convergence, et il adopte la conception du réalisable proposée par la théorie de la complexité algébrique<sup>7</sup>.

L'environnement de l'apprentissage est ainsi caractérisé par la donnée d'une distribution de probabilités  $D$  sur le domaine. Cette distribution est arbitraire et inconnue de l'apprenti, et les tirages des exemples  $y$  sont assujettis et sont indépendants les uns des autres. Dans l'exemple où le concept à apprendre est l'appartenance d'un nombre à l'ensemble des nombres premiers, on présente donc un par un à l'apprenti des couples formés par :

---

5. Inductive Logic Programming

6. En anglais : Version Spaces

7. La théorie de la complexité algébrique est l'étude nombre d'opérations minimales requises pour la résolution d'un calcul algébrique.

- un nombre  $x \in \mathcal{X}$  tiré au hasard selon  $D$  ( $\mathcal{X}$  représentant alors le domaine),
- une valeur  $+$  ou  $-$  (ou encore  $1$  ou  $0$ , *vrai* ou *faux*) selon que  $x$  est premier ou non.

Les tirages étant indépendants les uns des autres, l'apprenti doit, après avoir examiné un certain nombre d'exemples, proposer en un temps fini et avec une forte probabilité une hypothèse ayant un taux d'erreur maîtrisé.

Plus formellement, le domaine  $X = \{0, 1\}^n$  représente l'ensemble de toutes les affectations possibles sur  $n$  variables (ou attributs) booléennes. On définit un concept  $c \subset X$  et une hypothèse  $h \subset X$  comme des sous ensembles de  $X$ . L'erreur associée à  $h$  par rapport à  $c$  est définie par :

$$\text{erreur}(h) = \sum_{x \in h \Delta c} D(x),$$

où  $\Delta$  représente la différence symétrique<sup>8</sup>.

**Définition 2.2 (PAC-apprenabilité)** Une classe de concepts  $\mathcal{C}$  est PAC-apprenable par l'espace d'hypothèses  $\mathcal{H}$  s'il existe un algorithme polynomial  $\sigma$  et un polynôme  $p(.,.,.)$  tel que :

- pour tout  $n \geq 1$ ,
- pour tout concept  $c \in \mathcal{C}$ ,
- pour toute distribution  $\mathcal{D}$  sur  $X$ ,
- pour tout  $\varepsilon > 0$ ,
- pour tout  $\delta < 1$ ,

si  $\sigma$  prend en entrée au moins  $p(n, \frac{1}{\varepsilon}, \frac{1}{\delta})$  exemples indépendants de  $c$  tirés aléatoirement selon  $D$ , alors  $\sigma$  propose une hypothèse  $h \in \mathcal{H}$  avec une probabilité  $p(h) \geq 1 - \delta$  et une erreur  $\text{erreur}(h) \leq \varepsilon$ .

Valiant impose donc deux contraintes à son modèle qui est désormais connu sous le nom de PAC-apprenabilité<sup>9</sup> : une approximation paramétrée de la solution et une convergence en temps polynomial.

### Approximation paramétrée

Cette première contrainte est d'ordre statistique et implique que l'apprentissage soit fiable. Au contraire du modèle de Gold, qui demande que la présentation des exemples soit complète, celui de Valiant suppose que l'apprenti ne rencontre qu'un certain nombre d'exemples. On admet donc que l'apprentissage soit approximatif. En effet, il est toujours possible qu'un exemple rare,

8. Soit  $h \Delta c$  la différence symétrique entre  $h$  et  $c$  :  $x \in h \Delta c \Rightarrow x \in h \Leftrightarrow x \notin c$

9. PAC représente le critère d'apprentissage « Probablement Approximativement Correct »

représentant peut-être un phénomène contingent, soit mal classé par un apprenant qui ne l'a pas rencontré. Le concept appris n'est donc pas identique au concept cible, mais doit en être une approximation acceptable, cette approximation tenant compte de la probabilité de présentation des exemples. Le modèle apporte donc des garanties quant à la marge d'erreur acceptable de la solution apprise et à la convergence des hypothèses : plus la taille de l'échantillon grandit, moins le taux d'erreur doit être important.

### Convergence en temps polynômial

La deuxième est algorithmique et impose que le processus d'apprentissage s'arrête, au lieu de boucler sur un flux infini. Plus précisément, la convergence vers une hypothèse au taux d'erreur acceptable doit être atteinte en un temps fini pour que l'apprentissage soit qualifié de probablement approximativement correct. Ce temps doit être fonction des exemples et du concept cible, mais doit aussi tenir compte de l'approximation tolérée. Cependant, comme le modèle de Gold nécessitait que l'apprentissage soit correct pour tous les flux d'exemples possibles, celui de Valiant impose qu'il soit acceptable quelle que soit la distribution de probabilités imposée sur les exemples.

Le modèle proposé par Valiant permet donc d'évaluer le temps nécessaire à un algorithme pour converger vers une hypothèse en fonction de l'approximation requise et du nombre d'exemples. Ce critère d'arrêt qui permet désormais d'évaluer l'apprentissage, et dans lequel on retrouve une volonté de classer les problèmes d'apprentissage comme on classe les problèmes en calculabilité et complexité, a inspiré de nombreuses recherches et analyses sur l'acquisition de la capacité à prédire différentes classes de concepts.

Pourtant, ce modèle formel est plus éloigné de l'apprentissage « naturel »<sup>10</sup> et du processus de découverte que celui de Gold, puisqu'il décrit un processus fini. En fait, tous les concepts utilisés ici sont des concepts bornés. Valiant propose de gérer un contexte de connaissances incomplètes en quantifiant l'approximation de la solution dans ce contexte. Si cela s'avère utile dans bien des cas, je présenterai au Chapitre 3 une autre façon de gérer cette difficulté fondée sur l'utilisation de jugements logiques. La loi de distribution elle-même est difficile à établir, puisqu'elle ne peut être représentative que d'un échantillon fini du domaine. De plus, cette loi n'est valable que pour un apprentissage passif puisqu'une attitude active la modifierait considérablement.

En reprenant la métaphore du tapis roulant, le modèle de Valiant décrit tout à fait un observateur se tenant sur ce même tapis, mais pour un temps fini. Les situations lui sont désormais présentées selon une loi de probabilité établie sur les bases d'un ensemble fini d'observations, et la solution qu'il doit

---

10. ou apprentissage humain

proposer lorsque le tapis s'arrête doit être une approximation acceptable de ce qu'il a observé. Il me semble qu'une démarche scientifique de formulation et de validation empirique d'hypothèses, voire même de partage et de confrontation de théories via un processus de publication et de réfutation, est plus efficace pour aboutir à la découverte d'un concept cible. L'apprentissage à partir de requêtes, présenté ci-après, traduit mieux ce caractère actif, représentatif de la démarche scientifique, que nous souhaitons attribuer à l'apprenti durant le processus d'apprentissage. Il suppose toutefois l'existence d'un oracle, ce qui pose une difficulté que nous tâcherons de contourner au Chapitre 4.

## 2.3 L'Apprentissage à Partir de Requêtes (1988)

Cette approche théorique a été proposée à peu près en même temps que l'apprentissage *PAC* dans des travaux tels que [Shapiro, 1981] ou encore [Sammut & Banerji, 1986], mais Dana Angluin fût la première à proposer un cadre formel d'étude pour celui-ci [Angluin, 1988]. Ce modèle introduit une nouvelle source d'information pour l'apprenti : un *oracle* capable de répondre à différents types de requêtes formulées par l'apprenti. Au lieu de percevoir des exemples étiquetés qui lui sont fournis arbitrairement, l'apprenti peut donc désormais sélectionner lui-même l'information à laquelle il accède en questionnant l'oracle, et devient alors un acteur de son propre apprentissage. L'utilisation d'un tel cadre formel, faisant intervenir un apprenti pro-actif, peut être justifiée par le besoin pour l'apprenti d'être capable d'orienter ses observations en fonction des incertitudes de ses hypothèses, ce qui n'est pas possible dans le cadre *PAC* ou celui de l'identification à la limite, où il ne peut pas discriminer des perceptions ambiguës ou conflictuelles (ou peut-être que l'information discriminante ne lui sera présentée qu'au bout d'un temps très long).

L'introduction d'un *oracle* fait intervenir un nouveau paramètre dans l'évaluation de l'apprentissage et dans le calcul de sa complexité, puisqu'il faut désormais tenir compte de la complexité des tâches accomplies par cet oracle pour répondre aux requêtes.

Les deux types de requêtes qui se distinguent le plus l'un de l'autre, et les plus couramment utilisés, sont :

- l'appartenance d'un élément du domaine au concept à identifier (voir définition 2.3), et
- l'équivalence entre deux concepts (voir définition 2.4).

Ces deux types de requêtes sont représentés de façon ensembliste comme suit :

**Définition 2.3 (Requêtes d'Appartenance)** *Soit  $c$  le concept cible appartenant à une classe de concepts  $C$  et  $x$  un élément du domaine, une requête*

d'appartenance  $MQ(x)$  retourne « vrai » si  $x \in c$  (exemple positif) ou « faux » si  $x \notin c$  (exemple négatif).

**Définition 2.4 (Requêtes d'Équivalence)** Soient  $c$ , le concept cible, et  $c'$ , un concept quelconque, deux sous-ensembles d'un ensemble universel  $\mathcal{U}$ , une requête d'équivalence  $EQ(c')$  retourne « vrai » si  $c' \simeq c$  (hypothèse correcte), et « faux » si  $c' \neq c$  (hypothèse erronée). Dans ce dernier cas, un contre exemple  $x'$  appartenant à la différence symétrique entre  $c$  et  $c'$  est fourni à l'apprenti.

La requête la plus simple à traiter pour l'oracle est la requête d'appartenance (notée  $MQ$  pour Membership Query), puisqu'elle correspond à une simple vérification de la validité d'une instance donnée. Elle permet ainsi l'étiquetage des exemples (voir Section 2.2), et une stratégie d'apprentissage n'utilisant que ce type de requête s'appuierait éventuellement sur de la dichotomie en choisissant les exemples dont l'étiquetage permettrait l'élimination d'un maximum d'hypothèses. En se plaçant dans le cas où le domaine  $D$  des exemples est fini, une stratégie naïve parcourant  $D$  de façon brutale pour étiqueter les exemples reviendrait à utiliser le protocole  $PAC$  sans tenir compte de la loi de distribution (les exemples étant étiquetés un à un selon un ordre arbitraire) et utiliserait dans le pire des cas  $|D| - 1$  requêtes d'appartenance.

Au contraire, la requête d'équivalence (notée  $EQ$  pour Equivalence Query) appartient à une classe de complexité supérieure, puisqu'elle correspond à la preuve d'une théorie, à la vérification de la validité de l'hypothèse dont la requête fait l'objet sur l'ensemble des instances du domaine. Si l'ensemble  $H$  des hypothèses est fini, une stratégie naïve n'utilisant que des requêtes d'équivalence proposerait de façon exhaustive chaque hypothèse de  $H$ , utilisant ainsi dans le pire des cas  $|H| - 1$  requêtes d'équivalence, ce qui est parfois la meilleure stratégie que l'on puisse établir, en particulier lorsque l'ensemble des hypothèses est petit.

La différence apparente de complexité entre ces deux types de requêtes peut cependant être ignorée lorsqu'on suppose, comme c'est le cas dans [Angluin, 1988], que les hypothèses sont représentées par des indices et n'ont pas besoin d'être interprétées par l'oracle. Une  $EQ$  peut alors être résolue en comparant l'indice de l'hypothèse dont la requête fait l'objet avec l'indice de l'hypothèse cible connu par l'oracle.

D'autres types de requêtes correspondant aux relations ensemblistes peuvent être définis et sont moins catégoriquement opposés les uns des autres. On notera ainsi l'existence des requêtes d'inclusion ( $c' \subset c?$  : « y'a-t-il inclusion d'un concept dans un autre? »), de jointure ( $c' \cap c = \emptyset?$  : test d'intersection entre deux concepts) ou encore d'exhaustivité ( $c' \cup c = \mathcal{U}?$  : test d'union entre deux concepts), un contre exemple étant fourni pour chaque réponse négative.

Le critère d'apprentissage considéré est généralement l'identification à la limite, mais peut parfois être adapté à un apprentissage de type *PAC*. La principale contrainte porte en effet sur le nombre et la taille des requêtes qui doivent être polynomiaux, ainsi que sur l'existence d'un oracle capable de les traiter. Cette formalisation de l'apprentissage automatique constitue donc un outil supplémentaire d'étude et de caractérisation pour l'identification à la limite ainsi que pour la *PAC*-apprenabilité, et nous verrons dans la prochaine section que le rôle actif de l'apprenti représente également un premier pas vers un changement profond des paradigmes d'apprentissage automatique.

Toutefois, si une partie de la complexité d'un problème d'apprentissage peut être transférée sur un oracle, il faut rester prudent lors la mise en place concrète d'un tel paradigme, puisqu'une classe de concepts peut ainsi n'être apprenable qu'à la condition que l'oracle nécessaire au traitement des requêtes de l'apprenti existe et ne soit pas trop complexe.

En effet, on peut d'ores et déjà remarquer la difficulté d'appliquer ce paradigme au contexte de la découverte qui nous intéresse. Si on cherche à faire apprendre une solution connue à l'apprenti, alors on doit pouvoir supposer qu'un oracle peut être créé pour répondre à ses requêtes. Ceci est valable pour l'apprentissage machine, mais également pour l'apprentissage humain : les professeurs sont en général capable de répondre aux questions de leurs élèves. Toutefois, il semble raisonnable de considérer qu'il n'existe dans la nature aucune entité capable de répondre aux requêtes d'un chercheur (même si l'existence d'une telle entité, Dieu par exemple, n'a pas encore été réfutée). Ainsi, un chercheur ne demande pas à un oracle ce qui se passe exactement lorsqu'il lâche une pomme, il met au point un protocole expérimental lui permettant de simuler et d'analyser ce phénomène. Il en tirera des conclusions, mais celles-ci seront biaisées par les variables qu'il observe et par ses outils de mesure. De la même façon, il ne demande pas à un oracle de lui confirmer que sa théorie de la gravitation universelle est correcte, il la soumet au jugement de ses pairs. Ainsi, on ne peut supposer que les exemples et leurs descriptions sont finis, ce qui entraîne deux limites au formalisme proposé par Angluin : si l'ensemble des exemples est infini, l'oracle ne pourra pas traiter les requêtes d'équivalence, et si la description des exemples est elle aussi infinie (ce qu'il faut supposer dans un tel contexte puisque chaque objet étudié peut avoir une représentation infinie en fonction de l'échelle d'observation choisie), l'oracle ne pourra pas non plus traiter les requêtes d'appartenance. Je reviendrai sur ce problème au Chapitre 4.1 en tâchant d'y apporter une solution acceptable.

## 2.4 De l'Apprentissage par Simulation à l'Apprentissage par Imitation (Learning from different teachers)

Je n'exploiterai pas directement la formalisation rapportée ici, mais elle a inspiré certaines extensions présentées en 4. De plus, ces travaux novateurs méritent à mes yeux de figurer dans ce chapitre car ils ne manqueront pas d'influencer la communauté apprentissage. En particulier, il me semble que cette approche pourrait permettre de concilier des approches apparemment opposées de l'apprentissage, telles que les approches connectivistes et symboliques (brièvement discutées en introduction).

Les modèles formels d'apprentissage présentés dans la première partie de ce chapitre peuvent se décrire comme la transmission d'un programme représentant le concept à apprendre entre un tuteur et un apprenti, tous deux modélisés par des machines. La question de l'apprentissage peut alors se transformer en une question de simulation. En effet, si l'apprenti est théoriquement capable de simuler le fonctionnement du tuteur, alors l'apprentissage se limite à la transmission du programme à simuler du tuteur à l'apprenti, soit directement, soit indirectement sous la forme d'exemples.

Le point de départ des travaux présentés dans cette section est la constatation que l'apprentissage humain s'effectue avec des agents complexes, différents les uns des autres, aux capacités de modélisation limitées, et ayant une connaissance imparfaite d'eux-mêmes et des autres. De telles contraintes sur les agents - qui ne sont pas sans évoquer l'introduction par Simon de la notion de *rationalité limitée* dans la théorie économique - obligent à un changement de paradigme, la simulation devenant inaccessible aux agents dès que ceux-ci sont confrontés à l'ignorance de leur fonctionnement propre.

[Angluin & Krikis, 2003] tourne alors le dos aux cadres formels d'apprentissage se fondant sur la capacité à simuler et propose un formalisme intégrant un tuteur et un apprenti et fondé sur l'interaction entre ces deux agents. Il s'agit dès lors d'un apprentissage par imitation à la limite.

Pour illustrer ce « changement » de point de vue sur l'apprentissage, considérons l'exemple du jonglage en admettant que n'importe quelle personne sachant jongler constitue un tuteur (ou un modèle) acceptable. Une personne apprenant à jongler n'a qu'une connaissance limitée de la physique expliquant les mouvements des balles et de son corps, et une connaissance encore plus réduite de l'implantation neuronale des capacités de proprioception dans son cerveau ou dans celui du tuteur. Sans faire de théorie physique ni de copie de l'organisation des aires cérébrales impliquées dans la proprioception, il n'en est pas moins possible d'apprendre à jongler en cherchant à imiter le tuteur dans

ses mouvements.

Pourtant, la vue d'un oiseau ne pousse personne à se jeter par la fenêtre en battant des bras comme l'oiseau bat des ailes pour une raison évidente : les protagonistes n'ont cette fois pas les mêmes caractéristiques morphologiques. Cette capacité d'imitation est ainsi étroitement liée au fait que les fonctionnalités du modèle choisi ne semblent pas différentes ou incommensurablement supérieures à celles de l'apprenti. Ainsi, lorsque je vois quelqu'un jongler, cette tâche peut me paraître ardue, mais pas insurmontable puisqu'étant tous les deux des Hommes, le jongleur et moi avons à priori les mêmes pré-dispositions au jonglage. Toutefois, avant de jongler avec cinq balles, comme pourrait le faire le jongleur, je dois apprendre à jongler avec trois balles, ce qui peut-être considéré comme une tâche facile par le jongleur. Je pourrais ainsi considérer avoir atteint le niveau du jongleur pour cette tâche (jongler avec trois balles) lorsque sa réalisation me semblera aussi facile qu'à lui-même.

[Angluin & Krikis, 2003] formalise cette notion de difficulté relative entre une tâche effectuée par le tuteur et une tâche supposée équivalente mais effectuée par l'apprenti, et introduit une nouvelle requête de « complexité ». Comme nous allons le voir, cette requête étend celle d'appartenance : en plus de pouvoir demander si un concept appartient à la classe cible, l'apprenti peut désormais interroger son modèle à propos de la difficulté de décider si un concept appartient à la classe cible. On peut donc intuitivement voir une requête d'appartenance comme une vérification d'une instance, une requête d'équivalence comme une vérification pour chacune des instances du domaine, et une requête de « complexité » comme une estimation du temps nécessaire au tuteur pour répondre à une requête d'appartenance.

La première partie de cette section décrit comment [Angluin & Krikis, 2003] étend une formalisation standard d'un apprenti proposée par [Blum & Blum, 1975], afin de tenir compte du fait que l'apprenti n'a pas une connaissance parfaite des processus à sa disposition pour effectuer la tâche requise, par exemple le calcul d'une fonction. La seconde partie est consacrée au tuteur et à la prise en compte de son rôle dans le processus d'apprentissage. Enfin, j'illustrerai ce type d'apprentissage sur le jeu d'échecs. Une autre illustration, énoncée dans notre formalisme, est donnée au Chapitre 4 en tant que perspective.

### 2.4.1 Formalisation de l'Apprenti

Plusieurs définitions issues de [Blum & Blum, 1975] sont à rappeler afin de présenter les modifications proposées par [Angluin & Krikis, 2003].

### Cible de l'apprentissage :

La cible de l'apprentissage est une fonction  $f \in F$  telle que :

- $F$  est l'ensemble des fonctions partiellement définies sur l'ensemble  $\mathcal{N} = \{0, 1, 2, 3, \dots\}$  des entiers naturels,
- le domaine  $D_f \subset \mathcal{N}$  de  $f$  est le sous-ensemble de  $\mathcal{N}$  sur lequel  $f$  est définie,
- le graphe  $G_f$  de  $f$  est l'ensemble des paires  $(x, f(x))$  telles que  $x \in D_f$ ,
- soit  $G_g$  le graphe d'une fonction  $g$ , on dit que  $f$  étend  $g$  si on a  $G_g \subseteq G_f$ , et enfin
- la composition de  $f$  et  $g$  est une fonction telle que  $h(x) = f(g(x))$  si  $g(x)$  et  $f(g(x))$  sont définies,  $h(x)$  étant indéfinie dans le cas contraire.

Notons que les points  $e \in G_f$  du graphe d'une fonction peuvent servir d'exemples pour l'apprentissage de cette fonction, et que les graphes de fonctions peuvent être utilisés pour caractériser l'apprentissage (soit  $g$  la cible de l'apprentissage, celui-ci est exact si  $G_f = G_g$ , trop général si  $G_g \subseteq G_f$  ou trop spécifique si  $G_f \subseteq G_g$ ).

### Listing standard :

On se donne la liste particulière de toutes les machines de Turing, notée  $t_0, t_1, t_i, \dots$ , dans laquelle chaque machine calcule une fonction partielle  $f : \mathcal{N} \rightarrow \mathcal{N}$  et telle que :

- $f$  n'est pas définie sur les entrées pour lesquelles la machine de Turing correspondante ne s'arrête pas,
- les *fonctions récursives partielles* sont les fonctions partielles de  $\mathcal{N}$  dans  $\mathcal{N}$  calculables par une machine de Turing,
- les *fonctions récursives totales* sont les fonctions calculables par une machine de Turing pour toute entrée.

### Système de Programmation :

Un *système de programmation* est défini comme une suite  $\phi_0, \phi_1, \phi_i, \dots, \phi_n$  de fonctions récursives partielles contenant toutes les fonctions récursives partielles. La fonction universelle pour ce système de programmation,  $u(i, x) = \phi_i(x)$ , peut-être vue comme une fonction de commande permettant de lancer une machine particulière sur une entrée particulière. Si  $u$  est une fonction récursive partielle (*i.e.*, calculable par une machine de Turing), le système de programmation est dit *universel*.

**Système de programmation universel acceptable :**

Une fonction de composition  $c$  est une fonction totale telle que pour tout  $i$  et tout  $j$ ,  $i, j, k \in \mathcal{N}$ , si  $k = c(i, j)$ , alors  $\phi_k$  calcule la même fonction que  $\phi_i \circ \phi_j$ . Un système de programmation universel est dit *acceptable* s'il a une fonction de composition récursive, ce qui est supposé être le cas pour le listing standard.

[Machtey & Young, 1978] ont démontré que tous les systèmes de programmation acceptables ainsi définis se simulent mutuellement et sont en ce sens équivalents :

**Lemme 2.5** *Supposons que  $\phi$  est un système de programmation universel et  $\psi$  un système de programmation acceptable. Il existe alors une fonction de translation récursive  $r$  telle que pour tout  $i \in \mathcal{N}$ ,  $\phi_i = \psi_{r(i)}$ .*

**Une mesure de complexité adaptée :**

Étant donné un système de programmation acceptable  $\phi_0, \phi_1, \phi_2, \dots$ , une séquence  $\Phi_0, \Phi_1, \Phi_2, \dots$  est une *mesure de complexité calculatoire* si :

1. quel que soit  $i$ ,  $\phi_i$  et  $\Phi_i$  ont le même domaine :  $\forall i, \forall x$ ,  $\phi_i(x)$  est définie ssi  $\Phi_i(x)$  est définie,
2.  $\Phi_i(x) < y$  est décidable, i.e., «  $\Phi_i(x) < y$  » est un prédicat récursif total de  $i$ ,  $x$ , et  $y$ .

Par exemple, on peut choisir comme mesure de complexité pour le listing standard  $t_0, t_1, t_2, \dots$ , les fonctions  $T_0, T_1, T_2, \dots$ , où  $T_i(x)$  est le nombre de pas de calcul effectués par la  $i^{\text{ème}}$  machine de Turing pour traiter l'entrée  $x$ , ou reste indéfini si le calcul ne s'arrête pas. Chaque fonction récursive partielle pouvant être calculée par une machine de Turing, la séquence  $t_0, t_1, t_2, \dots$  est un système de programmation. La fonction universelle  $u(i, x)$  de ce système peut être calculée en simulant le fonctionnement de la  $i^{\text{ème}}$  machine de Turing  $t_i$  sur l'entrée  $x$ , ce qui rend le système de programmation universel. Une fonction de composition  $c(i, j) = k$  peut être calculée en simulant le fonctionnement de la machine  $t_k$  qui combine  $t_i$  et  $t_j$  en lançant  $t_i$  puis  $t_j$  sur le résultat de  $t_i$  si celle-ci s'arrête. Ce système de programmation est donc acceptable.  $T_i$  est récursive partielle, et possède le même domaine que  $t_i$  (condition 1), puisque l'on peut calculer  $T_i(x)$  en simulant  $t_i$  sur l'entrée  $x$ , et en comptant le nombre d'itérations effectuées par  $t_i$ . Si  $t_i$  ne s'arrête pas pour l'entrée  $x$ , alors la simulation non plus et  $T_i(x)$  est indéfinie, sinon le résultat de la simulation est le nombre total d'itérations effectuées par  $t_i$  avant de s'arrêter. De la même façon, étant donnés trois entiers  $i, x$ , et  $y$ , on peut décider si  $T_i(x) \leq y$  en simulant  $t_i$  sur  $y$  itérations. Si  $t_i$  s'arrête en un nombre d'itérations inférieur ou égal à  $y$ , alors le prédicat est vrai, sinon il est faux (condition 2).

[Machtey & Young, 1978] étend la correspondance entre deux systèmes de programmation acceptables à leur mesure de complexité :

**Lemme 2.6** *Si  $\phi$  et  $\psi$  sont des systèmes de programmation acceptables ayant  $\Phi$  et  $\Psi$  pour mesures de complexité respectives, et si  $r$  est une transcription récursive de  $\phi$  en  $\psi$ , alors il existe une fonction récursive  $b$  telle que pour tout  $i$  et pour tout  $x$  excepté pour un nombre fini de  $x$ ,  $\Psi_{r(i)}(x) \leq b(x, \Phi_i(x))$ .*

Il est ainsi prouvé qu'il existe une fonction récursive liant la complexité de deux programmes lorsque l'un est la transcription de l'autre dans un système de programmation équivalent : cette propriété forte est à l'origine des requêtes de « complexité » utilisées par Angluin pour limiter la complexité du programme simulé (hypothèse) en fonction de celle du programme original (calculant une fonction étendant la fonction cible). Cependant, le modèle de [Blum & Blum, 1975] suppose que l'apprenti connaît parfaitement le système de programmation dont il dispose pour énoncer ses conjectures : l'apprenti est une machine de Turing  $M$ , la cible de l'apprentissage est une fonction récursive partielle  $f$ , et les exemples sont les paires appartenant au graphe de  $f$ .  $M$  dispose d'une bande d'entrée accessible en lecture seule, supposée contenir une présentation<sup>11</sup> de  $f$ , et d'une bande de sortie accessible en écriture utilisée pour émettre ses conjectures concernant  $f$ . Chaque conjecture est un index  $j$  permettant d'identifier une fonction  $\psi_j$  dans son système de programmation  $\Psi_0, \Psi_1, \Psi_2, \dots$ . Le critère d'apprentissage étant l'identification à la limite (voir Section 2.1), on dit que  $M$  identifie une classe de fonctions récursives partielles  $\mathcal{C}$  si pour toute fonction  $f \in \mathcal{C}$  et toute présentation  $T$  de  $f$ , la séquence d'hypothèses proposées par  $M$  converge correctement vers une hypothèse  $j$  (*i.e.*,  $\psi_j$  étend  $f$ ).

Illustrons maintenant les limites de ce formalisme. On peut considérer que la liste de machines de Turing caractérisant un apprenti représente les opérations purement cognitives de celui-ci, alors que le système de programmation associé représente le répertoire des actions qu'il est capable de réaliser (la « pensée » opposée à l'« action », en quelque sorte). Ce modèle suppose alors qu'il n'est pas nécessaire pour l'apprenti d'apprendre à réaliser effectivement la tâche cible, il lui suffit de s'imaginer cette réalisation, *i.e.*, d'identifier une machine de Turing capable de simuler le programme cible (et simulant donc le programme calculant  $f$ ). Dans l'exemple du jonglage, on se contenterait ainsi de comprendre comment le jongleur s'y prend pour manipuler les balles en considérant qu'il n'est pas nécessaire de s'essayer à la tâche. Dans le cas de l'acquisition d'un langage, on peut distinguer le fait d'apprendre un vocabulaire

---

11. une présentation de  $f$  est une séquence infinie d'éléments pouvant être soit une paire  $(x, y) \in G_f$ , soit le caractère \* utilisé lorsque  $f$  n'est pas définie pour  $x$ .

et des règles grammaticales d'une part, et le fait d'apprendre à parler effectivement la langue d'autre part : prononcer, articuler, et accentuer correctement sont des tâches liées à l'action. Dans le cadre de la découverte scientifique, se limiter à la composante cognitive reviendrait à énoncer des théories et des modèles ne pouvant être expérimentés, donc potentiellement infalsifiables -ce sont les hypothèses transcendantales critiquées par Kant (voir Section 1.1.2). Il semble illusoire de considérer que la compréhension des étapes d'une résolution de problème et de leur enchaînement est suffisant, il faut également s'entraîner à les réaliser de façon satisfaisante, associer un comportement effectif au modèle théorique acquis. C'est le problème d'adéquation entre une réalité concrète et une abstraction théorique formulé en Section 1.3. L'hypothèse forte de ce modèle, et dont [Angluin & Krikis, 2003] souhaite s'abstraire, est donc que l'apprenti connaît son système de programmation, et qu'il existe une correspondance directe entre l'index d'une machine de Turing et la fonction qu'elle permet de calculer, une correspondance *innée* entre le fait de s'imaginer la réalisation d'une tâche et sa réalisation effective, même si ces deux aspects sont en réalité étroitement liés et évoluent généralement en parallèle lors de l'apprentissage.

### Un système de programmation de type boîte noire

Afin de rendre compte de la connaissance imparfaite de son fonctionnement propre, l'apprenti n'a plus un accès direct à son système de programmation. En remplacement, un système de programmation inconnu  $\phi$  ainsi que la mesure de complexité correspondant  $\Phi$  sont regroupés sous le terme de *boîte de fonctions*, désormais accessible en tant qu'oracle c'est à dire comme un système de type boîte noire. Une boîte de fonctions est ainsi définie comme une fonction récursive totale  $g(i, x, s)$ , avec  $i, x$ , et  $s \in \mathcal{N} \cup \{*\}$  et tels que :

- $i$  représente l'index d'une fonction  $g_i$  : on dit qu'une boîte de fonctions *contient* les fonctions  $g_0, g_1, g_2, \dots$ . Ces fonctions, récursives partielles, représentent l'espace des hypothèses pouvant être formulées par l'apprenti.
- $x$  désigne l'entrée désirée pour cette fonction  $g_i$ ,
- $s$  est une estimation de la complexité associée à  $g_i(x)$  :  $G_i(x)$  est la plus petite valeur de  $s$  telle que  $g(i, x, s)$  soit un entier naturel, le cas échéant, ou reste indéfini dans le cas contraire. Si  $G_i(x)$  est définie, alors  $g_i(x) = g(i, x, G_i(x))$ . On dit également que  $g$  *contient* les mesures  $G_0, G_1, G_2, \dots$

En d'autres termes, si l'on se donne  $i$  et  $x$ , et que l'on examine les valeurs successives de  $g(i, x, s)$  pour  $x = 0, 1, 2, \dots$ , on observe une séquence alternant des entiers naturels et des '\*'. Si cette séquence ne contient aucun entier naturel, alors  $g_i(x)$  et  $G_i(x)$  sont tous deux indéfinis. Dans le cas contraire, il existe un entier naturel particulier  $y$  qui est le premier à apparaître dans la séquence,

et l'on définit  $g_i(x)$  comme étant  $y$ , et  $G_i(x)$  comme la valeur de  $s$  lorsque  $y$  apparaît pour la première fois.

Cette définition généralise celle de la mesure de complexité proposée par [Blum & Blum, 1975] dans la mesure où il n'est pas requis que  $g$  contienne un système de programmation. Toutefois, les notions d'*universalité* et d'*acceptabilité* peuvent être définies de façon comparable pour les boîtes de fonctions, et on retrouve la précédente définition dans le cas d'une boîte de fonctions acceptable (*i.e.*, qui contient un système de programmation acceptable). De même, [Angluin & Krikis, 2003] paraphrase le lemme 2.6 et définit une *b-relation* pouvant être interprétée comme « ne pas être  $b$  fois plus lent que » de la façon suivante : Soit  $b(x, s)$  une fonction récursive totale. Une boîte de fonctions  $g$  est *b-reliée* à une boîte de fonctions  $g'$  si pour tout  $i$ , il existe un  $j$  tel que :

1.  $g_j$  étend  $g'_i$ .
2.  $G_j(x) \leq b(x, G'_i(x))$  pour tout  $x$  appartenant au domaine de  $g'_i$ , sauf pour un nombre fini de  $x$ .

En d'autres termes, une boîte de fonctions  $g$  est *b-reliée* à une boîte de fonctions  $g'$  si pour toute fonction  $g'_i$ , et pour presque tous les éléments du domaine de  $g'_i$ , il existe une fonction  $g_j \in g$  qui ne soit pas *b fois plus rapide* que  $g'_i$ . Si la boîte de fonctions d'un jongleur lui permet par exemple de lancer les balles à une certaine hauteur d'une part, et de les rattraper d'autre part, alors être *b-relié* à elle signifie pour ma boîte de fonctions qu'elle me permet de l'imiter dans une certaine mesure, de telle sorte que ma façon de faire ne soit pas *b fois pire* que la sienne.

Par opposition à l'apprenti proposé par [Blum & Blum, 1975], qui a une connaissance *experte* du système de programmation utilisé pour exprimer ses conjectures, [Angluin & Krikis, 2003] définit un *apprenti naïf* ainsi : un *apprenti naïf* possède la même architecture qu'un *apprenti expert*, mais au lieu d'accéder *directement* à son système de programmation, il interroge une boîte de fonctions arbitraire, comme il le ferait pour un oracle (voir Section 2.3) : en plus de ses deux bandes usuelles (celle où il lit les exemples, et celle où il écrit ses hypothèses), deux bandes supplémentaires sont dédiées à la communication par requêtes avec la boîte de fonctions.

## 2.4.2 Interaction tuteur-apprenti

Un tuteur ayant connaissance du programme calculant la fonction cible est ajouté au modèle afin d'aider l'apprenti naïf en répondant à ses requêtes de complexité. L'innovation consiste à utiliser, pour des raisons d'uniformité, la même architecture pour le tuteur que pour l'apprenti. Comme une connaissance du système de programmation ne peut pas rendre l'enseignement plus difficile,

cela n'affaiblit en aucun cas leurs résultats positifs (voir [Angluin & Krikis, 2003]).

### Le tuteur :

Le tuteur  $T$  a connaissance de la fonction à apprendre et il utilise cette connaissance pour aider l'apprenti naïf. Le tuteur a la même structure que l'apprenti naïf :  $T(G')$  dispose d'une boîte noire  $G'$  et il connaît une conjecture  $g'_i$  qui étend  $f$ .

### 2.4.3 Illustration de l'Interaction Tuteur-Apprenti

Illustrons ce protocole d'interaction entre un tuteur et un apprenti naïf avec l'exemple du jeu d'échecs comme ceux que l'ont trouve dans les magazines à la page loisirs : l'échiquier est dans une certaine disposition et le joueur blanc doit trouver comment placer le joueur noir en position d'échec et mat en un certain nombre de coups.

Imaginons qu'il existe une « théorie du jeu d'échecs »  $f$  qui permette de gagner à tous les coups lorsqu'on joue à ces jeux. Cette hypothèse n'est pas très réaliste, bien que les grands joueurs doivent avoir acquis quelque chose d'équivalent, mais elle est nécessaire pour définir les exemples comme étant les différentes dispositions initiales de l'échiquier auquel l'apprenti est confronté, et pour lesquels  $f$  permet de gagner de façon certaine. Imaginons que les boîtes de fonctions  $g$  et  $g^*$ , respectivement du tuteur  $T(g)$  et de l'apprenti  $L(g')$ , contiennent un programme  $\Psi_0, \Psi_1, \Psi_2, \dots$  pour chaque mouvement possible de chaque pièce présente sur l'échiquier, et que les complexités associées  $\Psi_0, \Psi_1, \Psi_2, \dots$  soient, pour chacun d'eux, le nombre de coups joués (ou de pièces déplacées), c'est à dire 1. Ces programmes (qui n'ont pas forcément le même index dans  $g$  et  $g'$ ) peuvent être composés pour former des enchaînements de coups, et la complexité d'un programme  $\Psi_c = (\Psi_a \circ \Psi_b \circ \dots \circ \Psi_n)$  sera  $\Psi_c = (\Psi_a + \Psi_b + \dots + \Psi_n)$ . De plus, on peut caractériser la différence de niveaux entre  $L(g')$  et  $T(g)$  par une fonction  $b(x, s) = 2s$ , par exemple, ce qui signifie que  $L(g')$  a le droit de proposer des solutions deux fois moins rapides que celle de  $T(g)$ . Imaginons maintenant que le tuteur connaît l'index  $i$  correspondant à un programme réalisant une extension de  $f$ , *i.e.*, calculant pour chaque disposition initiale (sauf pour un nombre fini d'entre elles) l'enchaînement de coups nécessaire pour vaincre le joueur noir.  $L(g')$  va, pour chaque disposition  $x$  lue sur sa bande d'exemples, demander à  $T(g)$  le nombre de coups qui lui sont nécessaires pour gagner.  $T(g)$  peut ainsi faire appel à sa boîte de fonctions pour déterminer  $g(i, x, s)$  pour  $s = 1, 2, 3, \dots$  jusqu'à ce qu'il trouve un  $s$  pour lequel  $g(i, x, s) \neq *$ , *i.e.*, le nombre de coups  $G_i(x)$  nécessaires pour mettre le joueur noir échec et mat. En réponse à la requête de  $L(g')$  concernant la disposition

$x$ ,  $T(g)$  répond donc  $b(x, G_i(x))$ , c'est à dire l'approximation de la pire solution que  $L(g')$  soit autorisé à proposer en fonction de leur différence de niveaux. Grâce à cette information,  $L(g')$  peut donc éliminer toutes les hypothèses  $j$  telles que  $G'_j(x) > b(x, G_i(x))$ , *i.e.*, toutes les « théories du jeu d'échec » qui ne permettent pas de gagner dans cette situation  $x$  en un nombre de coups comparable à celui trouvé par  $T(g)$ , afin de converger éventuellement vers une hypothèse n'étant pas  $b$  fois plus mauvaise que la solution de  $T(g)$ .

#### 2.4.4 Perspectives

Plutôt que d'identifier une machine de Turing permettant de calculer une fonction cible  $f$ , l'apprenti doit désormais identifier  $f$  avant d'identifier une machine capable de la calculer. Ce processus pourrait être comparé à la phase dite de « babillage » en apprentissage par renforcement, durant laquelle l'apprenti (un nouveau né, par exemple), effectue des mouvements désordonnés jusqu'à sélectionner ceux qui lui permettent d'arriver à ses fins (contracter l'iris pour évaluer les distances, ou tendre le bras pour attraper un objet par exemple). La borne que lui fournit son tuteur sur la mesure de complexité associée à la tâche lui permet d'éliminer les combinaisons d'actions trop complexes et de sélectionner celle(s) qui correspond(ent) à cette mesure, autorisant ainsi une solution différente de celle du tuteur, mais ayant une efficacité comparable. On passe donc bien de l'ère de la simulation à celle de l'imitation, et cette formalisation pourrait être appliquée à la supervision de systèmes d'apprentissage de types boîtes noires, tels que les réseaux de neurones artificiels issus des approches connexionnistes. Il pourrait ainsi ouvrir la voie à des systèmes d'apprentissage hybrides « à deux vitesses » gérant de façon logique et symbolique le produit d'un apprentissage de plus bas niveau réalisé par des approches réputées difficiles à superviser. Une approche similaire est proposée par [Bessiere *et al.*, 2007] ou encore [Paulin, 2008]. Toutefois, des *actions* de type boîtes noires ne sont pas nécessairement le fruit d'un autre apprentissage, elles peuvent simplement représenter des événements intervenant à une échelle de description inférieure à laquelle on n'a pas (encore) accès (ou dont on souhaite masquer les détails), par exemple des phénomènes émergents.

*La tendance de la nature à constituer une société hiérarchique de lois physiques est donc bien plus qu'une thèse scientifique à discuter. C'est la raison pour laquelle nous pouvons connaître le monde. C'est elle qui rend les lois les plus fondamentales non pertinentes, et les empêche de nous tyranniser. C'est grâce à elle que nous pouvons vivre sans comprendre les secrets ultimes de l'univers.*

Robert B. Laughlin, Un Univers Différent, 2005.

L'existence d'un tuteur est toutefois soumise aux mêmes réserves que celles énoncées pour l'oracle. Cependant les informations fournies par un tuteur sont beaucoup plus simples et bien moins improbables que celles fournies par l'oracle lorsqu'on se place dans un contexte de découverte scientifique. En effet, notre référentiel spatio-temporel offre des heuristiques de choix pouvant être utilisées comme mesure de complexité : le temps et l'espace. Ainsi, un chercheur peut mesurer le temps écoulé depuis le début d'une expérience, et le considérer beaucoup trop long par rapport à la durée du phénomène simulé.

Comparer la complexité d'une solution à une autre peut également équivaloir à se demander si l'on raisonne au niveau de précision adéquat, et éventuellement entraîner la révision du modèle de description. Si l'on se place dans l'étude des systèmes complexes et des phénomènes émergents, [Laughlin, 2005] souligne qu'il existe plusieurs formes possibles pour une même fonction. [Beurier, 2007], qui propose un modèle embryo-génétique pour modéliser les phénomènes émergents dans les systèmes complexes, a fondé l'ensemble de son travail sur cette distinction entre la forme et la fonction. Le premier intérêt de la formalisation présentée dans cette section est donc d'introduire cette distinction dans le processus d'apprentissage : l'objet de l'apprentissage n'est plus la forme mais la fonction. Ceci ouvre des perspectives dans le cadre de l'apprentissage par imitation et le raisonnement par analogie dans lequel la forme pourrait jouer un rôle prépondérant. Le second intérêt de cette approche est d'avoir fondé ce modèle sur les limites cognitives de l'apprentissage et d'introduire l'utilisation d'heuristiques telles que le temps. Enfin, il pourrait être une bonne incitation à rechercher des systèmes sociaux qui réalisent de la même façon que le tuteur et l'apprenti un système d'interaction à base de requêtes, mais avec un plus grand nombre d'agents, et de plonger cette communication dans la théorie des actes de langage [Searle, 1969, McCarthy, 1992], afin de l'implanter relativement directement dans un système multi-agents (voir Section 6.3).

## 2.5 Synthèse

Dans ce chapitre, j'ai exploré les principaux paradigmes d'apprentissage en gardant un oeil critique vis-à-vis de leur pertinence à la fois pour l'apprentissage humain et pour le contexte de la découverte. J'ai ainsi mis en évidence le fait que l'identification à la limite décrit un processus d'apprentissage infini qui se rapproche de ce fait de l'apprentissage humain (en effet, il n'est jamais trop tard pour apprendre, et seuls les imbéciles ne changent pas d'avis). Toutefois, ce paradigme attribue à l'apprenti un rôle passif qui ne permet pas de mettre en oeuvre une stratégie d'exploration des hypothèses. Avec l'introduction d'un oracle, [Angluin, 1988] transforme radicalement l'apprenti qui devient acteur

de son propre apprentissage. Ceci correspond tout à fait à nos conceptions actuelles de l'apprentissage humain. De plus, cette interaction par requête définit un protocole qu'il est facile de modéliser et d'implanter (voir Chapitre 6).

Les modèles de Gold et de Valiant ont pour but de faire résoudre un problème par une machine. Si l'environnement proposé par Valiant permet de réduire la difficulté des problèmes, et si l'introduction de l'Oracle par Angluin permet à nouveau, par l'interaction, de faire chuter de façon significative la complexité des problèmes formulés, l'utilité du caractère dialectique de la mise en interaction d'un apprenti et de son utilisateur par [Nobrega *et al.*, 2003] n'est confirmé que de façon empirique.

Le problème majeur de l'apprentissage par requêtes réside dans la définition d'un oracle capable de répondre à ces requêtes. J'ai souligné le fait que si la tâche à apprendre est connue de *quelqu'un*, alors cet individu peut tenir le rôle de l'oracle, mais que cet oracle ne peut exister dans le contexte de la découverte, lorsqu'il s'agit d'apprendre quelque chose de totalement nouveau. De plus, la démarche scientifique implique la remise en question des points de vue ou du modèle de description lorsque ceux-ci s'avèrent faux ou insuffisants. Il serait donc imprudent de supposer que les objets considérés ont une description finie et fixe. L'existence d'un oracle capable de vérifier une description infinie supposerait de la part de celui-ci des capacités improbables. Je reviendrai sur ce thème au Chapitre 4 en proposant une formalisation permettant d'illustrer ces paradigmes et leurs limites. J'expliquerai alors comment je propose de déporter la résolution des requêtes sur une communauté d'apprentis, afin de garantir la *réfutation à la limite* des conjectures. De façon générale, je considère dans la suite de ce document qu'on ne propose plus d'exemples à l'apprenti, et que celui-ci doit imaginer des protocoles expérimentaux et les mettre en oeuvre afin de générer des données pour valider ses hypothèses. Dans un tel contexte, il faut auparavant définir de façon logique les jugements pouvant être émis lors du processus de découverte, afin de pouvoir localiser de façon non ambiguë les oppositions qui peuvent exister entre eux (que ces jugements proviennent d'un seul individu ou de plusieurs). Ce travail fait l'objet du Chapitre 3.

J'ai également présenté une formalisation novatrice de l'interaction entre un tuteur et un apprenti, exposée dans [Angluin & Krikis, 2003]. L'apport majeur de ce modèle tient dans le changement d'objet considéré comme réalisant l'apprentissage : l'apprentissage n'est plus le fait d'un individu mais de deux individus en interaction. Ce changement de paradigme permet de focaliser nos efforts sur l'interaction qui est le noyau de cette entité apprenante. Je base certaines perspectives de travail sur cette formalisation qui permet en effet d'imaginer un protocole d'apprentissage dans lequel un apprenti ayant trouvé une solution acceptable pourrait jouer le rôle du tuteur pour un nouvel ap-

prenti. Ceci ouvre la voie vers une plate-forme multi-agents dans laquelle les agents peuvent choisir un autre agent comme tuteur, si cet agent semble plus efficace qu'eux pour un problème donné.



# Chapitre 3

## Une Localisation Dialectique des Contradictions

*la principale différence entre Einstein et  
une amibe (...) est que Einstein cherche  
consciemment l'élimination des erreurs*

Karl R. Popper.

LA LOGIQUE de la découverte scientifique telle qu'elle est présentée par [Popper, 1963] ou [Lakatos, 1976], et étudiée par le Cercle de Vienne, met en avant l'élaboration de normes et les points de ruptures qui accompagnent la formation de théories scientifiques. Ainsi, les contradictions sont souvent sources de découvertes, entraînant de nouvelles approches et la révision des cadres de référence.

Je trouve amusant de citer ici l'exemple géométrique choisi par [Delahaye, 1997] puisque le carré des oppositions est le thème récurrent de ce chapitre :

*Plusieurs fois dans l'histoire des mathématiques, des contradictions ont  
provoqué de graves inquiétudes. Les Grecs ressentirent la découverte de  
l'irrationalité de la diagonale du carré (autrement dit, que la racine carrée de  
2 n'est pas le quotient de deux entiers) comme une contradiction, car ils  
pensaient implicitement que toute grandeur pouvait être exprimée par une  
fraction. La diagonale du carré existait géométriquement, mais pas en tant  
que nombre ! Il fallait définir un autre type de nombre. Cela n'était pas aisé,  
et pendant des siècles les mathématiciens se méfièrent des extensions de la  
notion de nombre : ils se replièrent sur la géométrie, et il a fallu plus d'un  
millénaire pour qu'ils s'en dégagent.*

Jean-Paul Delahaye, Statut mathématique des contradictions..

L'utilisation de paradoxes pour guider la révision d'une théorie ne peut se produire que si la théorie pré-existante ne s'effondre pas totalement sous le

poinds de ses contradictions lorsque celles-ci sont mises en évidence. Afin de définir une logique pragmatique de la découverte, nous devons donc tenir compte de l'existence d'éventuelles contradictions et définir un système permettant de raisonner en présence de jugements opposés. Nous avons publié un tel exemple de logique pragmatique de la découverte dans [Afshar *et al.* , 2006].

En logique classique, une proposition ne peut prendre que deux valeurs de vérité (vrai ou faux). Cependant, durant le processus de découverte, un chercheur utilise des jugements beaucoup moins contrastés en fonction du degré de *certitude* ou de *preuve* qu'il accorde à un énoncé. La logique modale permet de formaliser ces différents degrés en déportant l'opposition entre le vrai et le faux sur les modalités associées aux énoncés plutôt que sur les énoncés eux-mêmes. Afin produire un système complet de jugements scientifiques permettant de codifier l'interaction entre un chercheur et son assistant, nous devons définir un ensemble de modalités reflétant ces jugements, telles que celles de preuve, de réfutation, de conjecture ou encore de postulat. De plus, cet ensemble doit permettre de raisonner de façon non triviale lorsque des contradictions apparaissent ou lorsque les informations disponibles sont insuffisantes pour décider.

Je ne m'engagerai pas dans une discussion sur la complexité d'une preuve (la complexité computationnelle du processus démonstratif a été très tôt et très constamment reconnue [Dubucs, 1992]), je me contenterai de signaler que les jugements modaux que nous définissons ici sont définis par les oppositions qui existent entre eux. La notion de réfutation peut ainsi être définie par rapport à celle de preuve, par exemple, mais l'inverse est également vrai et l'on peut définir la notion de preuve par opposition à celle de réfutation, selon une vision constructiviste de preuve empirique que l'on peut accorder à un énoncé qui résiste à la réfutation. Nous reviendrons sur cette difficulté au Chapitre 4.

Le but de ce chapitre est de montrer que le carré des oppositions est une structure qui s'abstrait de la logique sous-jacente et qu'il peut servir à définir les oppositions logiques entre les jugements modaux que nous souhaitons utiliser comme moteur dialectique de découverte. En s'appuyant sur des extensions paraconsistantes et paracomplètes de cette structure, nous utilisons un *cube des oppositions* pour représenter les oppositions existant entre les modalités qui nous intéressent. Nous verrons que cet ensemble de modalités est clos par négation et qu'une interprétation possible correspond à la définition usuelle des jugements scientifiques correspondants. Cette interprétation est donc non-ambiguë et partagée par un chercheur et son assistant, et leur permet de raisonner en termes de niveau de preuve sur les énoncés logiques qu'ils manipulent, fournissant un cadre logique à leur interaction et à la supervision de l'apprentissage.

Je commencerai par présenter le carré d'Aristote et les différentes formes

d'opposition qu'il représente, ainsi que les interprétations les plus courantes. J'illustrerai ensuite en Section 3.1.2 comment la logique modale est traditionnellement utilisée dans ce contexte, afin de définir les notions de preuve et de réfutation. Puis je présenterai les extensions que [Blanché, 1966], [Béziau, 2003] et [Moretti, 2004] ont proposé en travaillant sur les différentes décorations modales du carré classique (Section 3.1.3). Notamment, un hexagone paraconsistant et un hexagone paracomplet définissent les modalités que nous associons aux notions de conjecture et de postulat. Enfin, en s'appuyant sur [Moretti, 2004] et [Pellissier, 2006], je présenterai comment ces quatre modalités s'opposent dans un cube logique et j'en fournirai une interprétation en Section 3.2. Nos travaux publiés dans [Luzeaux et al, 2008] montrent que ce cube est une sous-structure *forte* d'une structure de plus haut niveau obtenue en appliquant les règles de la logique modale  $S_5$  (voir Section 3.2).

### 3.1 Le Carré des Oppositions

La notion d'opposition est extrêmement importante en philosophie où elle est nécessaire pour penser selon de multiples nuances la notion de différence. L'opposition est également cruciale en logique où elle a trouvé une formulation célèbre, par Aristote, en termes d'un objet géométrico-logique, le « carré des oppositions ». Ce carré logique gère, en mathématiques, rien moins que les rapports fondamentaux entre les quantificateurs (existentiel et universel), la négation et l'implication. Comme tel le carré est également un carrefour de la logique modale (celle-ci étant un calcul des quantifications possibles, sémantiquement interprété en termes de relations entre « mondes »). Cette structure oppositionnelle du carré a inspiré, en dehors de la logique, des formalisations nouvelles, entre autres en linguistique (le « carré sémiotique » de [Greimas & Courtès, 1993]), en psychologie (le « groupe INRC » de [Piaget, 1972a, Piaget, 1972b]) et en psychanalyse (les « formules de la sexuation » et le schéma *RSI* de [Lacan, 1966 (rééd. 1999)]). Des travaux plus récents ont été présentés lors du premier congrès international sur le carré des oppositions, à Montreux (Suisse) en Juin 2007 (voir Section 3.1.3).

La théorie du syllogisme d'Aristote, ainsi que l'ensemble du corpus syllogistique médiéval développé jusqu'au 19<sup>ème</sup> siècle, ne doivent cependant pas être vus comme une théorie formelle de la logique mais plutôt comme une théorie du raisonnement qui permet d'inférer la vérité de certaines propositions (et fondée sur les règles du *modus ponens* et du *modus tollens* qui forment les bases du calcul des propositions). De plus le processus de mise en évidence de la validité, ou de l'invalidité de certaines inférences et conversions (d'une

proposition et de sa négation) est basé sur de simples diagrammes. Le carré des oppositions apparaît donc comme une géométrisation du processus d'inférence.

### 3.1.1 Présentation

Le carré des oppositions (traditionnellement représenté par la Figure 3.1) a été proposé par Aristote au quatrième siècle avant Jésus Christ et apparaît dans les textes logiques depuis lors. Il connecte différentes propositions quantifiées et leurs négations par le biais de quatre notions d'oppositions parmi lesquelles figurent la *contradiction* (définition 3.1), la *contrariété* (définition 3.2), et la *sub-contrariété* (définition 3.3).

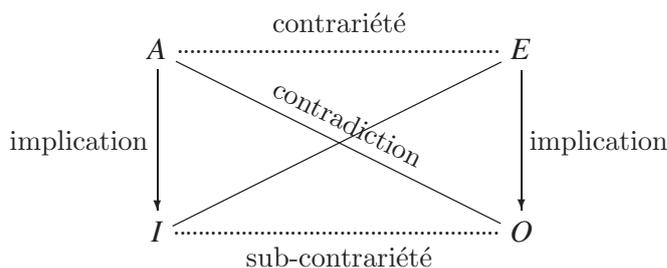


FIGURE 3.1 – Le Carré des Oppositions

**Définition 3.1 (Contradiction)** *La contradiction entre deux propositions est définie comme l'impossibilité pour elles d'être vraies en même temps, ou fausses en même temps.*

**Définition 3.2 (Contrariété)** *La Contrariété entre deux propositions est définie comme l'impossibilité pour elles d'être vraies en même temps, mais la possibilité d'être toutes les deux fausses.*

**Définition 3.3 (Sub-contrariété)** *La Sub-contrariété entre deux propositions est définie comme l'impossibilité pour elles d'être fausses en même temps, mais la possibilité d'être toutes les deux vraies.*

**Définition 3.4 (Sub-alternation)** *La Sub-alternation entre deux propositions se définit comme l'impossibilité d'avoir la première vraie sans que la seconde soit vraie également.*

Comme l'indiquent ces définitions, la notion d'opposition est basée sur différents degrés de vérité. En se plaçant dans le cadre propositionnel, par exemple, nous pouvons interpréter ces degrés de vérité de la façon suivante, ce qui nous amène à l'interprétation plus intuitive du carré fournie par la Figure 3.2.

- définition 3.1 :  $P \leftrightarrow \neg Q$
- définition 3.2 :  $\neg(P \wedge Q)$
- définition 3.3 :  $\neg(\neg P \wedge \neg Q)$
- définition 3.4 :  $P \rightarrow Q$

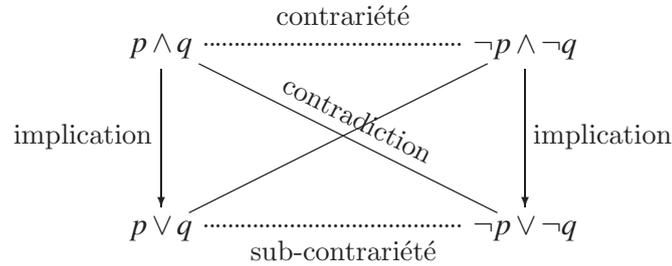


FIGURE 3.2 – Interprétation Propositionnelle du Carré des Oppositions

Au delà de la logique des propositions, le calcul des prédicats permet de s'intéresser à des énoncés relatifs à un ensemble d'objets grâce à l'introduction d'un quantificateur universel « quel que soit » et un quantificateur particulier « il existe ». Ainsi, une proposition universelle telle que « les schtroumpfs sont tous bleus » est une proposition qui devient fautive s'il existe au moins un schtroumpf « noir », par exemple, pour lequel le prédicat « Bleu » ne s'applique pas. La Figure 3.3 illustre cette opposition. Ces quantificateurs permettent la généralisation et la particularisation d'énoncés, qui sont au coeur du processus de découverte et de l'apprentissage à partir d'exemples.

Nous pouvons remarquer que la relation de sub-alternation correspond à un affaiblissement de l'expressivité des formules ainsi liées, et peut être interprétée de façon intuitive par la notion de conséquence. Cette interprétation correspond à la volonté initiale d'Aristote de distinguer non seulement les énoncés positifs des énoncés négatifs, mais aussi les énoncés universels des particuliers, puisqu'un énoncé vrai (respectivement faux) dans le cas général, l'est égale-

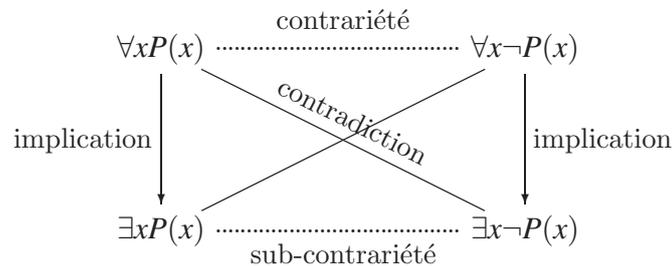


FIGURE 3.3 – Interprétation Prédicative du Carré des Oppositions

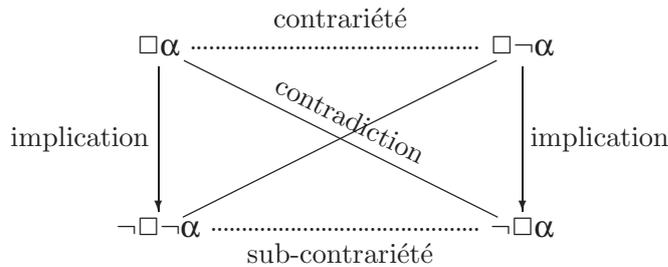


FIGURE 3.4 – Interprétation Modale du Carré des Oppositions

---

ment dans un cas particulier (si tous les oiseaux sont capables de voler, alors ce cygne en particulier, qui est un oiseau, l'est aussi).

L'interprétation la plus communément utilisée pour traduire ce carré est donc la suivante : Les colonnes ( $A, I$ ) et ( $E, O$ ) correspondent aux propositions affirmatives et négatives, tandis que les lignes ( $A, E$ ) et ( $I, O$ ) correspondent respectivement aux propositions universelles et existentielles (ou particulières). Le carré exprime ainsi graphiquement les relations quantificatives fondamentales.

### 3.1.2 Introduction des Modalités

La logique d'Aristote est dite *ontique* car chaque modalité peut être exprimée à partir d'une unique modalité ( $\Box$ ) et d'une négation ( $\neg$ ). Une interprétation modale du carré des oppositions peut être obtenue par clôture en doublant cette négation :  $\Box = \neg\neg\Box$ , comme le montre la Figure 3.4 :  $A$  est  $\Box\alpha$ ,  $E$  est  $\Box\neg\alpha$ ,  $I$  est  $\neg\Box\neg\alpha$  et  $O$  est  $\neg\Box\alpha$ , où  $\alpha$  est une proposition (de la logique classique).

Le recours à des opérateurs modaux [Goldblatt, 2005] permet d'accroître l'expressivité du langage formalisé par la logique propositionnelle et de rendre compte analytiquement de nuances à l'aide d'un formalisme relativement simple. Par rapport à la simple logique propositionnelle, la logique modale permet de soigneusement distinguer un énoncé de la forme « il ne faut pas qu'il pleuve » d'un autre de la forme « il faut qu'il ne pleuve pas ». Dans la langue courante, il n'est pas rare que ces deux énoncés soient utilisés l'un pour l'autre, mais il suffit d'y prêter un peu d'attention pour saisir la différence logique qui existe entre eux. Car le premier peut se paraphraser en « il ne faut pas nécessairement qu'il pleuve », qui ne traduit plus d'impératif, au contraire du second énoncé. Cette différence est encore plus nette sans doute, et source d'encre plus de confusions, lorsque l'on considère d'autres modalités, par exemple la modalité épistémique qui a trait à la connaissance. « Je ne crois pas que le chat est sur

le paillasson » n'est pas logiquement de même forme, et ne signifie pas la même chose que « je crois que le chat n'est pas sur le paillasson ».

De fait, la *modalisation* permet de formaliser toute situation où intervient une attitude propositionnelle, c'est-à-dire un point de vue à l'égard d'une proposition. Elle est donc appropriée au typage d'énoncés par des jugements, ce qui nous intéresse tout particulièrement. Différentes interprétations modales en découlent, selon les nuances que l'on souhaite apporter à la catégorisation des informations par ces modalités : les logiques déontiques traduisent différents niveaux d'autorisation, les logiques épistémiques traduisent des niveaux de savoir, alors que les logiques de la preuve traduisent différents niveaux de certitude dans les énoncés.

Nous pouvons d'ores et déjà fournir une interprétation *scientifique* des modalités représentées sur la Figure 3.4 :

- $\Box\alpha = \alpha$  est prouvé ;
- $\Box\neg\alpha = \alpha$  est réfuté ;
- $\neg\Box\alpha = \alpha$  n'est pas (encore) réfuté ;
- $\neg\Box\neg\alpha = \alpha$  n'est pas (encore) prouvé.

Cette première interprétation du carré nous permet de définir la preuve et la réfutation l'une par rapport à l'autre. L'emploi de la double négation définit une dichotomie très stricte (ce qui n'est pas *non-réfuté* - ou ce qui n'est pas *contingent*- est *prouvé*) pouvant être utile pour qualifier une situation figée, un état de connaissance. Cependant, cette dichotomie est trop restrictive pour un raisonnement scientifique qui se doit de prévoir et rechercher d'éventuelles contradictions. Cette interprétation ne laisse aucune place apparente pour les notions de postulat et de conjecture qui sont capitales pour la supervision d'un apprentissage comme nous l'avons vu au Chapitre 1. De plus, comme nous le verrons au Chapitre 4, la preuve ou la réfutation n'est que le résultat d'une interaction fondée sur la définition de postulats et la vérification de conjectures.

Le problème que nous venons de soulever n'est pas isolé, car le carré des oppositions peut être source d'ambiguïtés logiques et a été fortement critiqué au siècle dernier avec le développement de la logique formelle par Boole et Frege. Il a ainsi inspiré de nombreuses extensions. [Horn, 2007] dresse un état de l'art des extensions majeures, [Engel, 2007] met en évidence la diversité des degrés de croyances en logique épistémique et [Cavaliere & Donnarumma, 2007] propose de graduer les arêtes pour représenter la logique floue ; [Brenner, 2007] propose d'appliquer le carré à une notion de temporalité pour modéliser les différents états d'un processus, et [Beller, 2007] démontre en étudiant l'apprentissage des notions déontiques par des enfants que le carré est adapté à la représentation de normes sociales. Nous présentons maintenant celles sur lesquelles nous fondons notre proposition.

### 3.1.3 Extensions et Prise en Compte de Contradictions

La remarque la plus constructive concerne l'incompatibilité entre le carré des oppositions et le triangle de contrariété (Figure 3.5). En effet, l'un exprime une possibilité unilatérale tandis que l'autre, bien que moins connu, exprime de façon plus juste l'incompatibilité bilatérale entre une « possibilité » et l'impossibilité d'une part, la nécessité d'autre part, en définissant un sommet  $Y = I \wedge O$  qui n'était pas pris en compte dans le carré.

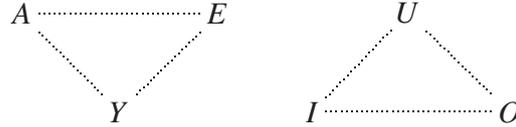


FIGURE 3.5 – Les Triangles de Contrariété (à gauche) et de sub-Contrariété (à droite)

Une solution à cette aporie fut fournie par Blanché en 1953, qui proposa de considérer une nouvelle figure, duale du triangle de contrariété : le triangle de sub-contrariété (également représenté sur la Figure 3.5) qui introduit un nouveau sommet  $U = A \vee E$ . Le bi-produit de ces deux triangles forme un hexagone logique (Figure 3.6) dans lequel les deux notions distinctes de possibilités (unilatérale et bilatérale) se trouvent désormais entrelacées. Comme je le présenterai en Section 3.2, l'extension que je propose de considérer résulte d'une généralisation de ce bi-produit.

[Blanché, 1966] décore ensuite les sommets de l'hexagone avec des modalités, et développe une théorie du raisonnement. Les sommets introduits par les triangles de contrariété et de sous-contrariété sont la conjonction et la disjonction des modalités voisines :  $I \wedge O$  est  $\neg \Box \neg \alpha \wedge \neg \Box \alpha$  et  $A \vee E$  est  $\Box \alpha \vee \Box \neg \alpha$ .

L'hexagone de Blanché n'est pas unique. En effet cinquante ans après Blanché, [Béziau, 2003] découvrit d'autres hexagones en travaillant sur les différentes décorations modales possibles de cet hexagone, et en considérant le fait

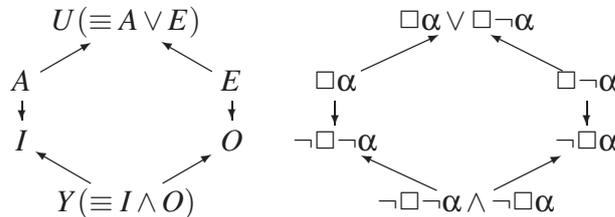


FIGURE 3.6 – l'Hexagone de Blanché (1953) et l'Hexagone Modal (1966)

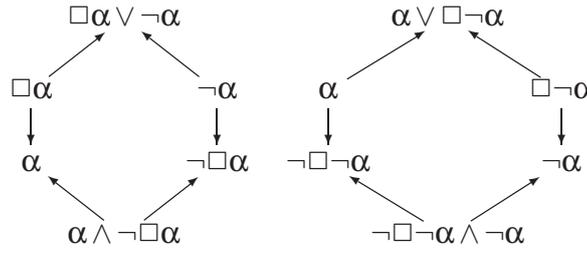


FIGURE 3.7 – les Hexagones Paraconsistant (à gauche) et Paracomplete (à droite).

qu'une proposition « nue », c'est à dire sans modificateur modal, est également une modalité. Il soutient que l'un de ces hexagones est paraconsistant et l'autre paracomplete (Figure 3.7), alors que l'hexagone de Blanché peut être considéré comme classique. Nous proposons d'interpréter cette modalité  $\alpha$ , qui est contrainte par les modalités  $\Box$  et  $\Box\neg$ , comme une « hypothèse » selon le sens usuel d'hypothèse de travail que l'on va chercher à prouver ou à réfuter ; par exemple, la description logique d'une observation.

**Définition 3.5 (Hypothèse (source : [www.wikipedia.fr](http://www.wikipedia.fr)))** *Une hypothèse est une proposition ou une explication que l'on se contente d'énoncer sans prendre position sur sa véracité, c'est-à-dire sans l'affirmer ou la nier. Une fois énoncée, une hypothèse peut être étudiée, confrontée, utilisée, discutée ou traitée de toute autre façon jugée nécessaire, par exemple dans le cadre d'une démarche expérimentale.*

*Une hypothèse destinée à être travaillée ou vérifiée est désignée par l'expression « hypothèse de travail » ; au contraire, une hypothèse utilisée sans intention de la vérifier [...] constitue un postulat.*

Il est intéressant de remarquer que chaque hexagone renferme trois carrés distincts. De plus, si l'on observe la colonne de termes négatifs *EO* dans chaque hexagone, on peut voir que la relation de sub-alternation  $\Box\neg\alpha \rightarrow \neg\Box\alpha$ , représentée dans l'hexagone dit *classique* se voit précisée dans les deux autres :

$$\Box\neg\alpha \rightarrow \neg\alpha \rightarrow \neg\Box\alpha.$$

En effet, ces termes expriment différentes sortes de négations en logique classique et modale, avec les relations d'affaiblissement correspondantes : [Béziau, 2005, Došen, 1984, Dosen, 1984] montrent que  $\Box\neg$  est une négation intuitionniste paracomplete, et [Béziau, 2002] montre que  $\neg\Box$  est une négation paraconsistante.

Il est également intéressant de remarquer que l'hexagone paraconsistant est obtenu à partir de l'hexagone paracomplete (et réciproquement) en remplaçant

$\alpha$  par  $\neg\alpha$  (le phénomène décrit par l'énoncé  $\alpha$  a-t-il été observé ou non?). Par conséquence, ni la modalité  $\neg\Box\neg$  (aussi notée  $\Diamond$ , interprétée comme *possible* ou *non réfuté*), ni la modalité  $\Box\neg$  (interprétée comme *réfuté*), c'est à dire les modalités possédant une négation à droite, n'intervient dans l'hexagone paraconsistant. En revanche cet hexagone introduit une nouvelle modalité *universelle*,  $\alpha \wedge \neg\Box\alpha$ , qui nous intéresse particulièrement puisqu'elle correspond à la définition usuelle d'une conjecture (« hypothèse non prouvée ») :

**Définition 3.6 (Conjecture (source : [www.wikipedia.fr](http://www.wikipedia.fr)))** *Jugement probable, opinion que l'on fonde sur des apparences.*

Une conjecture représente donc une hypothèse probable, qui n'a pas encore été prouvée.

Si l'on considère de la même façon, la modalité *universelle*  $\neg\alpha \wedge \neg\Box\neg\alpha$  introduite par l'hexagone paracomplet, on peut se rendre compte qu'elle correspond à la définition usuelle d'un postulat.

**Définition 3.7 (Postulat (source : [www.wikipedia.fr](http://www.wikipedia.fr)))** *On nomme postulat un principe utilisé dans la construction d'un système déductif, mais qu'on ne démontre pas lui-même, sans pour autant s'interdire la possibilité de s'y essayer plus tard [...].*

*On peut donc utiliser un postulat avec l'assentiment de l'auditeur, qui le prend comme un principe non démontré mais sans doute légitime, car semblant intuitivement non contestable [...]. La plupart des postulats sont des marques de bon sens, des appuis sur l'expérience.*

Au contraire d'une conjecture, qui est une hypothèse qui n'est pas encore prouvée, mais résistant à la réfutation et fondée sur une observation, le postulat décrit un énoncé qui n'a aucun fondement expérimental, mais qui n'est pas réfuté. Si l'on considère l'interaction entre un chercheur et son assistant rationnel (Chapitre 1), les règles provenant du chercheur sont postulées, celui-ci les énonce d'après son expérience personnelle, ses connaissances. En revanche, les énoncés représentant les règles logiques apprises par l'assistant sont conjecturées. Nous illustrerons en Section 3.2 le rôle de ces modalités dans la boucle d'interaction entre le chercheur et son assistant, mais également durant le processus d'apprentissage (de la machine) et le processus de découverte (du chercheur).

La question qui se pose maintenant est la suivante : « existe-t-il d'autres structures géométriques venant étendre le carré d'Aristote tout comme certaines logiques étendent la logique classique ? ». La réponse à cette question est oui. En effet, cette façon d'étendre le carré original a été étudiée et généralisée par [Moretti, 2004] et [Pellissier, 2006] après que Moretti et Béziau se soient

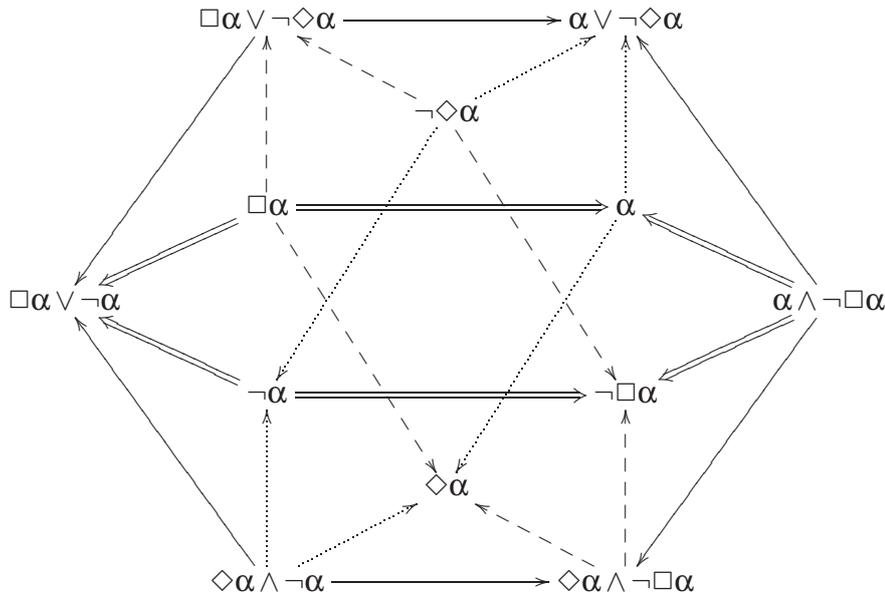


FIGURE 3.8 – Le Tétradécaèdre des Oppositions

rendus compte qu'associés à un quatrième hexagone (Moretti et Smessaert), une structure tri-dimensionnelle de plus haut niveau pouvait être représentée : un *tétradécaèdre* des oppositions (voir Figure 3.8), produit des quatre hexagones.

Le fait que l'on ait découvert très récemment que le carré des oppositions n'est qu'une particularité locale d'une branche nouvelle, très géométrisée, de la logique mathématique (la « théorie de la n-opposition », 2004) suggère à la fois des possibilités inédites de relecture de toute la logique modale abstraite (théorie des graphes modaux) et appliquée (géométrisation des logiques déontiques, temporelles, épistémiques et multi-modales). Entre autres, ces travaux nous ont permis de positionner notre approche dans un contexte plus général pour définir notre cube des jugements (3.2). Ils ont également servi de base à [Luzeaux et al, 2008] pour montrer que ce cube fait partie d'une structure de plus haut niveau dont la logique sous-jacente n'est autre que la logique  $S_5$  (cet article est présenté en annexe D).

## 3.2 Le Cube des jugements

En généralisant cette évolution des extensions du carré en structures géométriques de plus haut niveau, [Moretti, 2004] définit une *théorie de la n-opposition*.

Cette théorie, fondée sur l'organisation des simplex d'ordre inférieur, défi-

nit de façon récursive les structures géométriques pouvant être construites à partir des symétries liées aux oppositions qui existent dans un ensemble de modalités donné. Ainsi, le triangle de contrariété et le triangle de sub-contrariété se fondent en un hexagone des oppositions représentant les relations de sub-alternation entre les sommets du premier et ceux du second (voir Figures 3.5 et 3.6 à la section précédente). De la même façon, les quatre modalités universelles que sont la preuve  $\Box\alpha$ , la réfutation  $\Box\neg\alpha$ , la conjecture  $\alpha \wedge \neg\Box\alpha$  et le postulat  $\neg\alpha \wedge \neg\Box\neg\alpha$ , forment un tétraèdre de contrariété qui peut également être associé à son dual, formé des modalités opposées (Figure 3.9), pour former un cube des oppositions (Figure 3.10).

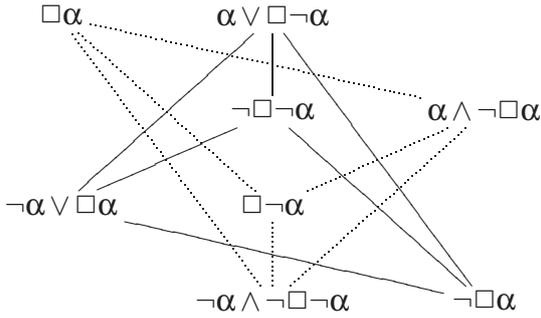


FIGURE 3.9 – Les Tétraèdres de Contrariété (en pointillés) et sub-Contrariété

Par construction, ce cube logique contient six carrés des oppositions (Figure 3.11), et représente plusieurs opérateurs de contingence ( $\alpha \wedge \neg\Box\alpha$  et  $\neg\Box\neg\alpha \wedge \neg\alpha$ ) qui précisent le cas de pure contingence  $\neg\Box\alpha \wedge \neg\Box\neg\alpha$  du carré modal. Ces opérateurs sont ceux introduits par les hexagones paracomplets et paraconsistants (cf. Section 3.1.3).

Cette structure n'est pas complète, puisqu'elle ne représente pas toutes les modalités possibles. Entre autre, l'hypothèse liée à l'observation,  $\alpha$ , qui nous a servi à la construire n'y figure pas. La raison est que ce cube est une sous-structure « forte » d'un tétraki-hexaèdre<sup>12</sup> des oppositions (voir Figure

12. un tétraki-hexaèdre (tetraicosahedron en anglais) est formé d'un cube augmenté d'une

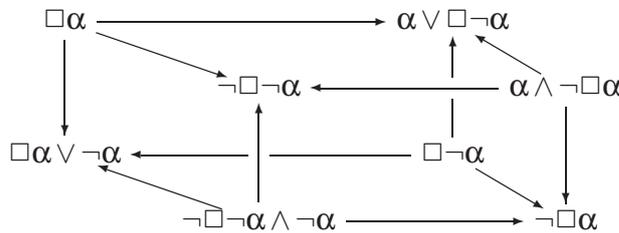


FIGURE 3.10 – Le Cube des Oppositions

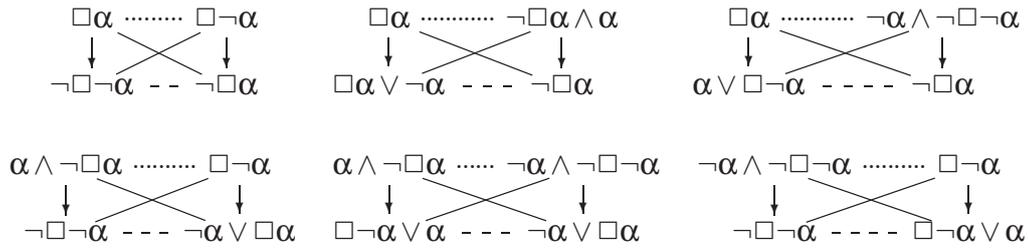


FIGURE 3.11 – Les six carrés des oppositions du cube logique.

3.12), « faible », dont les sommets sont obtenus par clôture de  $\{\alpha, \Box\alpha\}$  par les opérations logiques  $\{\neg, \wedge, \vee\}$ , en se plaçant dans le cadre de la logique  $S_5$ . Le terme *fort*, proposé par [Pellissier, 2006], caractérise les structures fondées sur des simplex de contrariété dont les termes ne sont pas seulement contraires deux à deux (*i.e.*, leur conjonction deux à deux est fausse), mais leur disjonction globale est vraie. Le terme *faible* s'applique dans des situations où la disjonction de tous les termes appartenant à un simplex de contrariété donné n'est pas nécessairement vraie. L'une des originalités de [Pellissier, 2006] réside dans cette distinction, alors que les précédentes approches avaient toujours supposé des constructions « fortes ».

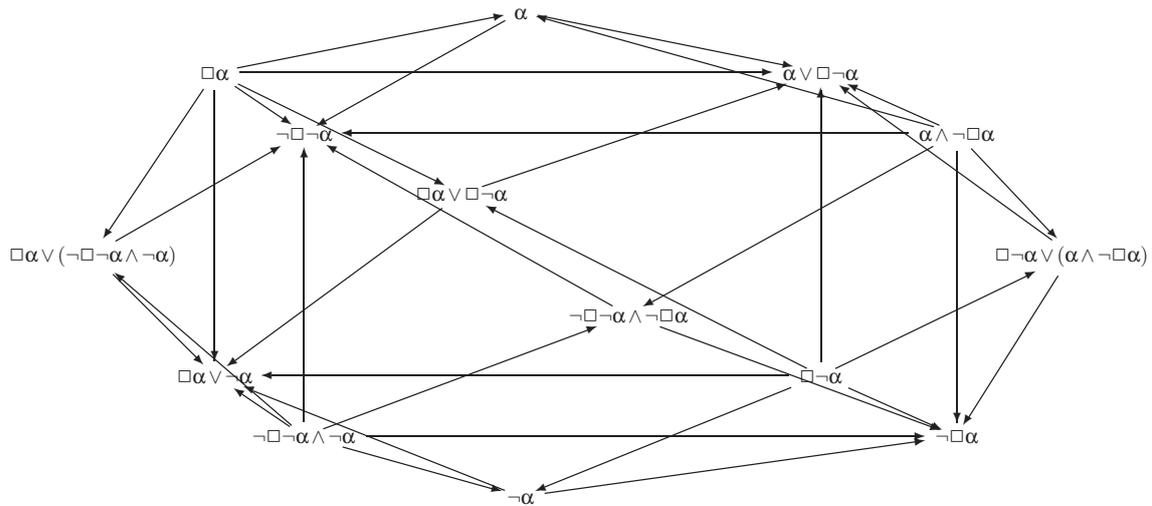


FIGURE 3.12 – Le tétraki-hexaèdre modal construit à partir des axiomes de la logique  $S_5$ .

On peut considérer ce cube des jugements non pas comme un modèle complet de raisonnement, mais plutôt comme un modèle d'interaction tenant pyramide à base carrée sur chaque face.

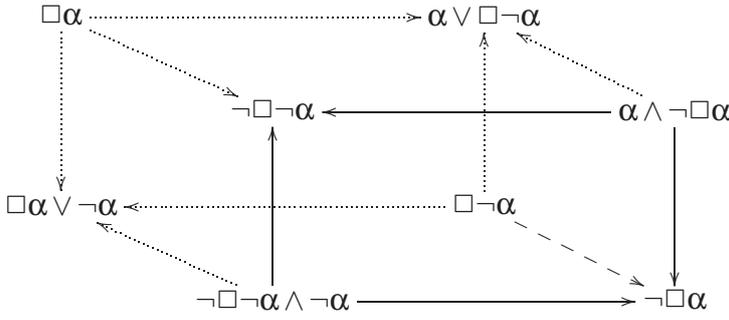


FIGURE 3.13 – Confrontation de Jugements Opposés

compte des différentes oppositions pouvant exister entre les jugements de deux agents, un chercheur et son assistant par exemple. Lorsque deux jugements sont en opposition, il faut admettre que l'un d'eux au moins est faux, et trouver un consensus permettant de continuer à raisonner logiquement jusqu'à ce que l'erreur soit circonscrite. Par exemple (Figure 3.13, si un agent considère un énoncé  $\alpha$  comme un postulat :  $\neg\Box\neg\alpha \wedge \neg\alpha$  et qu'un deuxième agent considère le même énoncé comme une conjecture :  $\alpha \wedge \neg\Box\alpha$ , ils peuvent s'entendre soit sur le fait que l'énoncé n'est pas prouvé :  $\neg\Box\alpha$  (connotation négative) ou qu'il n'est pas réfuté :  $\neg\Box\neg\alpha$  (connotation positive). Si un troisième agent considère l'énoncé comme réfuté, alors le consensus peut être trouvé en considérant l'énoncé comme peu probable ( $\neg\Box\alpha$ ) en exploitant les relations de subalternation du cube.

Terminons cette section en présentant brièvement la logique  $S_5$  pour comprendre l'intérêt qu'elle suscite chez beaucoup de logiciens.

En partant de la logique propositionnelle classique (avec les connecteurs  $\wedge$ ,  $\vee$ ,  $\neg$  et  $\rightarrow$  usuels ainsi que les théorèmes usuels tels que les lois de De Morgan, la contraposition et l'involutivité de la négation), on obtient un système qui ne comporte que deux modalités (l'identité  $\cdot$  et la négation  $\neg$ ) qui rendent compte de l'interprétation classique : une chose *est* ou *n'est pas*. La hiérarchie des systèmes logiques pouvant être définis à partir de ce système est la suivante :

- *système KT* : obtenu en ajoutant un opérateur modal ( $\Box$ ), les axiomes  $T$  ( $\Box\alpha \rightarrow \alpha$ ) et  $K$  ( $\Box(\alpha \rightarrow \beta) \rightarrow (\Box\alpha \rightarrow \Box\beta)$ ), ainsi qu'une règle d'inférence connue sous le nom de nécessité : « si  $\alpha$  est un théorème, alors  $\Box\alpha$  aussi ». Rappelons que  $\neg\Box\neg$  peut s'écrire  $\Diamond$ . ce système possède en fait une infinité de modalités affirmatives équivalentes (les modalités affirmatives sont les itérations finies de  $\Box$  et  $\neg$  possédant un nombre pair de  $\neg$ , tandis que les négatives en possèdent un nombre impair).
- *système  $S_4$*  : obtenu en ajoutant au système  $KT$  l'axiome 4 :  $\Box\alpha \rightarrow \Box\Box\alpha$ . On appelle aussi cet axiome l'axiome d'introspection : « si je sais quelque

chose, alors je sais que je le sais » (ou encore « si une chose est prouvée, alors il est prouvé qu'elle est prouvée »).

- système  $\mathcal{S}_5$  : obtenu en ajoutant à  $\mathcal{S}_4$  l'axiome 5 :  $\diamond\alpha \rightarrow \square\diamond\alpha$  (i.e.,  $\neg\square\neg\alpha \rightarrow \square\neg\square\neg\alpha$ ). On appelle aussi cet axiome l'axiome d'introspection négative : « si je ne sais pas quelque chose, alors je sais que je ne le sais pas » (ou encore « si une chose n'est pas prouvée, alors il est prouvé qu'elle n'est pas prouvée »).

Le système  $\mathcal{S}_4$  possède 14 modalités distinctes et irréductibles ( $\cdot$ ,  $\square$ ,  $\diamond$ ,  $\diamond\square$ ,  $\square\diamond$ ,  $\square\square$ ,  $\diamond\square\diamond$ , et leurs négations), alors que le système  $\mathcal{S}_5$  n'en possède que 6 ( $\cdot$ ,  $\square$ ,  $\diamond$ , et leurs négations).

$\mathcal{S}_5$  est la plus simple extension modale de la logique classique, en ce sens que toute combinaison finie de  $\neg$ ,  $\square$ ,  $\diamond$  précédant une formule peut se réduire à une combinaison d'au plus un  $\neg$  et un  $\square$ . Il est à ce titre largement utilisé. De plus, si l'on regarde plus attentivement les interprétations modales du carré d'Aristote qui ont abouti à la syllogistique médiévale, il s'avère que  $\mathcal{S}_5$  était le système utilisé : l'arête gauche ( $\square\alpha \rightarrow \diamond\alpha$ ) ne nécessite qu'une logique très faible (le système  $D$ ), mais il faut une logique plus forte pour résoudre les ambiguïtés logiques en rendant toutes les vérités modales nécessaires (axiomes 4 et 5), et  $\mathcal{S}_5$  est le système communément employé pour ça. En effet, ce système est célèbre pour sa compatibilité avec la sémantique des mondes possibles.

Il semblerait donc que la paraconsistance et la paracomplétude, qui nous ont permis de mettre en évidence les modalités de conjecture et de postulat, soient des sous-systèmes particuliers de  $\mathcal{S}_5$  [Béziau, 2007].

### 3.3 Synthèse

La découverte scientifique est un processus collectif rendu possible par la traçabilité des jugements, des résultats positifs et négatifs, des théories et des conjectures à travers leur publication et leur évaluation au sein d'une communauté. Sans cette traçabilité, les résultats scientifiques ne dureraient pas suffisamment longtemps pour influencer les travaux des autres membres de la communauté, et il n'y aurait pas de Science à proprement parler. Cette traçabilité est donc la clé de la localisation des points de débats entre les membres d'une communauté, de l'ouverture de nouveaux domaines de recherche, de la mise en évidence de problèmes et de paradoxes qui nécessitent une étude plus approfondie, et de l'établissement d'un cadre de référence consensuel. Ce processus collectif entraîne une organisation sociale au sein de laquelle certains membres se spécialisent dans la publication de résultats, la preuve ou la réfutation de théories ou de théorèmes, et acquièrent un certain crédit qui les définissent comme des références dans leur domaine.

Le carré d'Aristote, qui exprime graphiquement les relations quantificatives fondamentales nous a servi de point de départ pour la définition logique de ces jugements. Ayant souvent été source de discussions et de désaccords, il reste un outil pratique de par la simplicité de sa représentation géométrique et la généralité de ses interprétations. Parmi les extensions qui ont été proposées pour étendre le pouvoir expressif du carré, nous avons présenté deux hexagones, l'un paraconsistant et l'autre paracomplet. Pour offrir un cadre logique à la dialectique de la découverte scientifique, nous avons défini, en partant de ces extensions, un ensemble clos de modalités pouvant être interprétées comme divers degrés de vérité attribués aux énoncés traités durant le processus de formation de théories, depuis la formulation d'hypothèses de travail jusqu'à la réfutation de théories. Le cube logique qui en découle synthétise les différentes oppositions qui peuvent exister entre deux jugements scientifiques et permet de clore l'ensemble des jugements portant sur les énoncés traités durant un raisonnement logique, au lieu de clore le monde sur lequel on doit pouvoir raisonner. La restriction de l'interaction entre les acteurs d'une résolution de problème (un chercheur et son assistant rationnel par exemple) à l'échange de jugements ainsi définis assure une absence d'ambiguïtés dans le typage des connaissances.

# Chapitre 4

## Formalisation du Protocole

*You can tell whether a man is clever by his answers. You can tell whether a man is wise by his questions*

Naguib Mahfouz.

TOUT au long de ce document, j'ai discuté des différents enjeux de l'assistance à la découverte scientifique. Après avoir identifié les limites des approches actuelles et les pistes de recherche proposées par la communauté, j'ai mis en avant la nécessité de formaliser un protocole d'interaction adapté à la fois à l'apprentissage automatique et à l'apprentissage humain, puisque ces deux processus sont au coeur de la découverte scientifique computationnelle (voir Chapitre 1). Le premier enjeu de cette formalisation était de définir un cadre logique pour cette interaction permettant de circonscrire les oppositions existant entre différents jugements. J'ai présenté au Chapitre 3 une extension du carré d'Aristote fondée sur les oppositions entre des jugements modaux. Si cette structure n'est qu'une possibilité parmi d'autres, elle a le mérite de définir un ensemble clos de modalités suffisamment variées pour représenter les jugements nécessaires à la démarche scientifique.

Maintenant que les jugements permettant de confronter des points de vues durant le processus de découverte sont définis de façon non ambiguë, intéressons nous à la communication de ces jugements en définissant un protocole d'interaction qui permette l'apprentissage automatique et qui soit représentatif de la démarche scientifique. Parmi les protocoles existants, j'ai présenté au Chapitre 2 l'*identification à la limite*, l'*apprentissage par requêtes*, et l'apprentissage *PAC*, pour leur impact sur la communauté. J'ai mis en évidence le fait que la façon de modéliser l'interaction entre l'apprenti et son environnement a connu une évolution similaire à celle qu'a connu notre façon d'appréhender l'apprentissage humain, à savoir que le rôle attribué à l'apprenti est devenu de plus en plus actif. Au delà de cette différence, chacun de ces paradigmes définit

une certaine forme de réalité et un critère de réussite différent. Je pense qu'il serait utile d'homogénéiser ces paradigmes sous un formalisme commun afin de pouvoir les comparer avec plus de précision.

Ce travail, actuellement en cours de soumission pour [Dartnell et Martin, 2008], n'aurait pu être achevée sans l'aide d'Éric Martin qui lui a apporté rigueur et précision.

Je vais donc présenter en Section 4.1 une formalisation logique qui nous permettra de préciser la discussion informelle qui faisait l'objet du Chapitre 2. Cette formalisation reste la plus simple possible tout en permettant d'exprimer de façon homogène les paradigmes considérés, et elle est suffisamment expressive pour énoncer les difficultés liées au contexte de découverte. Entre autres, elle tient compte de l'enrichissement éventuel du vocabulaire de description par des prédicats *théoriques*, définis logiquement à partir de prédicats dits *observationnels* formant le langage de base du domaine. Pour des raisons de simplicité, nous avons choisi le domaine des cartes à jouer comme illustration des différentes notions définies ici. La forme de réalité que nous proposons se présente ainsi sous la forme d'une séquence infinie de cartes. Après avoir introduit un vocabulaire permettant de décrire une telle séquence, nous utiliserons la notion de programme logique pour distinguer un programme univoque (décrivant de façon finie un monde possible infini) d'un programme équivoque (décrivant de façon finie un ensemble de mondes possibles infinis). J'utiliserai cette distinction pour illustrer sur un jeu de cartes comment un changement de paradigme d'apprentissage influence les règles du jeu. Entre autres, j'exposerai en illustrant l'apprentissage par requêtes que supposer l'existence d'un oracle répondant aux requêtes de l'apprenti impliquerait que cet oracle possède des capacités supérieures à celle d'une machine de Turing. Pour contourner ce problème, je proposerai en Section 4.5 de distribuer la résolution de ces requêtes sur une communauté d'apprentis qui interagissent *via* la publication de jugements scientifiques. Enfin, nous verrons que le jeu *Éleusis+Nobel*, présenté au Chapitre 5, est une implantation partielle de ce protocole fondé sur la publication de conjectures et de réfutations seulement, dans une vision poppérienne de l'activité scientifique

## 4.1 Langage Formel

Tous les problèmes considérés dans ce chapitre seront fondés sur une représentation abstraite d'une certaine forme de réalité simple et intuitive. Nous proposons de représenter toute forme possible de réalité par une séquence infinie de cartes appartenant au jeu de 52 cartes classique, et dans laquelle chaque carte peut apparaître une infinité de fois.

### 4.1.1 Langage de base

Notre représentation d'une séquence de cartes est fondée sur un langage logique dont les deux seuls symboles fonctionnels sont une constante  $\bar{0}$  et une fonction unaire  $s$ . Le terme obtenu à partir de  $\bar{0}$  par  $n$  applications successives de  $s$  est dénoté  $\bar{n}$ . Intuitivement, ce terme fait référence à la  $n^{\text{ième}}$  carte de la séquence.

Pour décrire une carte, nous utilisons des prédicats observationnels *i.e.*, les membres des ensembles de symboles de prédicats unaires suivants :

$$\text{Prd}_{\text{suit}}(\mathcal{V}) = \{\text{hearts}, \text{spades}, \text{clubs}, \text{diamonds}\}, \text{ et}$$

$$\text{Prd}_{\text{rank}}(\mathcal{V}) = \{\text{ace}, \text{two}, \dots, \text{ten}, \text{jack}, \text{queen}, \text{king}\}$$

Ce vocabulaire est assez riche pour décrire de façon unique n'importe quelle séquence de carte. Par exemple, la séquence « dame de cœur, roi de pique, trois de carreaux, ... », est représentée par la théorie

$$\text{queen}(\bar{0}), \text{hearts}(\bar{0}), \text{king}(\bar{1}), \text{spades}(\bar{1}), \text{third}(\bar{2}), \text{diamonds}(\bar{2}), \dots$$

De façon plus générale, toute *séquence possible*, par analogie à la notion de monde possible introduite par Leibniz (voir Chapitre 1), peut être identifiée par une théorie respectant la Définition 4.1.

**Définition 4.1 (Séquence Possible)** Une séquence possible est une théorie telle que :

- pour tout  $n \in \mathbb{N}$ ,  $\mathcal{T}$  contient un membre de  $\text{Prd}_{\text{rank}}(\mathcal{V})$  et un seul,
- pour tout  $n \in \mathbb{N}$ ,  $\mathcal{T}$  contient un membre de  $\text{Prd}_{\text{suit}}(\mathcal{V})$  et un seul,
- $\mathcal{T}$  ne contient aucune autre formule.

Beaucoup de séquences possibles n'ont pas d'équivalent fini. En revanche, certaines d'entre elles admettent une représentation finie. Par exemple, la séquence constituée uniquement de dames de cœur est décrite par la formule

$$\forall x, \text{hearts}(x) \wedge \text{queen}(x). \quad (4.1)$$

Lorsque le problème est d'identifier une séquence possible à partir d'observations finies, comme c'est le cas dans le problème de l'identification à la limite, les séquences possédant une représentation finie sont d'un intérêt particulier. Nous ne nous limitons pourtant pas à cette classe de séquences possibles, puisque nous souhaitons pouvoir exprimer des problèmes de classification, dont le but est de déterminer si la séquence possible sous-jacente à une observation possède certaines propriétés (*i.e.*, appartient à une classe particulière), comme c'est le cas dans le problème de l'apprentissage par requêtes ou l'apprentissage à partir d'exemples. Dans ce contexte, l'intérêt ne se porte plus sur les

séquences possibles mais sur les classes de séquences possibles qui possèdent une représentation finie.

Je reviendrai sur ce sujet en illustrant les différents paradigmes d'apprentissage, mais considérons maintenant la propriété de posséder une représentation finie. Par exemple, la séquence qui commence par dix dames de cœur et qui continue par une répétition infinie du roi de pique peut se décrire par la théorie finie ci-dessous :

$$\begin{aligned} & \{queen(\bar{n}), hearts(\bar{n}) \mid n \leq 9\} \cup \\ & \quad \{king(\bar{10}), spades(\bar{10})\} \cup \\ & \quad \{\forall x, (king(x) \rightarrow king(s(x))) \wedge (spades(x) \rightarrow spades(s(x)))\} \end{aligned} \quad (4.2)$$

La séquence qui consiste en une alternance de dames de cœur et de rois de pique, en commençant par une dame de cœur, est décrite par la théorie finie suivante :

$$\begin{aligned} & \{queen(\bar{0}), hearts(\bar{0}), \\ & \quad \forall x, (queen(x) \rightarrow king(s(x))) \wedge (hearts(x) \rightarrow spades(s(x))) \\ & \quad \forall x, (king(x) \rightarrow queen(s(x))) \wedge (spades(x) \rightarrow hearts(s(x)))\} \end{aligned} \quad (4.3)$$

Il n'est pas difficile de constater que la classe des séquences possibles possédant une représentation finie est précisément la classe des séquences de cartes qui présentent des régularités, qui deviennent cycliques à partir d'un certain point. Ceci n'a rien d'étonnant si l'on considère la théorie de la complexité qui souligne que les *objets* les plus difficiles à comprimer (ou compresser) sont ceux qui sont organisés de façon aléatoire (voir Annexe B). Ainsi, la taille d'un programme capable de représenter un objet complètement désorganisé se rapproche de la taille de l'objet lui même, alors que si l'objet reste organisé à partir d'un certain point, alors sa description peut être comprimée. Par exemple, la classe des séquences possibles commençant par une sous-séquence finie de dames de coeur, de longueur arbitraire, suivie par une répétition de rois de pique, est décrite de façon finie par la théorie suivante :

$$\begin{aligned} & \{\forall x((king(x) \wedge spades(x)) \vee (queen(x) \wedge hearts(x))), \\ & \quad \exists x king(x) \wedge spades(x), \\ & \quad \forall x((king(x) \wedge spades(x)) \rightarrow (king(s(x)) \wedge spades(s(x))))\} \end{aligned} \quad (4.4)$$

Un dernier exemple intéressant est la classe des séquences possibles dans lesquelles une dame apparaît au moins une fois toutes les trois cartes est équivalente à la formule :

$$\forall x, queen(x) \vee queen(s(x)) \vee queen(s(s(x))).$$

L'ensemble des symboles de prédicats actuel ne permet cependant pas de définir de façon finie les classes de séquences telles que celles alternant la dame de cœur et le roi de pique, avec n'importe quelle séquence de cartes entre chaque occurrence d'une dame de cœur (respectivement, d'un roi de pique) et la prochaine occurrence du roi de pique (resp. de la dame de cœur). Afin de pouvoir fournir une description finie pour cette classe, nous devons enrichir le vocabulaire d'un nouveau symbole de prédicat binaire,  $<$ , interprété comme l'ordre naturel sur les entiers.

$$\begin{aligned} & \{\exists x(\text{queen}(x) \wedge \text{hearts}(x)), \\ & \quad \forall x(\text{queen}(x) \wedge \text{hearts}(x) \rightarrow \exists y(x < y \wedge \text{king}(y) \wedge \text{spades}(y)), \\ & \quad \forall x(\text{king}(x) \wedge \text{spades}(x) \rightarrow \exists y(x < y \wedge \text{queen}(y) \wedge \text{hearts}(y))\} \quad (4.5) \end{aligned}$$

### 4.1.2 Enrichissement du Vocabulaire

Nous proposons maintenant de formaliser l'enrichissement du vocabulaire de base (composé de  $\bar{0}$ ,  $s$ ,  $<$ , et des prédicats *observationnels*). Comme nous l'avons présenté au Chapitre 1, cet enrichissement du vocabulaire est inhérent au processus de découverte et correspond au principe de nominalisation selon lequel il est nécessaire de pouvoir représenter par des concepts abstraits certaines propriétés des objets observés. Nous proposons donc, comme le proposait déjà [Muggleton & Buntine, 1988], [Muggleton, 1994] ou encore [Banerji, 1997], d'enrichir ce vocabulaire par un certain nombre (possiblement aucun) de prédicats dont l'interprétation est fixe sur l'ensemble des séquences possibles. Ces *prédicats théoriques* enrichissent le pouvoir expressif du vocabulaire de la même façon que le symbole  $<$ , mais n'ont réellement d'effet que sur la taille et la lisibilité du programme résultant. Certains de ces prédicats n'apportent qu'une facilité de notation :  $\text{red}(x)$ , par exemple, peut servir d'abréviation pour  $\text{hearts}(x) \vee \text{diamonds}(x)$ .

Nous imposons que tout *prédicat théorique* soit décidable, comme présenté dans la Définition 4.2.

**Définition 4.2** *Soit  $\mathcal{V}$  un vocabulaire composé de l'ensemble des prédicats observationnels et enrichi d'un certain nombre de prédicats théoriques.  $\mathcal{V}$  est décidable ssi pour tout  $k \in \mathbb{N}$ , tout  $n_1, \dots, n_k \in \mathbb{N}$ , tout prédicat théorique  $k$ -aire  $q$ , et toute séquence possible  $\mathcal{G}$ , la valeur de vérité de  $q(\bar{n}_1, \dots, \bar{n}_k)$  dans  $\mathcal{G}$  est déterminée par la valeur de vérité dans  $\mathcal{G}$  de toutes les formules de la forme  $p(\bar{n}_j)$  dans lesquelles  $0 \leq j \leq k$ , et  $p$  est un prédicat observationnel.*

De façon équivalente, étant donné un prédicat théorique  $k$ -aire  $q$ , la formule atomique  $q(x_1, \dots, x_n)$  est une abréviation pour la combinaison booléenne

$\varphi$  des formules de la forme  $p(x)$ , où  $p$  est un prédicat observationnel, et  $x$  est l'un des  $x_1, \dots, x_n$ ;  $\varphi$  tient donc lieu de définition pour  $q$ . Notons que  $\varphi$  pourrait contenir une sous-formule qui définirait elle-même un autre prédicat théorique simplifiant la définition de  $q$ . Cette dernière ne serait donc plus composée uniquement de prédicats observables, mais également de prédicats théoriques, à conditions qu'ils soient eux-mêmes définis. Intuitivement, les prédicats théoriques définissent des relations entre les descripteurs des cartes, qui ne dépendent que de ce que ces cartes sont ou ne sont pas, et qui ne dépendent d'aucune autre carte qui pourrait apparaître ou non dans la séquence.

De plus, même si nous acceptons et soulignons l'utilisation d'un vocabulaire riche, nous devons garantir que toute séquence possible ou classe de séquence possible respecte l'ordre naturel sur les séquences de cartes. En d'autres termes, si nous devons énoncer que la dixième carte est une dame de cœur, ceci ne sera fondé que sur cette dixième carte ou l'une des cartes *la précédant*, en aucun cas sur une carte apparaissant éventuellement dans *la suite* de la séquence. Pour cette raison, nous nous limiterons à l'emploi d'un certain type de programmes logiques : les programmes logiques *itératifs*.

### 4.1.3 Programmes Logiques Univoques et Équivoques

Définissons tout d'abord la notion générale de *programme logique* :

**Définition 4.3** *Un programme logique est un ensemble fini de règles (définies sur un vocabulaire décidable) dont les têtes sont des formules atomiques, et telles que toutes les variables libres apparaissant dans le corps d'une règle apparaissent également dans sa tête.*

Toutes les variables libres de toute règle d'un programme logique sont quantifiées universellement de façon implicite. Il pourrait sembler trop restrictif de n'être capable de générer que des atomes. L'utilisation des prédicats théoriques nous permet de contourner cette restriction apparente. Supposons par exemple que nous voulions prédire que la  $n^{ième}$  carte est soit un cœur (*hearts*) soit un carreau (*diamonds*), en fonction des cartes précédentes. Nous ne pouvons pas utiliser une règle dont la tête est  $hearts \vee diamonds$ . Nous pouvons en revanche utiliser un vocabulaire contenant le prédicat théorique *red* défini comme précédemment par son sens commun, et utiliser ce prédicat dans la tête de notre programme logique.

Nous définissons désormais la notion de programme logique *itératif* afin de contraindre le champ d'application des règles à l'ordre naturel des cartes, comme nous l'avons présenté plus tôt :

**Définition 4.4** *Un programme logique  $\mathcal{P}$  est dit itératif ssi les conditions suivantes sont respectées : considérons une règle  $r \in \mathcal{P}$  dont  $\wp(x)$  est la tête et  $\Psi$  le corps (dans lequel n'apparaît aucune variable libre excepté  $x$  éventuellement). Soient  $n \in \mathbb{N}$  donné et  $\mathcal{G}$  une séquence possible. Alors la valeur de vérité de  $\Psi[\bar{n}/x]$  dans  $\mathcal{G}$  est déterminée par la valeur de vérité de toutes les formules de la forme  $\wp(\bar{m})$ , où  $m < n$  et  $\wp$  est un symbole de prédicat.*

Par exemple, un programme logique itératif peut contenir la règle suivante, qui exprime qu'une carte est un pique à condition qu'elle soit un roi et qu'elle soit précédée d'au moins un cœur :

$$[\exists y(y < x \wedge \text{hearts}(y)) \wedge \text{king}(x)] \rightarrow \text{spades}(x)$$

Soit  $\mathcal{P}$  un programme logique, nous dénotons par  $[\mathcal{P}]$  l'ensemble des atomes clos<sup>13</sup> qui peuvent être générés à partir de  $\mathcal{P}$ . Une propriété importante des programmes logiques itératifs est alors énoncée de la façon suivante :

**Propriété 4.5** *Soit  $\mathcal{P}$  un programme logique itératif défini sur un vocabulaire  $\mathcal{V}$  (décidable), alors  $[\mathcal{P}]$  est un ensemble récursivement énumérable.*

Cette propriété vient du fait que :

- les formules atomiques construites à partir d'un prédicat théorique peuvent être remplacées par des formules construites à partir de prédicats observationnels seulement ;
- pour chaque règle  $\wp \rightarrow p(x)$  d'un programme logique itératif et chaque  $n \in \mathbb{N}$ ,  $\wp[\bar{n}/x]$  peut être transformé de façon mécanique en un énoncé dépourvu de quantificateurs, et dont tous les termes appartiennent à  $\{\bar{m} \mid m \leq n\}$ .

Puisque  $[\mathcal{P}]$  n'est constitué que d'énoncés atomiques, il est toujours cohérent. Il peut toutefois être contradictoire si aucune interprétation ne peut lui être associée. Par exemple, pour un  $n \in \mathbb{N}$ ,  $[\mathcal{P}]$  peut contenir deux formules distinctes de la forme  $(\wp_1(\bar{n})$  et  $\wp_2(\bar{n}))$ , avec  $\wp_1$  et  $\wp_2$  appartenant tous deux à  $\text{Prd}_{\text{rank}}$  ou  $\text{Prd}_{\text{suit}}$ .  $\mathcal{P}$  peut également être en contradiction au vu d'une interprétation particulière, par exemple s'il contient  $\text{red}(\bar{10})$  et  $\text{spades}(\bar{10})$  dans notre exemple.

Afin d'achever cette formalisation, nous distinguons les programmes univoques des programmes équivoques : un programme logique est dit *équivoque* s'il possède un modèle dans la classe des séquences possibles  $\mathcal{W}$ , et il est dit *univoque* ssi il possède exactement un modèle dans  $\mathcal{W}$ .

**Propriété 4.6** *À la fois l'univocité et l'équivocité d'un programme logique sont co-semi-décidables.*

13. un atome clos est un atome dans lequel n'apparaît aucune variable libre.

À des fins d'illustration, décrivons désormais une classe de programmes logiques univoques. Soit  $<_{suit}$  et  $<_{rank}$  deux relations d'ordres arbitraires sur  $Prd_{suit}(\mathcal{V})$  et  $Prd_{rank}(\mathcal{V})$  respectivement. Par exemple,

- $hearts <_{suit} diamonds <_{suit} clubs <_{suit} spades$ ,
- $ace <_{rank} two <_{rank} \dots <_{rank} ten <_{rank} jack <_{rank} queen <_{rank} king$ .

Tout programme logique itératif de la forme suivante vérifie la propriété d'univocité :

$$\begin{aligned} \varphi_1(x) &\rightarrow hearts(x) \\ \neg\varphi_1(x) \wedge \varphi_2(x) &\rightarrow diamonds(x) \\ \neg\varphi_1(x) \wedge \neg\varphi_2(x) \wedge \varphi_3(x) &\rightarrow clubs(x) \\ \neg\varphi_1(x) \wedge \neg\varphi_2(x) \wedge \neg\varphi_3(x) \wedge \varphi_4(x) &\rightarrow spades(x) \end{aligned}$$

et de façon similaire pour les prédicats de  $Prd_{rank}(\mathcal{V})$ .

Maintenant que j'ai défini un cadre logique simple pour définir une certaine réalité et la notion de théorie, je vais illustrer les différences entre les paradigmes d'apprentissage cités au Chapitre 2, afin de bien situer le contexte de la découverte scientifique.

## 4.2 Apprentissage Passif

La notion de programme univoque nous suffit pour illustrer le problème de l'identification à la limite, qui correspond au jeu suivant.

Le maître du jeu choisit une séquence possible  $\mathcal{G}$  équivalente à un certain programme univoque. À chaque tour  $n$ , le maître du jeu révèle la carte  $\bar{n}$ , de telle sorte que l'apprenti  $\mathcal{L}$  les découvre une par une. Dans ce contexte, le but de l'apprenti est d'identifier, parmi un ensemble infini de programmes univoques  $\mathcal{H} = \{\mathcal{P}_0, \mathcal{P}_1, \dots\}$ , une hypothèse  $\mathcal{P}_{\mathcal{G}}$  équivalente à  $\mathcal{G}$ .

Après chaque nouvelle carte  $\bar{n}$ ,  $\mathcal{L}$  change donc éventuellement d'avis et propose, s'il est cohérent, un programme  $\mathcal{P}_{\mathcal{H}} \in \mathcal{H}$  qui décrit exactement la sous-séquence  $\bar{0}, \bar{1}, \dots, \bar{n}$ , *i.e.*  $[\mathcal{P}_{\mathcal{H}}] \supseteq \{\wp_1(\bar{0}), \wp_2(\bar{0}), \dots, \wp_1(\bar{n}), \wp_2(\bar{n})\}$ ,  $\wp_1 \in Prd_{suit}(\mathcal{V})$ ,  $\wp_2 \in Prd_{rank}(\mathcal{V})$ , jusqu'à converger éventuellement vers une hypothèse correcte.

Notons que cette forme d'apprentissage appartient au paradigme de l'apprentissage de fonctions, pour lequel présenter des données positives seulement est équivalent à proposer à la fois des données positives et négatives, puisque ces dernières peuvent être retrouvées à partir des premières.

**Proposition 4.7** *La classe des programmes univoques est identifiable à la limite.*

Ceci est une conséquence de la Propriété 4.6 qui permet d'appliquer la technique d'énumération classique.

Chaque hypothèse est donc réfutable à la limite, et chaque tour de jeu peut potentiellement invalider l'hypothèse courante. D'un autre côté,  $\mathcal{L}$  n'aura à aucun moment dans le jeu une preuve que son hypothèse courante est correcte.

Une stratégie acceptable pour ce jeu serait donc d'ordonner de façon arbitraire l'ensemble des hypothèses et de choisir, chaque fois qu'une nouvelle information invalide son hypothèse, la prochaine qui soit cohérente avec les connaissances courantes concernant  $\mathcal{G}$ .

Donnons un exemple de jeu. Soit  $\mathcal{G}$  une séquence possible alternant le quatre de pique avec la dame de cœur, en commençant par le quatre de pique, jusqu'à la 120<sup>e</sup> carte, après laquelle la dame de cœur est répétée indéfiniment. Soit  $\mathcal{H}$  l'ensemble d'hypothèses contenant les hypothèses suivantes :

- $\mathcal{P}_0$  : toute carte dans la séquence est un quatre de pique,
- $\mathcal{P}_1$  : toute carte dans la séquence est une dame de cœur,
- $\mathcal{P}_2$  : la séquence est une alternance de quatre de pique et de dames de cœur, commençant par le quatre de pique,
- $\mathcal{P}_3$  : La séquence commence par 60 répétitions de la sous-séquence « quatre de pique, dame de cœur », puis consiste en une répétition infinie de la dame de cœur.
- ...

Après la première carte,  $\mathcal{L}$  peut proposer n'importe quel programme dans  $\mathcal{H}$  sauf  $\mathcal{P}_1$ , par exemple  $\mathcal{P}_0$ . Après que la seconde carte a été révélée,  $\mathcal{P}_0$  n'est plus une hypothèse cohérente et  $\mathcal{L}$  doit *changer d'avis* et proposer  $\mathcal{P}_2$  ou  $\mathcal{P}_3$ , par exemple  $\mathcal{P}_2$ . On dit que le processus d'identification converge si  $\mathcal{L}$  change d'hypothèses un nombre fini de fois avant de proposer  $\mathcal{P}_3$ .

### 4.3 Apprentissage Actif

Nous avons illustré le problème de l'identification à la limite avec un jeu dans lequel une séquence de cartes infinie est présentée, carte après carte, à l'apprenti qui doit identifier un programme *univoque* décrivant cette séquence. En opposition à ce protocole passif, l'apprentissage à partir de requêtes permet à l'apprenti de poser des questions à un oracle pour tester ses hypothèses. Le problème est ici un problème de classification dans lequel l'apprenti doit identifier un programme *équivoque* (et non plus un programme univoque).

Soit  $P_{target}$  un programme logique équivoque décrivant un ensemble  $\mathcal{W}_{target} \subseteq \mathcal{W}$  de séquences possibles partageant certaines propriétés,  $\mathcal{W}$  étant l'ensemble de toutes les séquences possibles, et soit  $\mathcal{H}$  un ensemble potentiellement infini de programmes équivoques.

À chaque tour de jeu,  $\mathcal{L}$  peut poser une question à un oracle capable de répondre aux différents types de requêtes présentés au Chapitre 2 :

- *Appartenance* : l’oracle prend en entrée une séquence possible  $X \in \mathcal{W}$ , et répond « vrai » si  $X \in \mathcal{W}_{target}$ , ou « faux » si  $X$  est un contre-exemple.
- *Équivalence* : l’oracle prend en entrée un ensemble  $\mathcal{W}_{\mathcal{H}} \subseteq \mathcal{W}$  de séquences possibles, et répond soit « vrai » si  $\mathcal{W}_{\mathcal{H}} \equiv \mathcal{W}_{target}$ , soit un contre-exemple  $X$  tel que  $X \subseteq \mathcal{W}_{\mathcal{H}} \Delta \mathcal{W}_{target}$ .
- *Subset* : l’oracle prend en entrée un ensemble  $\mathcal{W}_{\mathcal{H}} \subseteq \mathcal{W}$  de séquences possibles, et répond soit « vrai » si  $\mathcal{W}_{\mathcal{H}} \subseteq \mathcal{W}_{target}$ , soit un contre-exemple  $X$  tel que  $X \in \mathcal{W}_{\mathcal{H}} - \mathcal{W}_{target}$ .
- *Superset* : l’oracle prend en entrée un ensemble  $\mathcal{W}_{\mathcal{H}} \subseteq \mathcal{W}$  de séquences possibles, et répond soit « vrai » si  $\mathcal{W}_{\mathcal{H}} \supseteq \mathcal{W}_{target}$ , soit un contre-exemple  $X$  tel que  $X \in \mathcal{W}_{target} - \mathcal{W}_{\mathcal{H}}$ .

La classification est considérée comme réussie si après un nombre fini de requêtes,  $\mathcal{L}$  converge vers un programme  $\mathcal{P}_{\mathcal{H}} \in \mathcal{H}$  tel que  $\mathcal{P}_{\mathcal{H}} \equiv \mathcal{P}_{target}$ .

Lorsque l’on utilise des objets infinis, comme c’est le cas dans notre définition des séquences possibles (Définition 4.1), toutes ces requêtes sont co-semi-décidables, c’est-à-dire que pour chaque requête, le processus de vérification ne s’arrêtera pas si la requête est correcte. En revanche, si un contre-exemple existe, alors il sera trouvé au bout d’un temps fini lors de ce processus. Les requêtes d’appartenance, en particulier, ne sont pas adaptées, puisqu’on ne peut décider si un programme équivoque est effectivement un programme univoque, c’est à dire, décrit une unique séquence possible. Supposer l’existence d’un oracle capable d’y répondre requerrait donc de sa part des capacités supérieures à celles d’une machine de Turing. Ce problème est représentatif du contexte de la découverte scientifique dans lequel la nature peut être considérée comme « silencieuse » puisqu’elle ne peut répondre aux requêtes de l’apprenti, et ne peut donc être considérée comme un oracle dans ce contexte. De plus, l’utilisation de tels objets infinis est nécessaire pour simuler ce contexte de découverte, afin de représenter le fait que chaque objet étudié peut avoir une description potentiellement infinie (voir Chapitre 1).

En tenant compte du problème soulevé, nous présentons en Section 4.5 une extension multi-agent de ce paradigme dans laquelle les apprentis n’ont pas accès à un tel oracle mais apprennent les uns des autres.

## 4.4 Qu’en est-il de l’apprentissage *PAC* ?

Dans le cas de l’apprentissage passif, une loi de distribution peut être définie pour déterminer la probabilité d’observer chaque élément d’information. Comme nous l’avons vu, [Valiant, 1984] a introduit une notion d’approxima-

tion pour refléter le fait que dans ce contexte, l'apprentissage peut ne pas être exact, soit à cause de la présence de bruit dans ces données, soit simplement parce que l'identification exacte n'est pas requise. Il définit ensuite un critère de succès fondé sur ces notions de distribution, d'approximation, et sur un critère de confiance. Celà étant, l'apprentissage *PAC* implique des notions bornées arbitrairement, alors que les modèles de réalité que nous avons défini en Section 4.1 sont infinis. Pour cette raison, la formalisation proposée ne semble pas adaptée au cadre *PAC*. Pourtant, il me semble intéressant de décrire les restrictions qu'il faut imposer à celle-ci pour illustrer l'apprentissage *PAC* qui est largement utilisé dans la communauté. De plus, ces restrictions sont comparables à celles que nous avons imposées lors des expériences présentées au Chapitre 7 dans un souci de simplification.

Considérons la distribution de probabilité naturelle  $\mathcal{D}$  sur  $\mathcal{W}$ , selon laquelle pour chaque atome clos  $\varphi_1$  construit à partir d'un symbole de prédicat de  $Prd_{suit}(\mathcal{V})$ , et chaque atome clos  $\varphi_2$  construit à partir d'un symbole de prédicat de  $Prd_{rank}(\mathcal{V})$ ,  $p(\varphi_1) = \frac{1}{4}$  et  $p(\varphi_2) = \frac{1}{13}$ . Pour chaque carte  $n \in \mathbb{N}$ , on a donc  $p(\varphi_1(\bar{n}) \wedge \varphi_2(\bar{n})) = \frac{1}{52}$ , ce qui peut servir de base à une distribution de probabilités sur  $\mathcal{W}$ , de façon similaire à la mesure de probabilités communément utilisée sur un espace de Cantor, si ce n'est que dans notre cas, nous avons à faire à un arbre possédant des ramifications à 52 branches au lieu de 2. On peut ainsi calculer, pour chaque paire  $(\mathcal{P}_1, \mathcal{P}_2)$  de programmes logiques équivoques, la probabilité qu'une séquence possible  $G$  donnée soit décrite par  $\mathcal{P}_1$  mais pas par  $\mathcal{P}_2$ , ou inversement. En d'autres termes cette distribution nous permet de déterminer un degré de similarité entre deux programmes logiques équivoques, ce qui nous sera utile pour définir le critère d'approximation.

Dans la formalisation proposée par Valiant, les exemples finis sont des points dans un espace à  $n$  dimensions, où  $n$  correspond au nombre de termes utilisés pour décrire ces exemples. Les concepts sont des ensembles finis de tels points, et les exemples sont classés en fonction de leur appartenance (ou non) au concept cible. En bornant notre formalisation à de telles notions finies, n'importe quel ensemble de segments initiaux de séquences possibles, de longueur arbitraire  $n$ , représente une classe naturelle de concepts. On peut alors considérer un ensemble fini  $X$  de séquences de cartes de longueur  $n$ , choisies aléatoirement selon  $\mathcal{D}$ , comme étant l'ensemble d'exemples. Si l'on définit la taille  $C$  de  $\mathcal{P}_{target}$  comme fonction du nombre de formules dans  $\mathcal{P}_{target}$  et de la taille maximale de ces formules, on a alors tous les paramètres nécessaires pour travailler dans un paradigme similaire à celui de l'apprentissage *PAC*.

Tout d'abord, choisissons un entier naturel  $n$ , puis pour chaque  $i$  tel que  $0 < i < n$ , choisissons aléatoirement  $\wp_1(\bar{i})$  et  $\wp_2(\bar{i})$  conformément à  $\mathcal{D}$ , avec  $\wp_1 \in Prd_{suit}$  et  $\wp_2 \in Prd_{rank}$ , de façon à laisser apparaître un ensemble  $S$  de

formules décrivant ces segments initiaux.

Le but de l'apprenti est alors de proposer, après avoir pris connaissance de l'ensemble d'exemples, un programme logique  $\mathcal{P}_H$  possiblement équivoque pour lequel presque tout  $x \in X$  est un modèle *i.e.*, un programme qui soit une approximation acceptable de  $\mathcal{W}_{target}$ , toujours conformément à  $\mathcal{D}$ , comme définis ci-dessous :

$$\text{error}(h) = \sum_{x \in \mathcal{W}_H \Delta \mathcal{W}_{target}} \mathcal{D}(\S),$$

où  $\mathcal{W}_H \Delta \mathcal{W}_{target}$  représente la différence symétrique entre la classification cible et celle de l'apprenti. Soit  $\varepsilon$  le critère d'approximation, et  $\delta$  le critère de confiance, la classification est considérée comme probablement approximativement correcte si  $\mathcal{L}$  propose, en un temps  $t$  et avec une probabilité  $\mu \geq 1 - \delta$ , un programme  $\mathcal{P}_H$  ayant un taux d'erreur  $\text{error}(\mathcal{P}_H) < \varepsilon$ , et tel que le temps  $t$  soit polynomial en la taille du programme cible, le nombre d'exemples,  $\varepsilon$ , et  $\delta$  :  $t = p(n, C, \frac{1}{\varepsilon}, \frac{1}{\delta})$ .

## 4.5 Apprentissage Interactif et Validation Sociale de Découvertes

Un scientifique ne travaille pas de façon isolée. Selon [Popper, 1963], la science se pratique au sein d'une communauté de chercheurs qui interagissent en échangeant des données et des théories en publiant et réfutant leurs conjectures. Les théories devraient ainsi être évaluées en fonction de leur capacité à résister à la publication. Une publication représente donc un état de l'art, une théorie courante ou une solution à un problème. Celle-ci est acceptée par la communauté jusqu'à ce qu'elle devienne insuffisante pour expliquer la nature et qu'elle soit réfutée. Nous incluons cet aspect important de la découverte scientifique dans notre protocole pour contourner les problèmes soulevés en Section 4.3 : la science est un processus limitant impliquant une communauté d'agents.

### 4.5.1 Restriction des Requêtes d'Appartenance

En s'appuyant sur le fait qu'un chercheur réalise ses observations à partir d'expériences finies et délimitées dans le temps, ne percevant ainsi qu'une partie de la réalité, nous proposons de restreindre l'utilisation des requêtes d'appartenance à des séquences finies. Les joueurs peuvent ainsi mener des *expériences* en demandant au maître du jeu si une séquence particulière  $X$  est cohérente avec  $\mathcal{P}_{target}$  ou non. Un résultat positif ne garantit pas que  $X$  ne soit pas un segment initial d'une séquence possible n'appartenant pas à  $[\mathcal{P}_{target}]$ ,

mais un résultat négatif garantit qu'aucune séquence possible ne peut être créée à partir de  $X$ , et invalide donc l'ensemble des hypothèses cohérentes avec  $X$ . La différence principale avec l'identification à la limite est donc l'utilisation d'expériences permettant à un apprenti d'explorer l'ensemble des séquences finies à sa guise, et d'employer éventuellement une stratégie d'exploration, des heuristiques, ou des connaissances préalables. Contrairement au contexte de l'identification à la limite, on ne présente pas à l'apprenti une unique réalité, de façon passive. L'apprenti définit lui-même ce qu'il souhaite observer, et ses biais de modélisation sont répercutés directement sur ses expériences. Contrairement au cadre *PAC*, seuls les exemples négatifs résultant de l'expérience sont exploitables pour réfuter, alors que les exemples positifs ne permettent que d'émettre de nouvelles hypothèses, et peuvent être remis en cause par la suite.

### 4.5.2 Une Dynamique Sociale de Réfutation à la Limite

Au lieu de supposer l'existence d'un oracle capable de répondre à des requêtes co-semi-décidables, nous préférons supposer que les apprentis peuvent publier leurs conjectures afin de demander à leur pairs de les réfuter. La publication d'une conjecture ne garantit pas qu'une réponse sera apportée, et cette requête reste co-semi-décidable : aucune réponse positive ne sera jamais apportée. Comme dans l'identification à la limite, un apprenti n'a donc jamais de certitude quant à la validité de sa conjecture. Toutefois, si cette conjecture est fautive, alors un contre-exemple apparaîtra après un temps fini. La distribution de la résolution de requêtes ne sera pas moins efficace que sa résolution par un oracle seul, mais pourrait être fortement accélérée dans le cas d'une grande communauté d'apprentis. Pour assurer que les conjectures incorrectes seront réfutées, nous créons une dynamique sociale comme cela est fait en théorie des jeux depuis [Neumann & Morgenstern, 1944]. Notons que plusieurs programmes cibles peuvent être mis à disposition des apprentis pour poser des problèmes stratégiques d'exploration de l'espace d'hypothèse ou de politique de publication, comme cela est présenté aux Chapitres 6 et 7.

Nous symbolisons le produit de l'interaction sociale par une fonction de gain, en attribuant ou en ôtant des points pour chaque requête en fonction de la réponse fournie par la communauté (réfutation ou non), de façon à pouvoir créer une atmosphère collaborative ou compétitive parmi les apprentis. Cette fonction de gain motive les apprentis à chercher les éventuels contre-exemples aux conjectures publiées par leurs pairs afin d'assurer que ces conjectures seront réfutées à la limite ou qu'elles demeureront en tant que références consensuelles et gagneront en crédibilité et en importance.

Le tableau 4.1 offre une vision synthétique des points discutés dans ce

	Mode d'interaction	Modèle de réalité	Durée du processus	Type d'apprentissage
Identification à la Limite	Passif	Infini	Non-bornée	Exact [Réalité] (simulation)
PAC	Passif	Fini	Finie	Approximatif [Réalité] (simulation)
Apprentissage par requêtes	Actif	Infini	Bornée	Exact [Réalité] (simulation)
Learning from different teachers	Actif	Infini	Bornée	Exact [Tuteur] (Imitation)
Apprendre les uns des autres	Actif / Interactif et Participatif	Infini mais biaisé	Non-bornée	Exact [pairs] (consensus)

TABLE 4.1 – Synthèse des différences entre paradigmes d'apprentissage

chapitre.

La colonne « temps de réflexion » concerne le temps laissé à l'apprenti pour formuler son hypothèse. Le fait qu'il soit polynomial dans le cas de l'apprentissage *PAC* représente le fait que le temps laissé à l'apprenti est défini *a-priori* en fonction du nombre d'exemples, du critère de confiance et du taux d'approximation comme définit au chapitre 2, alors que dans les autres paradigmes, aucune contrainte n'est imposée. Dans la colonne « durée du processus », je fais une distinction entre un temps fini déterminé en fonction du nombre d'exemples et un temps borné, c'est à dire fini mais inconnu *a-priori*. On constate l'évolution du rôle passif de l'apprenti à un rôle actif, et le caractère interactif définissant le protocole proposé semble naturel au regard de cette évolution. Cependant, il peut également être qualifié de participatif, puisque la découverte d'un contre-exemple est dépendante de la *motivation* des apprentis. Le modèle de réalité est caractérisé d'infini mais biaisé, ce qui exprime le fait qu'aucune restriction n'est posée à ce sujet. Cependant, les représentations de l'apprenti seront biaisées par son modèle de description si celui-ci n'est pas correct. Enfin, le critère d'apprentissage peut sembler surprenant. Dans tous les autres paradigmes, une entité connaît la règle cachée et peut déterminer si l'apprentissage est correct ou non. [Angluin & Krikis, 2003] définit un apprentissage vis-à-vis du tuteur, c'est à dire que la solution adoptée par l'apprenti est supposée équivalente à celle du tuteur, même si celle-ci n'est pas optimale. Dans le cadre de la découverte scientifique, il n'existe aucune entité possédant la connaissance absolue, et personne n'est donc en mesure de décider si l'apprentissage est correct. Le critère de réussite est donc l'atteinte d'un consensus

au sein de la communauté d'apprentis.

## 4.6 Perspectives

Le jeu *Éleusis+Nobel* présenté au Chapitre 5, et qui est le sujet des chapitres suivants, est une illustration parfaite de ce formalisme. Toutefois, je me suis limité ici à présenter une interaction fondée sur la publication et la réfutation de conjectures, et il me semble intéressant d'introduire la diversité des jugements définis au Chapitre 3 afin de mettre en oeuvre un véritable débat scientifique au sein d'une communauté d'apprentis. Cette perspective de travail est tout à fait réaliste et trouverait des applications concrètes sous la forme d'outil de collaboration au sein d'un groupe de travail, ou sous la forme d'une plate-forme de débats publiques, afin de proposer une alternative aux classiques forums de discussions qui deviennent trop souvent de véritables labyrinthes d'informations. Cette perspective représente déjà un projet concret, *Intermed*, auquel je m'associerai très prochainement.

L'introduction d'un niveau social dans les paradigmes d'apprentissage ouvre la voie à l'expérimentation de différentes fonctions de gains afin de déterminer dans quelles conditions une communauté d'apprentis converge le plus efficacement vers une solution acceptable. Nous avons proposé dans [Dartnell et Sallantin, 2005] d'associer un gain  $P = 1$  à la publication d'une conjecture, et un transfert de  $R = 2$  points de l'auteur d'une publication réfutée vers l'auteur de la réfutation. L'interaction avec les autres apprentis est donc nécessaire pour acquérir des points, mais elle est associée à une prise de risque directement liée au rapport  $\frac{P}{R}$ , et il appartient à chaque apprenti de gérer cette prise de risque. Nous verrons au Chapitre 7 que même si cette fonction de gain est très simple, elle permet d'observer chez des joueurs humains des comportements variés assurant une dynamique sociale (altruisme, prudence, opportunisme, etc.). Pourtant, des simulations automatisées doivent être menées pour vérifier que ce protocole est réellement adapté à un apprentissage efficace, et je reviendrais sur ce sujet en Section 6.3.

Chaque apprenti devant optimiser ses propres gains, cette fonction de gain ouvre également la voie vers des problèmes de gestion du risque et de stratégies d'exploration, mais nous n'explorerons pas ces pistes dans ce document.

Enfin, l'inspiration multi-agent pourrait être poursuivie afin de matérialiser une résolution distribuée de problèmes dans un espace à trois dimensions. En représentant l'espace des hypothèses comme un espace physique, et en modifiant l'altitude de l'espace dans les zones soumises à des conflits d'opinion, on pourrait ainsi maintenir une cartographie de la résolution de problème. Outre l'utilité d'un tel visuel, cette cartographie pourrait éventuellement être utilisée

pour guider une résolution de problèmes par des techniques courantes de la communauté mutli-agents, telles que les algorithmes de type fourmi, afin de guider des agents réactifs se déplaçant au gré des pentes vers les zones basses représentant le manque d'information ou les conflits d'opinions.

## 4.7 Synthèse

J'ai proposé dans ce chapitre une formalisation logique simple mais suffisante pour représenter les difficultés liées au contexte de découverte dans lequel nous nous plaçons. Cette formalisation du problème m'a permis de préciser les différences qui existent entre les principaux paradigmes d'apprentissage, et de positionner de façon plus claire les extensions que je propose afin de rendre ces paradigmes opérationnels dans un tel contexte, sous la forme d'un apprentissage interactif effectué par une communauté d'apprentis. Comme le propose Popper, ce protocole place la preuve expérimentale hors de portée des apprentis, et s'oriente vers la réfutation à la limite des conjectures et l'accréditation sociale de celles qui résistent à la réfutation. J'ai présenté plusieurs perspectives de travail, à plus ou moins long terme, parmi lesquelles figurent bien sur l'introduction des jugements modaux définis au Chapitre 3 pour diversifier l'interaction et les formes de publication. Je n'ai malheureusement pas encore exploité cette formalisation comme je l'aurais souhaité, par manque de temps. Elle est cependant suffisamment élaborée pour caractériser un prototype de plate-forme de résolution distribuée.

La seconde partie de ce document présente les aspects pratiques liés à la définition de ce prototype, et en particulier comment mes choix de modélisation m'ont amené à définir un modèle *AGR* de ce protocole d'apprentissage interactif.

Deuxième partie  
Aspects Pratiques



# Chapitre 5

## Présentation du Prototype

### *Éleusis+Nobel*

*« Il semble que le savoir scientifique soit toujours essayé, toujours contrôlé, toujours critiqué... en critiquant sans cesse son propre savoir »*

G. Bachelard.

LA PREMIÈRE partie de ce document m'a permis de traiter les aspects théoriques de ma recherche. Je vais désormais présenter des aspects plus pratiques : la modélisation et l'implantation d'un prototype, *Éleusis+Nobel* (Chapitre 6), et les résultats obtenus grâce à ce dernier (Chapitre 7). Ce prototype, que j'ai développé tout au long de ma thèse, a servi de support concret aux formalismes proposés et a permis de concentrer les efforts des différents collaborateurs. Je vais donc avant toute chose décrire dans ce chapitre comment l'idée originale a vu le jour.

*Éleusis+Nobel* est le fruit d'un travail pluri-disciplinaire débuté avec Sylvain Charron et David Chavalarias<sup>14</sup>, sur les bases du mémoire de DEA de ce dernier [Chavalarias, 1997]. Les aspects conceptuels et formels présentés en première partie de ce document ont pu être approfondis grâce à l'aide précieuse de Dominique Luzeaux<sup>15</sup>, Éric Martin<sup>16</sup> et Jean Sallantin<sup>17</sup>, et vont maintenant être illustrés. Enfin, la récente exploitation de ce jeu par Hélène Hagège<sup>18</sup>

---

14. Centre de Recherche en Épistémologie Appliquée, UMR 7656, École Polytechnique/CNRS

15. Centre Technique des Systèmes d'Information, Arcueil, France

16. School of Computer Science and Engineering, University of New South Wales, Sydney, Australie

17. Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, Université Montpellier II, France

18. Laboratoire Interdisciplinaire de Recherche en Didactique Éducation et Formation,

lors de simulations a permis de valider ses intérêts épistémiques et didactiques. Les résultats encourageants de ces simulations nous laissent espérer que ce prototype pourra servir de base à des outils pédagogiques, à des plates-formes de travail collaboratif, ou sera décliné sur des thèmes plus concrets que celui des cartes pour créer d'autres jeux sérieux.

Imaginé afin de reproduire certains processus de la découverte scientifique, tels que la génération et la validation d'hypothèses, l'exploration de l'espace de recherche par la mise au point d'expériences, la formulation de théories ainsi que leur publication ou leur réfutation, *Éleusis+Nobel* est la fusion naturelle des deux jeux *Éleusis* [Gardner, 1959] et *Nobel* [Chavalarias, 1997] que je vais présenter en sections 5.1 et 5.2.

En section 5.3, je présenterai le jeu en lui-même, en faisant le lien avec le formalisme présenté au Chapitre 4. Je fournirai également quelques pistes concernant l'étude des systèmes complexes. En effet, cette conceptualisation des étapes nécessaires à la création et à l'évolution de théories scientifiques intègre certaines des problématiques majeures des systèmes complexes :

- l'interaction entre plusieurs agents,
- la construction d'un langage partagé et adapté à la description des phénomènes observés ainsi qu'aux différents niveaux de représentation nécessaires à leur explication,
- les problèmes de bifurcations et de concurrence liés aux différents comportements possibles du système dans chacun de ses états.

## 5.1 Eleusis

Le texte ci-dessous est inspiré de l'article de Martin Gardner [Gardner, 1959] dont la version française est parue dans « La mathématique des jeux », Bibliothèque pour la science, et de la pages web personnelle de Thomas Kauffmann [Kauffmann, 1999].

### 5.1.1 Les origines du jeu

*Éleusis* a été inventé en 1956 par Robert Abbott, alors qu'il était étudiant à l'Université du Colorado. L'étude de ce jeu a apporté des vues nouvelles et profondes dans la connaissance de ce que les psychologues appellent parfois l'effet « Haha », l'« Euréka » d'Archimède. Comme nous l'avons vu en Introduction, de grands bouleversements dans la science s'articulent souvent autour de tels bonds de la compréhension, dont la genèse reste mystérieuse. *Éleusis* s'est révélé une simulation étonnante de cet aspect épistémologique, bien qu'

Abbott ne l'a pas inventé dans ce but. En 1963, les règles complètes du jeu de Abbott ont paru dans son livre *Abbott's New Card Games* (volume relié, édité par Stein et Day, broché par Funk et Wagnalls).

Martin Kruskal, brillant physicien de l'Université Princeton, s'est intéressé à ce jeu et en a amélioré différents aspects [Kruskal, 1979]. De nombreux professeurs de lycées américains utilisèrent alors ce jeu pour illustrer la méthode scientifique. Puis des chercheurs en intelligence artificielle écrivirent des programmes d'ordinateur pour y jouer, et Abbott reçut des lettres du monde entier suggérant des améliorations pour rendre le jeu plus attrayant.

En 1973, Abbott travailla à l'analyse de ce jeu avec John Jaworsky, un jeune mathématicien anglais qui avait élaboré un programme d'*Éleusis* utilisé pour enseigner la méthode inductive. Sur la base de cette analyse, Abbott entreprit de refondre le jeu pour lui apporter le plus d'améliorations possibles. La nouvelle version, dont la mise au point prit trois ans, n'est pas seulement plus intéressante, elle est aussi beaucoup plus riche sur le plan philosophique. Avec l'introduction du Sage et du mauvais prophète, ce jeu simule désormais la recherche de n'importe quelle forme de vérité. Voici donc, à partir d'informations communiquées par Abbott, les règles du nouveau jeu d'*Éleusis* tel que le pratiquent aujourd'hui les aficionados et dont nous nous sommes inspiré.

## 5.1.2 Les règles du jeu

*Éleusis* se joue à quatre joueurs au moins ; à partir de huit joueurs, le jeu devient trop long et confus. On utilise deux jeux ordinaires de 52 cartes battus ensemble. Une partie consiste en une ou plusieurs donnes distribuées par des donneurs différents. Le donneur peut être appelé Dieu, la Nature, Tao, Brahma, l'Oracle (comme à Delphes) ou tout simplement le Donneur.

### 5.1.2.1 La règle secrète

Le premier travail du donneur est d'élaborer une « règle secrète ». C'est tout simplement une règle définissant la suite des cartes « correctes » qui doivent être jouées au cours d'une partie. Le jeu consiste pour les autres joueurs à découvrir cette règle. Plus vite un joueur trouve la règle plus sa marque est élevée.

La caractéristique la plus subtile du jeu d'*Éleusis* est le décompte des points décrit plus loin ; ce décompte avantage le donneur qui a su inventer une règle ni trop facile, ni trop difficile à deviner. Sans ce garde-fou, le donneur serait tenté de fabriquer des règles si complexes que personne ne pourrait les découvrir, ce qui rendrait le jeu ennuyeux et frustrant. Un exemple de règle trop simple est « jouer une carte de couleur différente de celle précédemment jouée » ;

l'alternance des couleurs serait immédiatement évidente. Une meilleure règle, « jouer de façon que la hauteur de la carte jouée soit alternativement un nombre premier ou un nombre non premier » serait peut-être encore trop simple pour un mathématicien ; pour un profane, elle risquerait d'être trop difficile. Un exemple de règle trop compliquée est « multiplier les nombres des trois dernières cartes jouées puis diviser par 4. Si le reste est 0, jouer une carte rouge ou supérieure au 6. Si le reste est 1, jouer une carte noire ou une figure. Si le reste est 2, jouer une carte paire ou une carte inférieure au 6. Si le reste est 3, jouer une carte impaire ou un 10. » Personne ne trouvera une telle règle et la marque du donneur sera faible. Voici trois exemples de bonnes règles pour jouer avec des débutants.

1. Si la dernière carte correcte jouée est impaire, jouer une carte noire ; dans le cas contraire, jouer une carte rouge.
2. Si la dernière carte correcte jouée est noire, jouer une carte de valeur supérieure ou égale. Si la dernière carte jouée est rouge, jouer une carte de valeur inférieure ou égale ; les valeurs attachées au valet, à la dame, au roi et à l'as sont respectivement 11, 12, 13 et 1.
3. La carte jouée doit être, soit de la même famille (trèfle, carreau, etc.), soit de la même hauteur que la dernière carte correcte jouée.

En principe, la règle secrète ne doit dépendre que de la suite des cartes correctes jouées. Des joueurs confirmés peuvent s'autoriser des règles qui dépendent de l'ensemble des cartes disposées sur le tapis, correctes ou incorrectes, mais de telles règles sont très difficiles à deviner et ne sont pas autorisées dans la forme habituelle du jeu. La règle secrète ne peut en aucun cas dépendre d'autre chose que des cartes jouées, comme par exemple du sexe du dernier joueur, de l'heure du jour, d'une inspiration divine, etc... La règle secrète doit être écrite de façon claire et précise sur une feuille de papier que l'on met de côté pour vérification à la fin de la partie. Comme le propose M. Kruskal, le donneur peut fournir une indication avant le début du jeu. Par exemple, il peut dire : « la règle ne fait pas intervenir la couleur » ou « la règle ne dépend que des deux cartes jouées précédemment » .

### 5.1.2.2 Déroulement du jeu

Après que la règle secrète a été inscrite, le donneur bat les deux jeux de cartes et distribue 14 cartes à chaque joueur, aucune à lui-même. Il place une carte appelée « carte de départ » à l'extrême gauche du tapis de jeu.

Pour déterminer le joueur qui commence, le donneur compte dans le sens des aiguilles d'une montre à partir du joueur placé immédiatement à sa gauche, en s'excluant lui-même. Il compte jusqu'à ce qu'il ait atteint la valeur de la

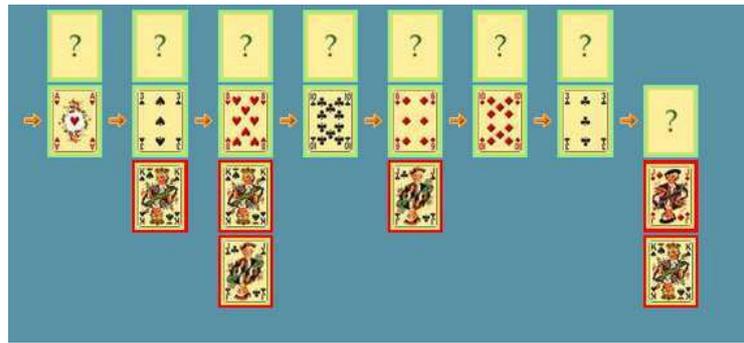


FIGURE 5.1 – Ligne Principale de Cartes Valides et les Lignes Secondaires ayant été Refusées.

carte de départ. Le joueur sur lequel tombe ce compte commence le jeu, qui se poursuivra en suivant le sens des aiguilles d'une montre.

Lorsque vient son tour de jouer, chaque joueur doit abattre de son jeu une ou plusieurs cartes, qu'il montre à tout le monde.

- si, selon la règle secrète, la carte est correcte, le donneur dit « vrai » ; la carte est alors placée à la droite de la carte de départ sur la « ligne principale » des cartes correctes qui s'allonge horizontalement de la gauche vers la droite,
- si la carte ne correspond pas à la règle, le donneur dit « faux » ; dans ce cas, la carte est placée juste au-dessous de la dernière carte jouée.

Les colonnes verticales de cartes incorrectes sont appelées « lignes secondaires » (voir Figure 5.1). Les cartes incorrectes jouées consécutivement se placent sur la même ligne secondaire de haut en bas. Au joueur qui propose une carte incorrecte, le donneur inflige une pénalité de deux cartes, qui viennent augmenter ainsi le nombre de cartes de ce joueur malheureux. Si cette possibilité de jouer plusieurs cartes en même temps semble intéressante, elle n'a pas été retenue dans notre prototype, aucune image n'est donc disponible.

Si un joueur pense avoir découvert la règle secrète, il peut jouer une « suite » de 2, 3 ou 4 cartes à la fois. Il présente les cartes de la suite légèrement décalées mais regroupées pour préserver leur ordre et les montre à tout le monde. Si toutes les cartes de la suite sont conformes à la règle, le donneur dit « vrai ». Alors, toutes les cartes sont placées côte à côte sur la ligne principale, sans se chevaucher, comme s'il s'agissait de cartes isolées correctes qui auraient été jouées séparément. Si une ou plusieurs cartes de la suite sont incorrectes, le donneur déclare fausse la suite entière, sans préciser quelles sont les cartes incorrectes. La suite est alors déposée, cartes toujours imbriquées, au-dessous de la dernière carte jouée afin qu'il reste trace de la suite jouée. Ce joueur perdant reçoit alors deux fois le nombre de cartes qu'il avait jouées dans sa

suite.

Les joueurs augmentent leur marque en se débarrassant du plus grand nombre possible de cartes et, naturellement, ils peuvent le faire d'autant mieux qu'ils ont deviné la règle secrète. Au commencement d'une partie, le choix de la carte à jouer se fait sur la base d'informations insuffisantes et nécessairement au hasard. Au fur et à mesure que la partie avance, la disposition des cartes donne de plus en plus d'informations qui vont faciliter la découverte de la règle.

Un joueur peut croire qu'il a découvert la règle secrète mais juger que son jeu ne comprend pas de cartes correctes à jouer. Il a alors la possibilité de déclarer « je passe ». Dans ce cas, il abat son jeu. Si le donneur déclare « vrai » et si le jeu de ce joueur contient moins de quatre cartes, les cartes sont remises dans le pot et la partie terminée. S'il a raison, mais a cinq cartes ou plus, alors ses cartes sont remises dans le pot et on lui distribue un nouveau jeu comportant quatre cartes de moins qu'il n'en avait précédemment. Si le joueur s'est trompé en déclarant « je passe », le donneur lui prend une de ses cartes correctes et la place sur la ligne principale. Le joueur conserve le reste de son jeu et se voit infliger une pénalité de cinq cartes supplémentaires. Un joueur pensant qu'il n'a pas de cartes correctes, mais qui n'a pas encore tout à fait deviné la règle secrète, n'a pas intérêt à passer, mais plutôt à jouer une carte au hasard.

Lorsqu'un joueur pense avoir trouvé la règle secrète, il a la possibilité de l'annoncer et d'accroître ainsi sa marque. Il le fait en se proclamant Sage. Le Sage prend immédiatement les prérogatives du donneur : il annonce « vrai » ou « faux », et distribue les cartes de pénalité aux joueurs. Il ne peut se déclarer un Sage que si et seulement si les conditions suivantes sont remplies : (1) il vient de jouer (correctement ou incorrectement) et le joueur suivant n'a pas encore joué ; (2) il n'y a encore aucun Sage existant ; (3) entre le donneur et lui, deux joueurs au moins ont encore à jouer ; (4) il n'a encore jamais été un Sage durant cette partie. Lorsqu'un joueur se déclare un Sage, il dépose une marque quelconque sur la dernière carte qu'il a jouée (un pion d'échecs, par exemple, qui servira pour le décompte des points). Le Sage conserve ses cartes, mais ne joue plus jusqu'à ce qu'il soit destitué. La partie se poursuit normalement en passant le tour du Sage.

Chaque fois qu'une carte ou une suite est jouée, le Sage proclame « vrai » ou « faux ». Le donneur entérine ou contredit l'affirmation du Sage en disant « oui », ou « non ». Si le Sage a raison, la carte, ou la suite, est placée sur le tapis, sur la ligne principale si elle est correcte, sur une ligne secondaire si elle est incorrecte et le Sage distribue toutes les cartes de pénalité requises. S'il se trompe (le donneur dit « non »), le Sage est immédiatement destitué :

il est déclaré un mauvais prophète. Le donneur enlève de la carte la marque du mauvais prophète et lui distribue cinq cartes de plus. En signe d'ignominie il lui est interdit de se déclarer Sage jusqu'à la fin de la partie. La connotation religieuse est évidente mais, comme le fait remarquer Abbott, il y a une amusante analogie avec les pratiques des milieux scientifiques : « le Sage est le scientifique qui publie », et « le mauvais prophète est le scientifique qui publie trop précipitamment ». Devenir un Sage ou démasquer un mauvais prophète est ce qui est le plus amusant dans le nouveau jeu d'Éleusis.

Après la destitution du Sage, le donneur reprend ses anciennes prérogatives. Il complète le jeu qui a confondu le Sage en plaçant la carte, ou la suite, à sa bonne place sur le tapis. Toutefois si la carte ou la suite jouée est incorrecte, il n'y a pas de cartes de pénalité. Le but de cette exemption est d'encourager les joueurs à jouer de manière inhabituelle, même délibérément de façon incorrecte, pour essayer de destituer le Sage. Dans l'optique de Karl Popper, c'est pour encourager les scientifiques à rechercher les moyens de « contredire » les théories discutables de leurs collègues. S'il y a un Sage et qu'un joueur pense ne pas avoir de carte correcte à jouer, les choses se compliquent. Au moment où un joueur déclare « je passe » il y a quatre possibilités :

1. Le Sage dit « vrai » et le donneur dit « oui ». Le Sage procède simplement comme il a été dit auparavant.
2. Le Sage dit « vrai », et le donneur dit « non ». Le Sage est immédiatement destitué. Le donneur reprend la direction du jeu comme auparavant. Le joueur qui vient de jouer n'est pas pénalisé.
3. Le Sage dit « faux », et le donneur dit « non ». En d'autres termes, le joueur a raison et le Sage se trompe. Le Sage est destitué et le donneur reprend la direction du jeu comme avant.
4. Le Sage dit « faux », et le donneur dit « oui ». Dans ce cas, le Sage doit prendre une carte correcte dans la main du joueur et la placer sur la ligne principale. S'il le fait correctement, il distribue les cinq cartes de pénalité et le jeu continue. Toutefois le Sage peut se tromper en prélevant une carte incorrecte. Si cela arrive, le Sage est destitué. La mauvaise carte est rendue au joueur et le donneur reprend la direction du jeu, sans donner de cartes de pénalité au joueur fautif.

Quand 30 cartes ont été jouées, et si aucun Sage ne s'est déclaré, les joueurs sont éliminés de la partie lorsqu'ils jouent un mauvais coup, c'est-à-dire une carte incorrecte ou une déclaration de passe erronée. On distribue au joueur éliminé les cartes de pénalité correspondant à son coup final et il quitte la partie tout en gardant sa main pour le décompte final des points. S'il y a un Sage, on diffère le processus d'élimination jusqu'à ce que 20 cartes au moins aient été jouées après celle qui porte la marque du Sage.

Une partie peut par conséquent faire alterner des phases éliminatoires et des phases normales. Par exemple, s'il y a 36 cartes sur le tapis et pas de Sage, Pierre est éliminé lorsqu'il joue incorrectement. Ensuite Paul joue correctement et se proclame Sage. Si Vincent se trompe maintenant, il n'est pas éliminé car 20 cartes n'ont pas encore été déposées après la marque du Sage, Paul.

### 5.1.2.3 Fin de la partie

Une partie peut se terminer de deux façons : (1) lorsqu'un joueur s'est débarrassé de toutes ses cartes ou (2) lorsque tous les joueurs (à l'exclusion du Sage s'il y en a un) sont éliminés. Le décompte des points au jeu d'Éleusis est le suivant :

1. On compte le nombre de cartes du joueur (Sage inclus) qui en a le plus : c'est le « maxi ». Chaque joueur (Sage inclus) soustrait du « maxi » le nombre de cartes de sa main, il obtient ainsi son décompte de points. S'il n'a plus de carte, on lui attribue un bonus de quatre points.
2. Le Sage, s'il y en a un, obtient aussi un bonus. C'est le nombre de cartes de la ligne principale au-delà de sa marque auquel s'ajoute deux fois le nombre de cartes sur les lignes secondaires après sa marque. En d'autres termes, on compte pour cette prime un point par carte correcte et deux points par carte incorrecte jouées après que le Sage s'est déclaré.
3. La marque du donneur est égale à la marque la plus élevée, avec une seule exception : s'il y a un Sage, la marque est obtenue en comptant le nombre des cartes (bonnes ou mauvaises) qui suivent la marque du Sage et en doublant ce nombre ; si le résultat est inférieur à la meilleure marque, la marque du donneur est ce plus petit nombre.

Si le jeu continue, un nouveau donneur est choisi. En principe, le jeu s'achève lorsque tous les joueurs ont été donneurs mais cela peut prendre presque une journée. Pour terminer le jeu avant que tout le monde ait été donneur, chaque joueur additionne les points qu'il a obtenus à chaque partie, plus dix points s'il n'a pas été donneur ; ce handicap compense le fait que les donneurs ont tendance à obtenir une marque plus élevée que la moyenne.

L'utilisation qui est faite des cartes dans le jeu d'*Éleusis* nous a séduit : l'aspect ludique de ces dernières ne rebute que très peu de gens, et pourtant, le fait de découvrir leur signification et leur rôle dans une règle donnée n'est pas trivial. De plus, la disposition des cartes sur le plateau constitue un excellent visuel de la temporalité et de la chronologie du jeu. Comme je le présenterai en section 5.3.2, j'ai un peu adapté cette disposition afin qu'elle représente également les bifurcations dans l'espace défini par la règle cachée. J'ai également

modifié le calcul des scores et le système de donne en incluant *Éleusis* dans un jeu plus vaste : *Nobel*.

## 5.2 Nobel

*Nobel* est un jeu inventé par David Chavalarias à la suite de son stage de DEA [Chavalarias, 1997]. Il est un protocole de jeu à  $n$ -joueurs permettant de reproduire une situation de recherche collective, et dont l'objectif scientifique est de recueillir des données sur les comportements humains dans de telles situations de recherche.

Ce jeu s'appuie sur une conception poppérienne de la recherche scientifique selon laquelle, l'activité d'une communauté de chercheurs consiste en la formulation de conjectures et la pratique de la réfutation. Le protocole, totalement paramétrable, consiste en :

- un ensemble de lois et un mode d'organisation de joueurs regroupés en équipes ( $m$ -joueurs,  $m \geq 1$ ) en compétition entre elles pour découvrir ces « lois ». Celles-ci sont des énoncés formulés dans un langage  $L$  connu des joueurs. Ceux-ci peuvent faire des expériences pour tester leur théorie courante relative à une loi, chaque expérience fournissant un résultat positif ou négatif selon que le test est conforme ou non à la loi.
- Chaque équipe peut à tout moment publier sa théorie courante pour une loi donnée, ou encore un exemple falsifiant une loi publiée. Ces deux types de publications deviennent alors connaissance commune. Celles-ci peuvent ainsi être réutilisées par chacun des joueurs dans leurs recherches.
- Un ensemble de récompenses/pénalités :  $P$  pour chaque théorie qui n'est pas falsifiée à la fin du jeu,  $-R$  pour une théorie qui a été falsifiée au cours du jeu et  $+R$  pour un exemple qui falsifie une théorie proposée. L'équipe qui a obtenu le gain maximum au bout d'un temps fixé, connu des joueurs, partage équitablement le « prix Nobel »  $N$ .

Quelques remarques a priori peuvent être faites sur la structure des gains de ce jeu :

- Maximiser les gains totaux (y compris le prix final) équivaut logiquement à maximiser les gains de la partie. Il s'ensuit que les comportements ne changent pas avec la valeur de  $N$  et dépendent seulement de  $P$  et  $R$ . Ensuite, les comportements ne sont pas modifiés si on multiplie  $P$  et  $R$  par une même valeur et ne dépendent donc que du seul rapport  $P/R$ . Il suffit donc de faire varier ce seul paramètre.
- Ce rapport  $P/R$  a une autre signification fondamentale : il régule le compromis entre explorer les théories possibles en continuant à expérimenter ou exploiter les résultats déjà obtenus pour publier. Cette question se

pose au niveau de chaque équipe qui doit ainsi arbitrer entre le nombre de ses publications et leur fiabilité.

- Différents modes d'organisation des joueurs sont possibles. Le mode le plus simple consiste à faire jouer chacun individuellement. Un second mode permet aux joueurs de former des équipes au sein desquelles ils échangent, par message, les théories courantes de chacun et les contre-exemples : l'avantage essentiel de ce mode est de permettre une observation des hypothèses de théories de chacun même lorsqu'il n'est pas suffisamment sûr pour publier. Ces deux modes sont relatifs à un petit nombre de sujets et peuvent être expérimentés en laboratoire. Un troisième mode consiste à faire jouer un grand nombre d'individus sur le web en leur laissant la possibilité de former des équipes, et d'en changer en cours de jeu. Je reviendrais sur ces perspectives au Chapitre 6.

## 5.3 Eleusis+Nobel

J'ai mis au point ce jeu en réponse au besoin, émis par les membres du CREA, de résultats empiriques sur le jeu *Nobel* pour lequel aucune implantation n'existait alors. Dans ce jeu, les joueurs doivent deviner des lois à partir d'expériences au caractère plus ou moins aléatoire. J'ai donc proposé d'intégrer le jeu *Éleusis* pour simuler les phases d'expérimentation, afin de remplacer les tests aléatoires par des expériences sur des séquences de cartes, et se rapprocher ainsi de notre problématique de découverte et de construction de connaissances. *Éleusis+Nobel* implante donc le protocole de publication défini par *Nobel* tout en intégrant les règles cachées du jeu d'*Éleusis* afin de rendre les phases d'expérimentation plus concrètes et plus ludiques. Le résultat, visible sur la figure 5.3, est détaillé et discuté dans cette section.

### 5.3.1 Les règles cachées

Comme dans *Nobel*, le langage servant à décrire les cartes est donné (langage de formulation des lois), et l'ensemble des règles cachées est partitionné en trois groupes : les règles pour lesquelles il n'existe aucune théorie publiée, les règles pour lesquelles une théorie est publiée et non falsifiée, et enfin les règles pour lesquelles la théorie publiée a été falsifiée par un contre-exemple (voir '1' sur la figure 5.3). La première modification apportée à *Éleusis* par sa fusion avec *Nobel* est donc la multiplicité des règles cachées. Les joueurs n'essaient plus de découvrir une règle cachée unique, mais plusieurs en parallèle. Chacun doit donc choisir une stratégie de publication afin d'optimiser son score en répartissant son temps de jeu entre l'exploration d'une nouvelle règle et la

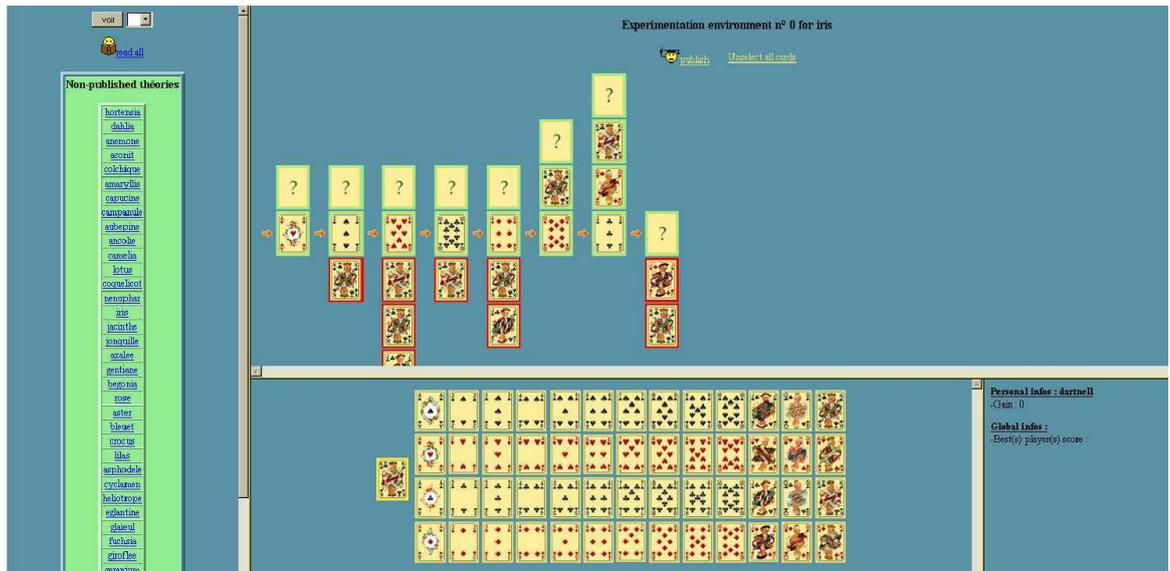


FIGURE 5.2 – Eleusis+Nobel

vérification d'une théorie publiée. Selon la terminologie introduite au Chapitre 4, les règles cachées sont des programmes logiques équivoques.

Puisque chaque joueur a désormais le loisir de naviguer entre les règles cachées, la recommandation au sujet de la difficulté des règles peut-être négligée et les règles implantées pour les expérimentations sont de tous types et vont des permissives avec peu d'exceptions jusqu'aux très contraignantes. Toutefois, ces règles sont supposées complètes, c'est à dire que quelle que soit la carte choisie en première position (et quelle que soit la règle considérée), il est possible de former au moins une séquence valide à partir de celle-ci afin de ne jamais bloquer le processus d'expérimentation qui alimente le jeu. Les théories publiées sont donc censées prédire la validité de n'importe quelle séquence de carte. Ainsi, « jouer une carte rouge après une carte noire » n'est pas une théorie complète, puisqu'elle ne permet pas de prédire la validité d'une séquence débutant par autre chose qu'une carte noire. Il est donc facile de montrer qu'elle n'est pas complète en fournissant une séquence débutant par une carte rouge. Pourtant aucune vérification automatique n'est faite lors de la publication, et je reviendrai également sur ce sujet au Chapitre 6. On distingue donc deux types de contre-exemples : les falsificateurs (une séquence formée par deux cartes noires dans l'exemple précédent), et les inattendus (une séquence débutant par une carte rouge dans l'exemple précédent).

Afin de garder une dimension humaine pour nos expérimentations, la longueur des séquences concernées par les règles cachées est fixée à 2. J'aborderai à nouveau les pré-supposés concernant la difficulté des règles cachées en lorsque

au chapitre 6.

### 5.3.2 Les espaces d'expérimentation

Chaque joueur peut sélectionner une règle cachée pour laquelle il va tester une séquence de cartes. J'ai modifié la disposition des cartes proposée par Kruskal afin de représenter visuellement le plus d'informations possibles. Il est donc désormais possible de tester une carte « à la place » d'une carte déjà jouée et appartenant à la séquence principale. Si la séquence formée avec la carte précédente sur la ligne principale est fautive, la carte est placée en dessous comme précédemment, sinon, elle est placée au dessus de la carte de la ligne principale, formant ainsi une colonne de cartes valides à cette position. Un second test est alors effectué lorsque la séquence principale continue après la position de la nouvelle carte. Si la séquence formée par la nouvelle carte et par celle qui la suit sur la séquence principale est valide, la carte est entourée en vert, sinon elle est entourée en orange. Une carte orange indique donc une bifurcation dans l'espace de recherche : deux cartes jouent le même rôle par rapport à la carte précédente, mais entraînent des comportements différents par la suite, et les séquences formées à partir de l'une ou l'autre n'auront peut-être pas la même valeur de validité.

Les résultats concernant les séquences ainsi formées sont privés, et chaque joueur a accès à ses propres espaces d'expérimentation, ainsi qu'à un jeu de carte infini (chaque carte peut être jouée autant de fois que souhaité). Ceci est la seconde modification importante introduite par la fusion d'*Éleusis* et *Nobel*. Au lieu de partager les exemples de séquences valides sur un plateau commun, les joueurs ont chacun leur propre « paillasse » qui leur permet d'explorer l'espace de leurs hypothèses comme bon leur semble. De plus, les séquences qu'ils peuvent former ne sont plus limitées par les cartes qu'ils ont en main, puisqu'ils ont accès à la totalité du jeu de carte à chaque coup joué, mais bel et bien par leurs choix exploratoires. Les joueurs peuvent changer de règle à tout moment, et retrouver leur paillasse dans l'état dans lequel ils l'avaient laissée.

### 5.3.3 Le protocole de publication

L'analogie entre le joueur qui se proclame Sage et le scientifique qui publie, soulevée en 5.1, se précise puisque le joueur qui pense avoir trouvé une règle cachée a désormais la possibilité d'éditer et de publier un ensemble de clauses logiques décrivant sa théorie (voir 6.2.3.3), les soumettant ainsi au jugement de ses pairs. Pour destituer un sage, il faut donc trouver un contre-exemple à sa théorie et le publier sous la forme d'une réfutation. Les gains  $P$  et  $R$  sont immédiats et fixés selon les règles du jeu *Nobel* ( $P = 1$ ,  $R = 2$ ). Cette

façon d'interagir pour établir une théorie et la soumettre à une validation correspond au protocole d'apprentissage interactif défini au Chapitre 4 et étend l'apprentissage à partir de requêtes.

## 5.4 Synthèse

Le jeu d'*Éleusis* a été mis au point pour simuler le phénomène de découverte d'une unique règle cachée déterminant les séquences de cartes pouvant être jouées par les participants. Le jeu *Nobel* propose un protocole de publication inspiré de la théorie des jeux et permettant d'étudier l'évolution des stratégies d'exploration et de publication dans une communauté de chercheurs. En fusionnant ces deux jeux, *Éleusis+Nobel* offre un environnement de type « jeu sérieux » pour illustrer les différents paradigmes d'apprentissage présentés au Chapitre 2, mettre en oeuvre la formalisation proposée au chapitre 4, et fournir une plate-forme robuste d'expérimentation.

Je vais maintenant présenter la modélisation et l'implantation de ce prototype au Chapitre 6 avant de présenter les résultats obtenus au Chapitre 7.



# Chapitre 6

## Implantation

*Les instruments ne sont que des théories matérialisées*

Gaston Bachelard.

LA MODÉLISATION qui a fait l'objet de ce document jusqu'à présent nécessite l'implantation d'un certain nombre de fonctionnalités. En outre, les cadres formels d'apprentissage sur lesquels je me suis appuyé pour spécifier le mode d'interaction entre les acteurs d'une résolution de problème distribuée utilisent naturellement les termes et les concepts du paradigme agent. En effet, ces paradigmes décrivent comment un agent apprenti interagit avec son environnement pour recueillir des données et tester ses hypothèses. L'extension que j'ai présentée étend cette interaction à une communauté d'agents et introduit un deuxième niveau d'interaction entre les apprentis eux-mêmes. Ce chapitre, qui constitue ma deuxième contribution, montre comment la distribution des requêtes d'un apprenti sur une communauté d'apprentis permet de rendre opérationnel les paradigmes d'apprentissage dans le cadre de la découverte scientifique.

Il convient tout d'abord de respecter ce paradigme de modélisation lors de l'implantation de la plate-forme ([Michel, 2004]). Pour des raisons de clarté et pour sa simplicité de mise en oeuvre, je me suis tourné vers le formalisme *AGR*<sup>19</sup>, spécifié par [Ferber & Gutknecht, 1998] en vue d'identifier et de manipuler les niveaux d'organisation d'un système ainsi que les comportements des entités du système au sein de ces niveaux d'organisation. J'aurais pu utiliser un autre paradigme tel que *VOYELLES* [Demazeau, 2001], qui offre une méthode de modélisation d'un système multi-agents sous la forme d'un ensemble d'Agents (A) plongé dans un environnement (E), soumis à une organisation (O), interagissant entre eux (E), et centré utilisateur (U). Cependant, cette

---

19. Agent-Groupe-Rôle

méthode, qui n'est pas si différente, n'est pas accompagnée d'un outil facilitant la mise en oeuvre et l'implantation du système comme l'est MADKIT pour *AGR*.

En plus de gérer la communication entre les agents, la plate-forme doit proposer des interfaces de saisie et de visualisation de ces communications, à savoir les publications et réfutations d'une part, les expériences et leur résultat d'autre part. J'ai dû garder à l'esprit que cette implantation devait rester la plus générique possible afin de pouvoir adapter l'outil à d'autres contextes d'utilisation (que celui des cartes), et ce à moindre coût. La modélisation agent présentée dans ce chapitre facilite cette généralité en garantissant l'autonomie des agents, et donne un cadre organisationnel à l'interaction sans créer de dépendance aux objets représentant les phénomènes étudiés par les agents (leur environnement). Le code spécifique dédié à la problématique choisie se trouve donc efficacement séparé de celui déployé pour gérer la plate-forme multi-agents et son fonctionnement.

La Section 6.1 présente une modélisation *AGR* de notre protocole, et l'implantation d'un prototype est détaillée en Section 6.2.

## 6.1 Vers une Modélisation *AGR* d'un Protocole d'Apprentissage Distribué

Le modèle formel d'interaction dont nous avons discuté au Chapitre 2 sous-entend une certaine organisation des agents, et une communication par requêtes. Selon le formalisme *AGR* les *agents* peuvent créer, joindre, ou quitter des *groupes* au sein desquels ils jouent un *rôle* donné. L'interaction se fait par envoi de *messages* prévus par un protocole d'interaction, ce qui assure l'intégrité des agents. En d'autres termes, les agents peuvent influencer le comportement des autres agents *via* l'interaction, mais ne les contrôlent pas directement et les agents restent maîtres de leurs actions. Un message peut être :

- envoyé directement à l'agent concerné si son identité est connue,
- envoyé à l'agent jouant un rôle connu,
- diffusé à l'ensemble des membres d'un groupe.

Les *alias* d'interaction que sont les noms de rôles et de groupes garantissent l'autonomie des agents et le respect du protocole d'interaction en différenciant la fonction d'un agent de son fonctionnement interne. Par exemple, si un agent *Voyageur* a besoin d'obtenir un horaire de train, de la part d'un agent *ChefDeGare*, alors quel que soit l'agent qui joue ce rôle dans le système, et quel que soit son fonctionnement ou la méthode qu'il utilise pour retrouver l'information voulue, *Voyageur* peut envoyer un message à *ChefDeGare* qui lui répondra selon le protocole établi. Imaginons maintenant qu'une défaillance

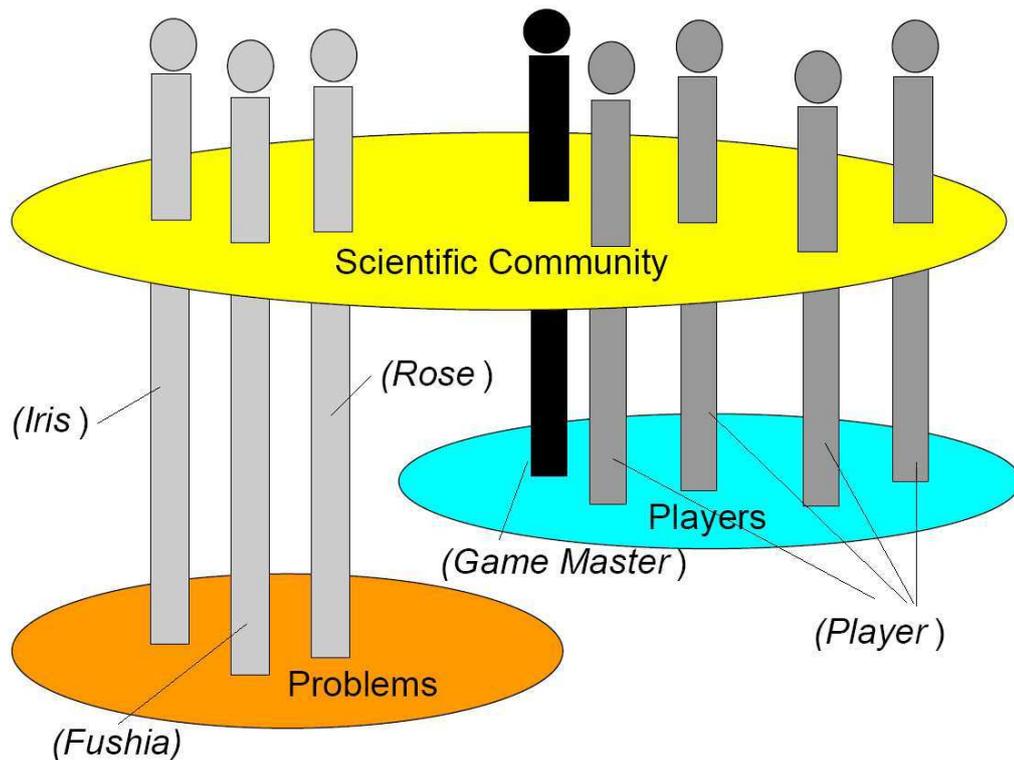


FIGURE 6.1 – Modélisation AGR d'Éleusis+Nobel

technique retarde un train *TrainX*, *ChefDeGare* peut alors diffuser un message au groupe *VoyageursTrainX* afin de prévenir tous les agents de ce groupe, quels qu'ils soient, du retard de leur train. Ce formalisme permet de modéliser facilement des systèmes ouverts et hétérogènes. Le terme « ouvert » désigne un système dont le nombre d'agents est variable, que de nouveaux agents peuvent intégrer de façon dynamique (en cours de simulation), et le terme « hétérogène » désigne des agents dont les architectures internes ne sont pas forcément identiques. En contraignant l'interaction entre agents en termes d'organisation et de rôle, le formalisme AGR permet de s'abstraire des détails d'implantation de chaque agent pour spécifier un protocole d'interaction, un vocabulaire de communication, des actes de langage [Searle, 1969, McCarthy, 1992].

La modélisation AGR d'Éleusis+Nobel présentée dans la suite de cette section est schématisée par la Figure 6.1.

### 6.1.1 Agents

Trois types d'agents (*Learner*, *Problem*, et *Game Master*) sont nécessaires :

- chaque apprenti est un agent *Learner* ;
- pour chaque problème à résoudre, un agent *Problem* est créé, doté d'un programme logique équivoque permettant de simuler les expériences des agents *Learner* ;
- un agent *Game Master* gère le jeu en créant les agents *Problem* et *Learner*, et en appliquant la fonction de gain pour mettre à jour les scores des agents *Learner* lors de leurs interactions.

### 6.1.2 Groupes

Trois groupes (*Scientific Community*, *Problems*, et *Learners*) sont également nécessaires afin de structurer la communauté efficacement :

- un groupe *Scientific Community* regroupe l'ensemble des agents et des groupes impliqués dans la résolution de problèmes,
- le groupe *Problems*, rassemble les différents problèmes étudiés par la communauté. Il est partitionné en trois sous-groupes *UnPublished*, *Published*, et *Refuted*, représentant l'état des théories pour ces problèmes. Ce groupe permet d'obtenir aisément un état de l'art et une vision générale de l'état d'avancement de la communauté ;
- le groupe *Learners* rassemble les joueurs afin d'offrir un cadre d'interaction à l'échange de conjectures. Le *Game Master* appartient également à ce groupe afin d'avoir connaissance des interactions entre les joueurs et de pouvoir attribuer les points correspondants.

### 6.1.3 Rôles

À chaque *Problem* est attribué un nom aléatoire et non représentatif du problème, par exemple un nom de fleur. Chaque *Problem* tient le rôle correspondant à son nom au sein du groupe *Problems* afin de permettre aux agents de type *Learner* de leur envoyer des messages.

### 6.1.4 Messages et Protocole d'Interaction

Des messages correspondant à l'expérimentation, la publication et la réfutation ont été définis sous la forme d'actes de langage. Les actes de langage [Searle, 1969, McCarthy, 1992] sont utilisés pour donner un statut aux informations échangées (affirmation, question, réponse, promesse) lors d'une communication [Castro *et al.* , 2002]. Cette notion a été développée en informatique dans le domaine des agents et a notamment abouti à la définition des actes de communication des langages KQML [Finin *et al.* , 1994] et ACL [FIPA, 2002]. De façon générale, ces initiatives cherchent à établir un cadre de communication

entre agents artificiels, mais des initiatives dites mixtes *i.e.*, la modélisation de l'intervention de l'utilisateur, commencent à apparaître [Castro *et al.*, 2000, Cerri & Maraschi, 2001, Maraschi *et al.*, 2001, Castro *et al.*, 2002]. Les actes de langages peuvent être définis comme un typage des informations, destiné à induire une réaction prévue de la part du destinataire. Ainsi, *Ask* et *Tell* - *Ask* induisant une réaction de type *Tell* - sont des actes de langage couramment utilisés dans le cadre d'un échange d'information, et les notions de publication et de réfutation peuvent leur être comparées.

#### 6.1.4.1 Expériences

Les messages de type *Experiment*, contenant une séquence finie représentant l'expérience de l'agent *Learner*, sont envoyés directement au *Problem* chargé de simuler les expériences pour le problème concerné (grâce au nom du rôle qu'il occupe). Le *Problem* tenant le rôle spécifié vérifie la validité de la séquence grâce au programme équivoque dont il dispose. L'émetteur reçoit alors la réponse « oui » ou « non » en fonction de la consistance de la séquence, et ce résultat est affiché dans son espace d'expérimentation comme décrit au Chapitre 5 et illustré sur la Figure 5.3. Les séquences sont construites en ajoutant une à une des cartes à la séquence existante. Les cartes formant une séquence possible sont placées à la position proposée et entourées en vert, alors que les cartes formant une séquence impossible sont placées sous la séquence principale à la position proposée et entourées en rouge. Puisque le traitement des messages *Experiment* est décidable, ces échanges sont synchronisés (l'émetteur attend la réponse du destinataire et reste bloqué dans cet état jusqu'à ce que celle-ci arrive) de telle façon que chaque *Learner* ne peut faire qu'une expérience à la fois. En termes d'actes de langage, l'acte *Experiment* induit une réponse de la part de l'agent *Problem*, tout comme *Ask* induit une réponse de type *Tell*. Cette partie du protocole est privée et assure que chaque *Learner* explorera de façon indépendante l'espace de ses hypothèses et aura un background (une expérience) différent des autres pour chaque problème.

#### 6.1.4.2 Publications et Réfutations

Après avoir évalué le risque associé à la publication d'une conjecture *i.e.*, après avoir mené *suffisamment* d'expériences pour considérer que la publication est rentable, un *Learner* peut envoyer un message de type *Publication* au groupe *Scientific Community*. Un tel message contient une description finie du programme équivoque représentant l'hypothèse courante du *Learner* (voir Section 6.2.1). Chaque *Learner* de la communauté reçoit cette requête publique et peut y apporter une réponse négative en sélectionnant un contre-exemple dans son propre espace d'expérimentation et en l'incluant dans un message de

type *Refutation*. Dans un souci de clarté, chaque réfutation référence la théorie publiée ainsi que les clauses rendues inconsistantes par le contre-exemple. Le *Problem* faisant l'objet d'une publication ou d'une réfutation réagit simplement en archivant celle-ci et en adoptant le rôle *Published* ou *Refuted* afin que l'état de l'art soit toujours visible. À cause du risque associé à la publication, l'acte *Publication* ne peut être complètement confondu avec *Ask*, puisque une réponse, toujours négative, entraînera un gain de connaissance mais également une perte de points et n'est donc pas toujours souhaitée. *Publish* entraîne donc bien un changement de comportement de la part des destinataires, mais a plutôt un sens de défi.

## 6.2 Implantation

J'ai développé *Éleusis+Nobel* comme un plugin MADKIT [Gutknecht & Ferber, 1997]. Cet outil est une plate-forme multi-agents respectant le formalisme *AGR* et permet de créer facilement en Java des systèmes multi-agents ainsi que les modules décisionnels des agents le constituant. Ainsi, la modélisation que nous avons proposée ne définit qu'un protocole de découverte, et il appartient au développeur d'implanter un raisonnement logique pour chaque *Learner*, ainsi que les méthodes d'apprentissage qu'il souhaite utiliser. Pour réaliser les expériences décrites au Chapitre 7, ce module décisionnel est lié directement à une interface web qui permet à un utilisateur de contrôler un agent de type *Learner*.

### 6.2.1 La Plate-Forme + Interfaces

Le caractère communautaire de la plate-forme justifie une architecture client-serveur afin de centraliser les objets d'interaction publics (publications et réfutations) tout en permettant l'individualisation des espaces d'expérimentation privés. La première version du jeu fût celle qui tenait le plus compte de l'activité individuelle de chaque joueur, et proposait donc un client lourd capable de s'interfacer avec des outils appliqués au domaine de recherche (outils d'analyse de séquence ADN, par exemple), offrant une représentation topologique des résultats expérimentaux, et permettant de les intégrer dans les publications. Pour des raisons d'accessibilité, et puisque l'orientation du projet s'est orientée beaucoup plus largement vers une campagne d'expérimentations sur le jeu de carte que sur l'adaptation du jeu à d'autres domaines plus sérieux, afin d'étudier les comportements des joueurs ainsi que les changements dans leur conception de la science, une version web du client a été développée en parallèle avant de devenir la version principale de la plate-forme. Toutefois, la version

actuelle du serveur, qui fonctionne avec le serveur *Apache Tomcat*, permet la communication avec les deux types de clients. Je ne présenterai que le client léger, puisque le client lourd, développé au début de cette thèse, n'est plus maintenu. Il nous a cependant permis d'identifier les composants nécessaires au bon fonctionnement de l'outil.

### 6.2.2 Le Serveur

Le serveur de jeu est une application web hébergée par un serveur Tomcat-Apache, réalisée avec des jsp (JavaServer Pages). Cette technologie, basée sur le langage Java, permet aux développeurs de générer dynamiquement du code *HTML*, *XML* ou tout autre type de page web. À chaque première utilisation d'une JSP, celle-ci est interprétée par le serveur d'application qui la convertit en *Servlet* Java et la compile. En ce sens, les JSP offrent une méthode relativement simple pour générer les servlets qui sont à la base de l'architecture client/serveur de Java. Ces servlets permettent de gérer des sessions utilisateur afin d'identifier l'auteur des requêtes *HTTP*, de conserver des informations sur celui-ci, ou de passer des arguments aux méthodes des jsp traitant les requêtes.

À l'initialisation de l'application web, la servlet principale lit un fichier de configuration et crée un booter madkit afin d'obtenir un lien sur la plate-forme multi-agents. Le fichier de configuration contient :

- une description de la fonction de gain,
- la localisation d'un répertoire destiné à stocker les publications,
- la liste des problèmes à présenter à la communauté d'apprentis. Cette liste associe le nom fictif des problèmes aux noms de classe des agents qui les implantent.

Le booter madkit est utilisé pour instancier une plate-forme MADKIT sur le serveur, créer l'agent *GameMaster*, ainsi que les agents correspondants aux problèmes spécifiés dans le fichier de configuration. Pour chaque joueur se connectant à l'application, un message de connexion est envoyé au *GameMaster* qui crée un agent *Learner*. Cet agent est associé à la session http de ce joueur de façon à lui permettre de contrôler l'agent le représentant au sein de la plate-forme multi-agents. À chaque création d'un agent, celui-ci crée ou rejoint le groupe qui lui correspond avant de demander à jouer son rôle au sein de ce groupe.

L'objet de session **Joueur** possède :

- un agent madkit,
- une table de hashage associant les problèmes étudiés à un contexte d'expérimentation,
- un score,

- une méthode *experiment*, encapsulant l’envoi d’un message de type *Experiment* via l’agent madkit,
- une méthode *publish*, encapsulant l’envoi d’un message de type *Publication* via l’agent madkit,
- une méthode *refute*, encapsulant l’envoi d’un message de type *Refutation* via l’agent madkit.

Un client web a été développé en jsp de façon à rendre l’application accessible *via* n’importe quel navigateur internet<sup>20</sup>. Toutefois, d’autres clients pourraient être développés : soit un client lourd accédant à ce même serveur, soit un client MADKIT (en cours de développement), qui offrirait une interface équivalente pour se connecter directement au plugin *Éleusis+Nobel*, sans passer par le serveur d’application qui n’est là que pour relayer les informations entre le client et MADKIT. Une autre forme d’utilisation, comme je l’ai déjà fait remarquer, serait de d’implanter un comportement pour les apprentis directement dans le module décisionnel de l’agent MADKIT pour lancer une simulation automatisée.

### 6.2.3 Le Client

Une fois connecté via un navigateur internet, le joueur se trouve face à une page HTML divisée en trois zones : un espace public présentant les différents problèmes et le statut des théories les concernant, un espace privé dans lequel il peut mener ses expériences et rédiger ses publications, et enfin une vision synthétique du jeu lui même, à titre informatif.

#### 6.2.3.1 L’Espace Public

Cette page est rafraîchie automatiquement toutes les 10 secondes (cette fréquence est paramétrable), et lorsque le joueur publie ou réfute une théorie. Elle n’utilise que les données publiques du serveur d’application et affiche les noms des agents *Problem* appartenant aux trois groupes définis (*Unpublished*, *Published*, et *Refuted*), afin de symboliser l’état de l’art de la communauté. Chaque nom de problème est affiché sous forme d’un lien *HTML* dont l’activation entraîne la sélection de la paillasse correspondante de l’utilisateur, ou sa création si le joueur n’a jamais étudié ce problème (voir Section 6.2.3.2). À côté de chaque problème figure un lien permettant de visualiser l’historique des publications pour ce problème, et un lien général en haut de page permet d’ouvrir dans une nouvelle fenêtre une page contenant l’historique des publications pour chaque problème. Comme nous le verrons en 6.2.3.3, les publications sont stockées au format *XML*. Elles peuvent ainsi être affichées en utilisant

---

20. à l’adresse [www.lirmm.fr/kayou/netoffice/eleusis/](http://www.lirmm.fr/kayou/netoffice/eleusis/)

une transformation *XSL* permettant de générer du *HTML* à partir d'un document *XML* en respectant un certain style, afin d'être présentées à l'utilisateur. Notons que cette transformation n'est utile que dans le cas où le *Learner* est contrôlé par un joueur humain, et qu'un agent logiciel peut interpréter directement le document *XML*. Un exemple de publication est donné en Section 6.2.3.3.

### 6.2.3.2 L'Espace d'Expérimentation (privé)

Cette page est elle-même divisée en deux (voir Figure 5.3). La page **Inventory**, en bas, représente l'ensemble des objets auquel le joueur a accès (dans le cas présent, des images de cartes à jouer). La page **ExperimentContext**, en haut, représente la *paillasse* associée au problème sélectionné. Cette paillasse interactive résume l'ensemble des expériences concernant le problème, comme présenté en 5.3, et permet de mener des expériences pour tester la validité d'une hypothèse (dans le cas présent, la consistance d'une séquence finie de cartes). La classe **ExperimentContext** tient donc à jour deux tableaux à deux dimensions : l'un pour les cartes valides, l'autre pour les cartes refusées. Ainsi, à chaque position dans la séquence principale est associée une colonne (contenant potentiellement une seule carte) de cartes valides, et une colonne (potentiellement vide) de cartes refusées.

De plus, puisque les expériences ne sont générées qu'à partir des cartes de la séquence principale, l'utilisateur a la possibilité de sélectionner une carte valide dans l'une des colonnes qui remplacera la carte correspondante de la séquence principale dans un nouveau contexte d'expérimentation. Chaque *paillasse* peut donc contenir plus d'un *ExperimentContext*, laissant alors le choix à l'utilisateur de sélectionner celui qu'il veut utiliser pour son expérience, d'en créer un nouveau, ou d'en supprimer un. Cette fonctionnalité a toutefois été désactivée dans les expériences relatées au Chapitre 7, de façon à ne pas trop compliquer le jeu.

L'implantation des cartes, objets d'expérimentation, respecte les interfaces **Manipulable** et **Perceptible** afin de les manipuler comme des objets génériques possédant un identifiant et un ensemble de descripteurs (les prédicats observationnels). Leur affichage sous la forme de liens *HTTP* permet l'interaction avec l'utilisateur. En cliquant sur l'une des images de l'inventaire (page du bas), l'activation du lien appelle une méthode modifiant la carte sélectionnée pour ce contexte. En cliquant sur une case vide de la paillasse, l'activation du lien appelle la méthode *experiment* (qui renvoie un booléen), en fournissant en paramètre la séquence composée des cartes précédant la position choisie et de la carte sélectionnée. En cliquant sur une des cartes de la paillasse, le joueur sélectionne l'ensemble de la séquence principale contenant cette carte (si c'est

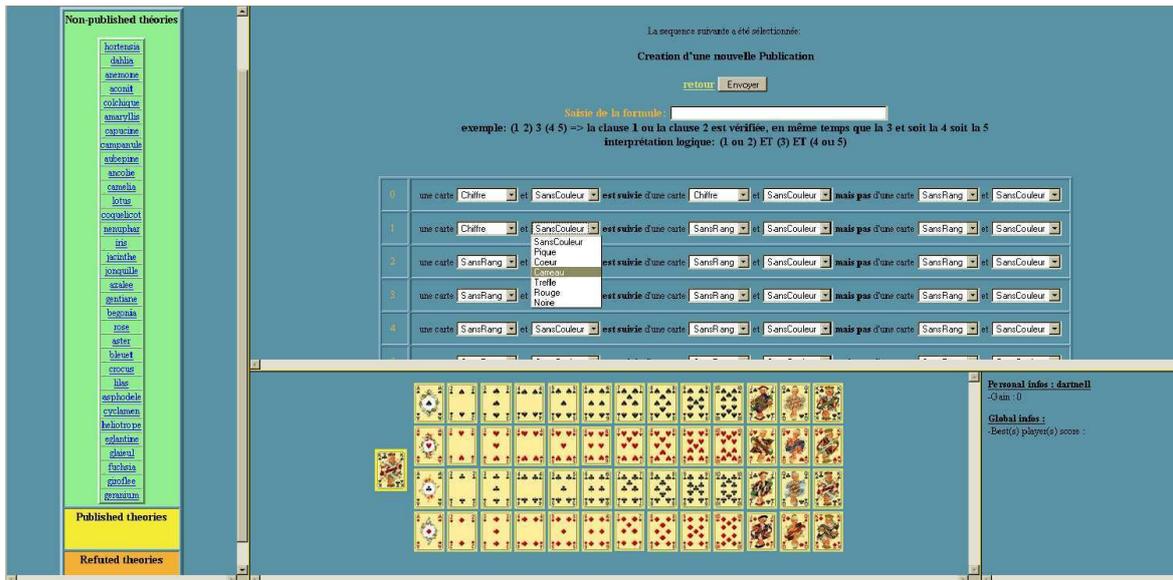


FIGURE 6.2 – Interface de Publication du Jeu *Éleusis+Nobel* .

une carte valide), soit le segment initial de la séquence principale jusqu'à la position ainsi que la carte (si c'est une carte refusée). Ce procédé permet de sélectionner les contre-exemples permettant de réfuter une théorie publiée ou de sélectionner les données expérimentales sur lesquelles le joueur fonde la théorie qu'il va publier.

### 6.2.3.3 L'Espace de Publication (privé)

Une fois qu'une séquence de carte a été sélectionnée, l'utilisateur peut accéder à l'espace de publication permettant de rédiger une nouvelle théorie (si la théorie est actuellement non publiée ou réfutée), ou une réfutation (si la théorie est actuellement publiée). Dans le cas d'une publication, l'utilisateur a accès à un ensemble de clauses numérotées et pré-formatées permettant d'exprimer l'ensemble des problèmes implantés (voir Figure 6.2. Ce formatage est le suivant :

Une carte [SelectionRang] et [SelectionCouleur] est suivie par  
 une carte [SelectionRang] et [SelectionCouleur],  
 mais pas par une carte [SelectionRang] [SelectionCouleur]

Un modèle de description exhaustif commun à l'ensemble des joueurs, contenant à la fois des prédicats observationnels et théoriques, permet de sélectionner le rang (un, deux, ..., roi, pair, impair, figure, chiffre, ou sans rang particulier) et la couleur (pique, cœur, carreau, trèfle, rouge, noir, ou sans

couleur particulière) pour chaque membre de chaque clause. Notons que l'ontologie est donc connue *a-priori* et commune à l'ensemble des joueurs, ce qui est un des biais majeurs de cette implantation. Je reviendrai sur ce sujet en Section 6.3, et me contenterai de signaler qu'étant donné que nous avons également expérimenté le jeu avec des collégiens, cette ontologie *a-priori* était nécessaire pour ne pas trop compliquer le jeu. La dernière partie de la clause a été rajoutée pour exprimer plus facilement certains des problèmes contenant des exceptions, sans avoir à énumérer tous les cas possibles.

En plus des clauses, la séquence sélectionnée est affichée pour rappel en haut de page, et une zone de texte permet de saisir librement une formule logique utilisant ces clauses, pour créer des conjonctions ou des disjonctions. Cependant, la plupart des problèmes implantés se formulent sous forme de simple conjonction, et rares sont les joueurs qui l'ont utilisé. Je reviendrai également sur ce sujet en Section 6.3.

Une fois les clauses éditées, un bouton « Send » appelle la méthode **publish** qui récupère l'ensemble des clauses éditées (en faisant abstraction des autres) et les formate pour créer un document *XML*. Ce document est diffusé au groupe *Players*, et le *GameMaster* (qui appartient à ce groupe) est en charge mettre à jour le score de l'émetteur et de transférer le message au *Problem* concerné qui archive la publication avant de changer de rôle. Le *GameMaster* peut éventuellement effectuer un contrôle des publications avant de les rendre effectives, en vérifiant par exemple que celles-ci sont complètes.

L'interface de réfutation est semblable à celle-ci et affiche la séquence sélectionnée comme contre-exemple en haut de page. La page est générée directement à partir du document *XML* représentant la publication, à l'aide d'une transformation *XSL* qui affiche un bouton de sélection devant chaque clause afin de permettre au joueur de sélectionner celles qui sont réfutées par le contre exemple. Un bouton « Envoi » appelle la méthode **refute** qui modifie le document *XML* en taguant les clauses réfutées (qui seront désormais affichées en rouge). Là encore, le message *Refute* est diffusé au groupe *Players*, et le *GameMaster* met à jour les scores avant de transférer le message au *Problem* concerné qui change alors de rôle.

Lorsqu'un agent MADKIT reçoit un message de type *Publication* ou *Refutation*, la page publique affiche une pop-up en se rafraîchissant, afin d'avertir le joueur qui peut alors changer de comportement, en fonction de sa stratégie.

Voici un exemple de publication : le joueur Dartnell publie la théorie « alternance de cartes rouges et de cartes noires », pour le problème « iris », en s'appuyant sur la séquence « Deux de coeur, Trois de piques, Roi de carreau, Dame de pique » :

<EN-Document>

```
<publi>
  <head>
    <type>Publication</type>
    <auteur>dartnell</auteur>
    <nom>iris</nom>
  <formule></formule>
</head>
<clauses>
  <clause index="0" truthtype="conjecture">
    <condition color="Rouge" rank="SansRang"></condition>
    <consequence color="Noire" rank="SansRang"></consequence>
    <exception color="SansCouleur" rank="SansRang"></exception>
  </clause>
  <clause index="1" truthtype="conjecture">
    <condition color="Noire" rank="SansRang"></condition>
    <consequence color="Rouge" rank="SansRang"></consequence>
    <exception color="SansCouleur" rank="SansRang"></exception>
  </clause>
</clauses>
<facts>
  <context>
    <carte location="img/Cartes/Deco.jpg" valid="true"></carte>
    <carte location="img/Cartes/trpi.jpg" valid="true"></carte>
    <carte location="img/Cartes/Roca.jpg" valid="true"></carte>
  </context>
  <fact> <carte location="img/Cartes/Dapi.jpg" valid="true"></carte></fact>
</facts>
<archives></archives>
</publi>
</EN-Document>
```

Lors d'une réfutation, le paramètre **truthtype** de la clause rendue inconsistante prendra la valeur « refuted », la séquence représentera le contre-exemple, et la publication actuelle sera copiée entre les balises **<archives></archives>** pour rappel.

### 6.2.4 Les Agents *Problem*

Les agents *Problem* sont des agents MADKIT adaptés à la simulation d'expériences. Une classe abstraite définit le comportement général d'un agent *Problem*, à savoir la réaction à des messages de type *Publication* ou *Refutation*,

et prévoit l'appel à une méthode **boolean validate(Sequence s)** qu'il faut redéfinir pour chaque problème, afin de vérifier la consistance de la séquence passée en paramètre avec le programme logique spécifié pour ce problème. Les programmes logiques sont implantés avec l'*API Intègre* développée par *Normind*. Cette *API* permet de gérer facilement un système de normes logiques et de vérifier la consistance d'une base de connaissance lorsque de nouvelles observations y sont ajoutées. Elle n'est donc pas exploitée à sa juste valeur et seule l'*API* « notio » permettant la gestion des graphes conceptuels est utilisée pour formuler les clauses des programmes logiques. Ce choix n'est vraiment justifié que par la collaboration qui existe entre Normind et le LIRMM. L'utilisation qui est faite d'*Intègre* peut-être améliorée, notamment lorsque l'on s'intéresse à un apprentissage portant sur des objets hautement structurés. Toujours est-il que l'architecture des agents *Problem* permet d'imaginer la création dynamique de ces agents à partir du graphe conceptuel définissant les normes, qui peut être énoncé sous la forme d'une simple chaîne de caractères. Cette perspective sera détaillée en 6.3.

### 6.2.5 Le Panneau d'Informations

Une page, affichée en bas à droite de la page principale, affiche des informations générales sur le jeu ainsi que quelques statistiques : le nombre de joueurs, le score du joueur, le score du meilleur joueur, ainsi que le nombre moyen d'expériences réalisées avant publication par le joueur d'une part, par le meilleur joueur d'autre part. Ces informations peuvent être utiles pour simuler un réseau social localisant les joueurs dans un graphe de proximité, et en ne divulguant que les informations concernant les joueurs *voisins* du joueur, afin d'observer l'effet du mimétisme dans les stratégies de découverte, comme cela est étudié par [Chavalarias *et al.* , 2005].

## 6.3 Extensions

### 6.3.1 Ontologie Libre

Comme je l'ai fait remarquer, le fait de proposer aux joueurs une ontologie suffisante pour décrire l'ensemble des problèmes à traiter est un biais considérable de notre prototype. Toutefois, ce biais est calculé et il était nécessaire afin de ne pas trop compliquer le jeu pour les expériences que nous avons mené. Ces expériences, comme nous le verrons au Chapitre 7, avaient pour but de tester le protocole sur des joueurs humains, et d'analyser les changements de conceptions qui s'opèrent chez eux durant le jeu afin de valider les fondements épistémologiques du protocole ainsi que son intérêt pédagogique. Toutefois, il

semble naturel de laisser la possibilité aux joueurs de définir eux-mêmes le modèle de description qui leur semble le plus pertinent pour chaque problème. La formalisation logique que j'ai proposée au Chapitre 4 s'avère utile dans cette perspective, puisqu'elle fait la distinction entre les prédicats observationnels, qui sont fournis à chaque *Learner* (dans ce cas, le rang et la couleur de la carte), et les prédicats théoriques qui peuvent être dérivés à partir de ceux-ci (par exemple le fait qu'une carte est rouge ou noire, paire ou impaire, que le rang correspond à un nombre premier ou non, etc.). Afin d'implanter cette fonctionnalité, il faut reconsidérer l'interface de publication ainsi que le format des documents *XML* échangés. En effet, les prédicats théoriques auraient alors un sens ambigu puisqu'ils ne seraient plus définis à-priori. Chaque publication devra donc, avant de présenter les clauses du programme logique, proposer une définition pour chaque prédicat théorique utilisé dans celles-ci, afin qu'ils puissent être interprétés de façon non-ambiguë par chaque *Learner*, exactement comme je l'ai proposé dans la formalisation.

Notons que pour les mêmes raisons de simplification, tous les problèmes proposés portent sur des séquences de deux cartes. Si cette restriction n'enlève aucun intérêt au jeu lorsqu'il est joué par des humains, on peut penser que l'espace des hypothèses ainsi défini est trop restreint pour une machine apprenante. Ce biais aussi peut être levé en supprimant la restriction apparaissant dans le pré-formatage des publications actuelles quant à la notion de successeur. Encore une fois, cette restriction n'apparaît pas dans la formalisation proposée et l'interface de publication pourrait être adaptée afin de lever ce biais. Il est à noter que de nouveaux problèmes devraient alors être implantés afin d'offrir des règles portant sur des séquences de longueur variables.

Enfin, si l'on suppose que l'interface de publication a été mise à jour pour refléter plus précisément la formalisation des programmes logiques, alors on peut imaginer que chaque joueur puisse proposer, lors de sa connexion, un ensemble de problèmes à soumettre à la communauté, ce qui éviterait d'avoir à implanter de nouveaux problèmes. Cette idée est d'ailleurs présente dans le jeu d'*Éleusis* original qui proposait de réguler la difficulté des problèmes posés par les joueurs (qui ne doivent être ni trop faciles, ni trop difficiles pour rendre le jeu intéressant), en adaptant le score d'un joueur proposant un problème en fonction de la difficulté que les autres joueurs ont eu à le résoudre. Il faudrait pour cela automatiser la création d'un programme logique à partir d'une publication.

### 6.3.2 Stratégie et Politique

Le protocole que j'ai présenté est destiné à rendre utilisable l'apprentissage par requête en distribuant celles-ci sur une communauté de joueurs. L'orienta-

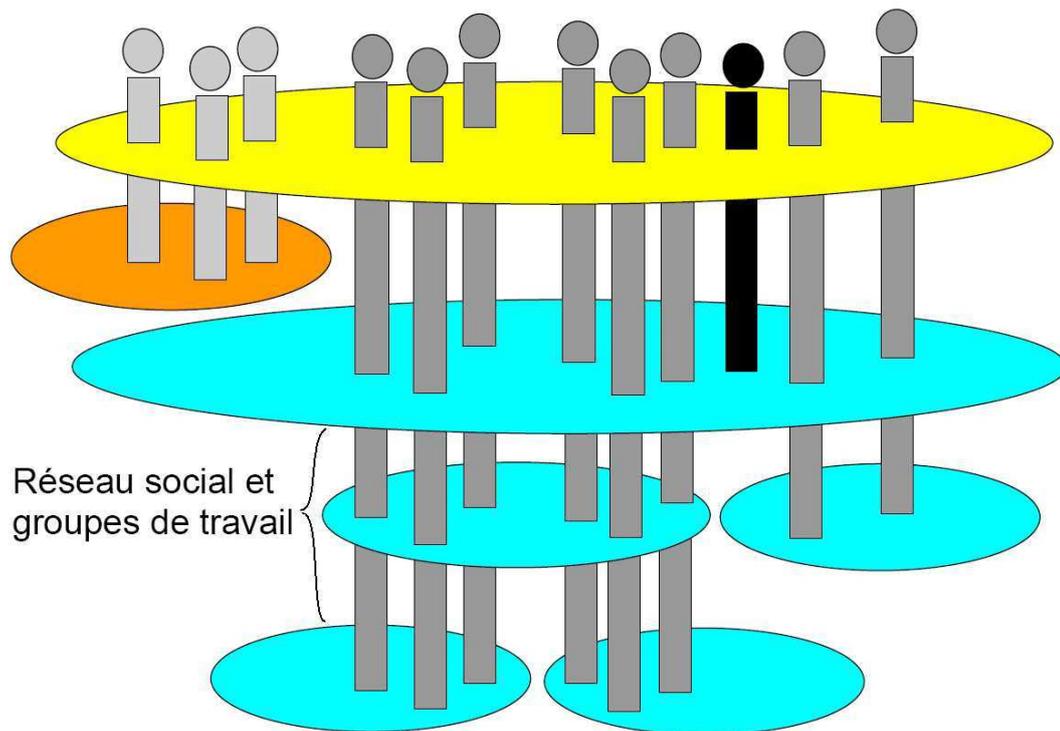


FIGURE 6.3 – Modélisation *AGR* d’une communauté d’apprentis organisée en groupes de travail

tion vers les systèmes multi-agents, même si elle semblait naturelle, n’est pas exploitée entièrement. En effet, pourquoi limiter l’échange de requêtes à une seule communauté? Une évolution possible de la plate-forme serait de permettre aux agents *Learner* de créer des groupes de travail, afin de partager résultats expérimentaux et conjectures. On peut alors imaginer une attribution de points différente pour les requêtes émises au sein du groupe de travail afin de favoriser le partage des données. On peut également imaginer que cette politique de répartition des points au sein de chaque groupe soit laissée à la discrétion du groupe lui-même, qui définirait alors ses propres règles de fonctionnement. Les agents pourraient ainsi changer de groupe si la politique de publication ne leur convient pas, et il serait intéressant d’observer les réseaux sociaux qui seraient ainsi établis. L’utilisation du formalisme *AGR* est particulièrement adapté à cette vision de la communauté scientifique. La Figure 6.3.2 montre une organisation possible des joueurs en groupes de travail, et la Figure 6.3.2 montre comment un *GameMaster* peut planifier une résolution de problèmes distribuée en créant plusieurs communautés de joueurs et en leur associant des problèmes.

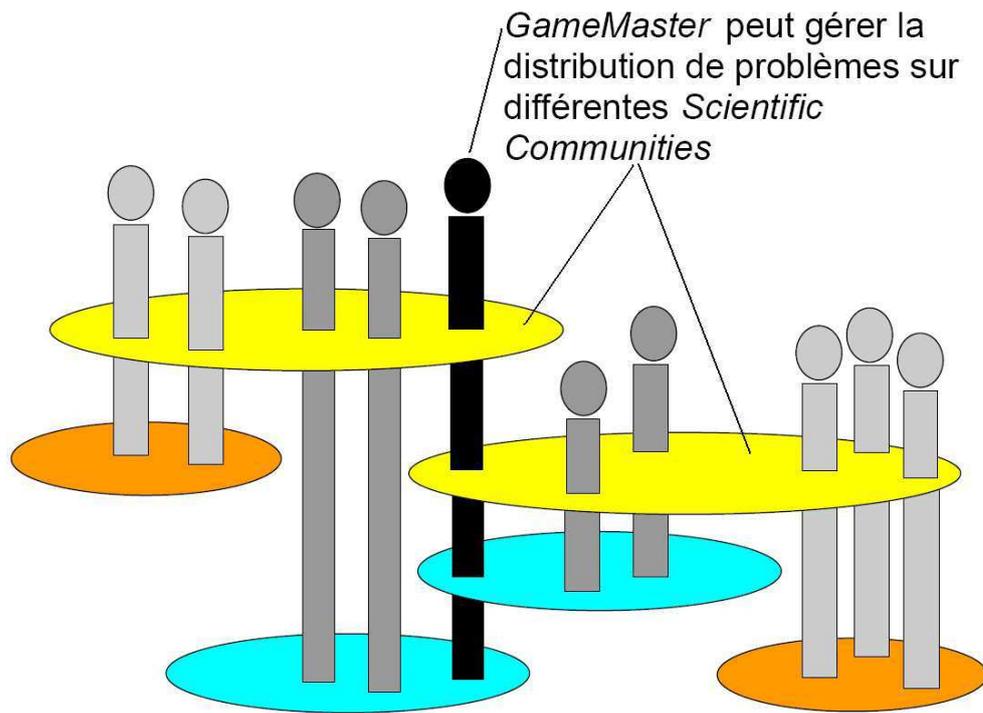


FIGURE 6.4 – Modélisation *AGR* d'une résolution de problème distribuée

### 6.3.3 Introduction des Assistants Rationnels

Maintenant que nous sommes en possession d'un outil mettant en oeuvre un protocole d'apprentissage distribué, il semble naturel d'y intégrer des agents capables d'apprendre. Comme je l'ai fait remarquer, l'implantation décrite dans ce chapitre permet aisément à un développeur d'intégrer des formes de raisonnement et des méthodes d'apprentissage aux agents *Learner*, en redéfinissant la méthode `doIt()` de ces agents, comme c'est le cas pour tous les agents *MadKit*. Il semble également naturel d'implanter des agents rationnels pour assister les joueurs, comme décrit au Chapitre 1, puisque c'est le thème de cette thèse. L'interaction entre le joueur et son assistant pourrait alors intervenir à plusieurs niveaux, et l'assistant pourrait :

- suggérer ou déconseiller des cartes à jouer (en les entourant d'une certaine couleur dans le panneau d'inventaire), et ainsi guider le joueur durant la phase d'expérimentation,
- proposer des conjectures, ou réfuter celles du joueur, en adaptant simplement le protocole à l'interaction entre le joueur et son assistant, comme cela a été proposé en 6.3.2,
- apprendre la théorie du joueur et devenir le garant de cette théorie, en répondant lui même aux messages de type *Experiment*.

Cette dernière fonctionnalité est porteuse d'un sens profond, puisqu'elle garantirait la réfutabilité des théories publiées. En effet, nos expériences ont montré une limite de ce prototype : malgré tous nos efforts pour simplifier le jeu, l'ontologie que nous avons proposée s'est avérée ambiguë et nous avons pu observer de fausses réfutations *i.e.*, des joueurs ayant mal interprétés la théorie publiée et fournissant des séquences étant effectivement des contre-exemples pour leur propres interprétations, alors qu'elles ne l'étaient pas pour les théories publiées. L'apparition de menteurs (n'étant pas toujours mal intentionnés) nous a fait prendre conscience de la nécessité de mettre en oeuvre un contrôle des réfutations. Ce problème peut être résolu en permettant aux joueurs de spécifier un programme logique permettant de tester directement la théorie publiée. Ceci peut être fait de deux façons : soit en écrivant directement le programme logique et en le « donnant » à un agent, soit en le faisant apprendre à son assistant. Dans tous les cas, le garant de la théorie pourrait alors se soustraire à l'agent *Problem* pour répondre aux messages de type *Experiment*, et incarnerait alors l'oracle décrit dans le jeu d'*Éleusis* au Chapitre 5. L'agent *Problem* pourrait alors comparer les réponses du garant aux siennes et le réfuter automatiquement lorsqu'il se trompe, ce qui garantirait la réfutabilité des théories au sens de Popper.

Le protocole d'interaction devrait être légèrement modifié en conséquence : à la réception d'un message de type *Experiment*, un *Problem* devrait transférer

ce message à l'agent garant de la théorie (dont l'adresse ou le nom serait spécifié dans la publication) afin de comparer la réponse de celui-ci avec la sienne. Si celles-ci sont identiques, alors le résultat est renvoyé normalement, sinon la séquence ayant mis en évidence la fausseté de la théorie est utilisée comme contre-exemple dans une réfutation, et le vrai résultat est renvoyé à l'auteur de l'expérimentation.

### 6.3.4 Introduction d'Heuristiques

Comme je l'ai expliqué au Chapitre 2, des travaux récents proposent une formalisation novatrice de l'apprentissage. En partant de considérations cognitives selon lesquelles un apprenti, ou tout autre agent, ne devrait pas avoir une connaissance parfaite de son propre fonctionnement, Angluin formalise les capacités d'un agent sous la forme d'un ensemble de boîtes noires qu'il doit apprendre à maîtriser pour accomplir une tâche. Notre prototype pourrait être adapté pour tenir compte de cette formalisation, en considérant que les agents de type *Learner* et *Problem* ont accès à un ensemble de programmes logiques (la boîte de fonctions) qu'ils ne connaissent pas parfaitement, et admettons que cet ensemble représente les hypothèses de l'agent, ou les mécanismes d'inférence à partir desquels il peut construire ses hypothèses. Supposons qu'il existe un tuteur ayant découvert l'indice  $i$  d'un programme logique  $g'_i$  de cet ensemble, qui étende la règle cachée  $f$ . Imaginons que la complexité associée à chaque programme de la boîte est le nombre de descripteurs de chaque objet  $x$  impliqué dans l'application de la règle cachée. Par exemple, si  $g'$  contient deux programmes  $\psi_a$  calculant la fonction « une carte rouge est suivie d'une carte noire » et  $\psi_b$  calculant la fonction « une carte carreau et paire est suivie d'un pique », leurs complexités respectives seront  $\Psi_a = 2$  et  $\Psi_b = 3$ . Notons qu'une fonction de composition peut être définie pour calculer des disjonctions de règles, et que la complexité d'un tel programme est celle du programme calculant la règle de la disjonction qui s'applique à un exemple donné. Imaginons également qu'une fonction  $b(x, s)$  décrit la différence de vocabulaire dans les modèles de description du tuteur  $T(g')$  et de l'apprenti  $L(g)$ . Par exemple,  $b(x, s) = 2s$  peut signifier que  $L(g)$  utilise deux fois plus de descripteurs que  $T(g')$ .

Une autre mesure de complexité pourrait être la taille minimale d'une séquence permettant de décider de sa validité, ou le nombre de cartes effectivement utilisées pour décider de cette validité, ou alors une association des deux, par exemple « il faut au minimum cinq cartes pour décider si la séquence est valide, et parmi ces cinq cartes, trois seulement sont utilisées pour décider ».

On peut ainsi adapter les agents de type *Problem*, ou introduire de nouveaux agents de type *Teacher* (ou encore considérer que le garant d'une théorie

peut tenir ce rôle), et modifier les messages de type *Experiment* en y introduisant une mesure de complexité à tester. Cette évolution permettrait de tenir compte du temps lié aux expériences, ce qui est totalement ignoré dans le prototype actuel.

### 6.3.5 Adaptation à d'autres Domaines

L'implantation du protocole est indépendante du domaine des cartes. Comme je l'ai précisé, les objets manipulés par les agents *Learner* lors de leurs expériences ne sont décrits qu'à l'aide de prédicats observables, au minimum leur nom. On peut donc imaginer faire abstraction des cartes, en considérant que l'espace d'expérimentation offre un aperçu intéressant des séquences d'actions pouvant être enchaînées ou non. Dans cette optique, on peut imaginer se servir de ce visuel pour des problèmes de planification, et un jeu possible serait de fixer 2 actions, par exemple, avec un certain nombre d'actions libres entre les deux, le jeu étant alors de trouver une séquence d'actions permettant d'agir dans une situation *A* pour arriver à une situation *B*, en *n* étapes. Il suffirait pour cela de modifier l'interface d'expérimentation pour en tenir compte.

Nous avons entre autre été contactés par un éditeur de logiciel de montage vidéo qui imagine utiliser *Éleusis+Nobel* comme didacticiel, afin de permettre aux utilisateur d'échanger les différentes *bonnes manières* permettant de réaliser certaines actions proposées par le logiciel.

## 6.4 Synthèse

Afin de rendre les paradigmes d'apprentissage opérationnels dans le cadre de la découverte scientifique, j'ai proposé de distribuer la résolution des requêtes d'un apprenti sur une communauté d'apprentis, dans un jeu social de publication et de réfutation de conjectures. Dans ce chapitre, j'ai proposé une modélisation *AGR* pour représenter l'organisation de cette communauté d'apprentis, et j'ai présenté un prototype respectant cette modélisation. Cette implantation est telle que les expériences peuvent être aisément adaptées à différents contextes. En effet, chaque expérience est décrite comme une séquence temporelle d'objets (ou d'évènements) identifiés comme des instances uniques par le biais des prédicats observationnels présentés au Chapitre 4. Les agents de types *Learner* et *Problem* sont ainsi libres de dériver n'importe quelle représentation pour ces objets, en fonction de leur propres langages de représentation et théories, grâce à l'utilisation de prédicats théoriques. Comme je vais le présenter au Chapitre 7, ce prototype nous a permis de valider les fondements épistémologiques de notre protocole de découverte, et nous a également

permis de constater que le jeu *Éleusis+Nobel* semble adapté pour l'enseignement de l'épistémologie, le développement de jeux sérieux, ou encore des outils de travail collaboratif. J'ai également proposé différentes améliorations de ce prototype, en décrivant pour chacune d'elles comment le modifier. Parmi ces perspectives de travail figurent la levée de certains biais (comme l'ontologie connue à-priori et la longueur fixe des séquences concernées par les problèmes) qui étaient nécessaires pour l'obtention de ces premiers résultats, ou encore l'introduction d'heuristiques dans le processus de résolution distribuée, telles que le temps lié aux expériences.

# Chapitre 7

## Expérimentations et Résultats

A FIN de fournir des éléments d'évaluation, je vais désormais décrire une série d'expériences qui me permettent d'apporter un soutien empirique aux propos théoriques que j'ai tenus jusqu'ici. Ces expériences, qui constituent ma troisième contribution, ont pu être réalisées grâce à la participation de deux membres du CREA, Sylvain Charron et David Chavalarias, ainsi qu'un membre du LIRD, Hélène Hagège.

Ces expériences nous ont aidé à répondre aux questions suivantes :

- L'interaction fondée sur l'expérimentation et l'échange de requêtes d'équivalence permet-elle effectivement un processus de découverte tel que celui qui est réalisé par une communauté scientifique ?
- Les fondements épistémologiques de cette interaction sont-ils corrects ?
- Des phénomènes émergents pourraient-ils servir à caractériser une cognition sociale ?

Nous avons donc fait jouer des collégiens, des étudiants en licence de Biologie, et enfin des amis et des membres du laboratoire à *Éleusis+Nobel*. Certains, lors de la première expérience, ont simplement joué et généré des traces d'utilisation que nous avons recueillies et analysées afin d'étudier l'impact de la distribution des requêtes sur une communauté d'apprentis (Section 7.1). D'autres ont bien voulu répondre à un questionnaire destiné à qualifier les changements éventuels de conceptions chez les joueurs (Section 7.2). Je tiens donc ici à remercier l'ensemble des participants et leurs encadrants, ainsi que les organisateurs des expériences.

### 7.1 Impact de la Résolution Distribuée des Requêtes d'Équivalence

L'objectif fixé lors de notre première expérience était de tester des hypothèses relatives à la cognition sociale et de valider la distribution des requêtes

sur une communauté d'apprentis comme moyen de rendre opérationnel un protocole d'apprentissage distribué dont les limites théoriques ont été établies (voir Chapitre 4). L'expérience qui fait l'objet de cette première section ne s'est pas déroulée dans les meilleures conditions mais a tout de même permis de recueillir des résultats encourageants. L'enjeu pour les membres du CREA qui co-organisaient cette expérience était de proposer des pistes pour la conception de méthodes d'exploitation des données recueillies, orientées vers l'analyse de phénomènes sociaux tels que le mimétisme.

### 7.1.1 Conditions expérimentales

Trois tests ont été réalisés, dans des conditions comparables : 10 joueurs, âgés de 22 à 34 ans, presque tous étudiants. Dans le premier test, les joueurs peuvent mener des expérimentations, mais sont isolés de façon à ce que leurs publications ne puissent être réfutées que par eux-mêmes. Dans les deux autres tests, 10 joueurs utilisent la plate-forme complète selon le protocole proposé (voir Chapitre 6), et les publications sont diffusées à la communauté. L'*univers* étudié par les apprentis est composé de 34 problèmes de difficultés variables (des problèmes permissifs avec peu d'exceptions, des problèmes contraignants n'offrant que peu de libertés, et enfin des problèmes intermédiaires). Le nombre de problèmes a été fixé après une série de tests préalables pour une durée de jeu estimée à environ deux heures.

### 7.1.2 Problèmes rencontrés

Le premier test s'est déroulé sans incidents. Les deux suivants, toutefois, ont dû être interrompus avant la fin du jeu. Le second, effectué en juin 2004 au CREA a été interrompu au bout de 20 minutes de jeu pour cause de problèmes de réseau et d'instabilité du serveur de jeu (la plateforme, alors implantée sans l'utilisation de MADKIT, n'avait jamais été testée avec plusieurs joueurs auparavant ; cette expérience, qui a en fait servi de test de charge, a justifié l'utilisation future de MADKIT). Les défaillances du serveur ont eu pour conséquence la perte de près de la moitié des actions de jeu, et ce test n'a pas permis d'obtenir des données exploitables. Après quelques corrections (mais toujours sans l'utilisation de MADKIT), le troisième et dernier test effectué le 11 août 2004 au LIRMM s'est parfaitement déroulé pendant 40 minutes avant de devoir être interrompu pour régler les erreurs générées par un problème inattendu discuté au Chapitre 6 : les fausses réfutations. Le jeu a pu ensuite reprendre, mais plusieurs publications avaient été perdues et la ré-attribution des publications (faussement réfutées) aux joueurs a fortement perturbé la suite du jeu. De plus, le serveur montrant à nouveau des signes d'instabilité, nous avons arrêté le jeu

20 minutes après sa reprise. Pourtant, même si ce troisième test n'a pu être mené à terme, nous avons recueilli des données sur 66 minutes de jeu effectif. Ces données ont pu être exploitées par Sylvain Charron pour son rapport de stage [Charron, 2004], et je vais présenter et commenter certains des résultats auxquels il est arrivé.

### 7.1.3 Exploitation des Données

Deux types de données ont été recueillies lors de la simulation : les données relatives aux publications et les données relatives aux expériences réalisées par les joueurs (les agents de type *Problem* qui sont sollicités lors de ces deux types d'événements consignent les informations dans des fichiers de logs). Ces données étaient sélectionnées pour étudier :

- la dynamique du processus social de découverte et l'effet de la distribution des requêtes d'équivalence d'une part (Section 7.1.4),
- la dynamique individuelle et l'évolution des stratégies de publication influençant les durées d'expérimentation d'autre part (Section 7.1.5).

L'un des résultats attendus était l'influence mutuelle des comportements sociaux et des stratégies individuelles.

### 7.1.4 Dynamique Sociale

Dans cette partie de l'étude nous avons cherché à quantifier l'évolution des connaissances produites par la communauté.

#### 7.1.4.1 Traitement des données

À partir des théories dont le pré-formatage permet une formulation en langage naturel et une interprétation en logique propositionnelle (cf. Chapitre 6), un indice de performance a été défini par Sylvain Charron grâce à une représentation matricielle des problèmes proposés. Il est à noter que cette étude n'est possible que parce que nous avons posé de fortes contraintes sur les séquences possibles pour se restreindre à des notions finies (voir Chapitre 6), nous rapprochant ainsi du cadre *PAC*.

Cette représentation matricielle des transitions autorisées (Figure 7.1.4.1) permet de calculer la distance entre les théories publiées au cours du jeu et les lois auxquelles elles se rapportent, par un calcul des distances de Hamming, et le calcul d'un indice global quantifiant le niveau de connaissance de la communauté et l'évolution de la dynamique sociale de découverte.

Cet indice est un compromis entre la qualité des publications et la vitesse de convergence vers l'horizon du jeu (état d'équilibre sur le flux des publications/réfutations).

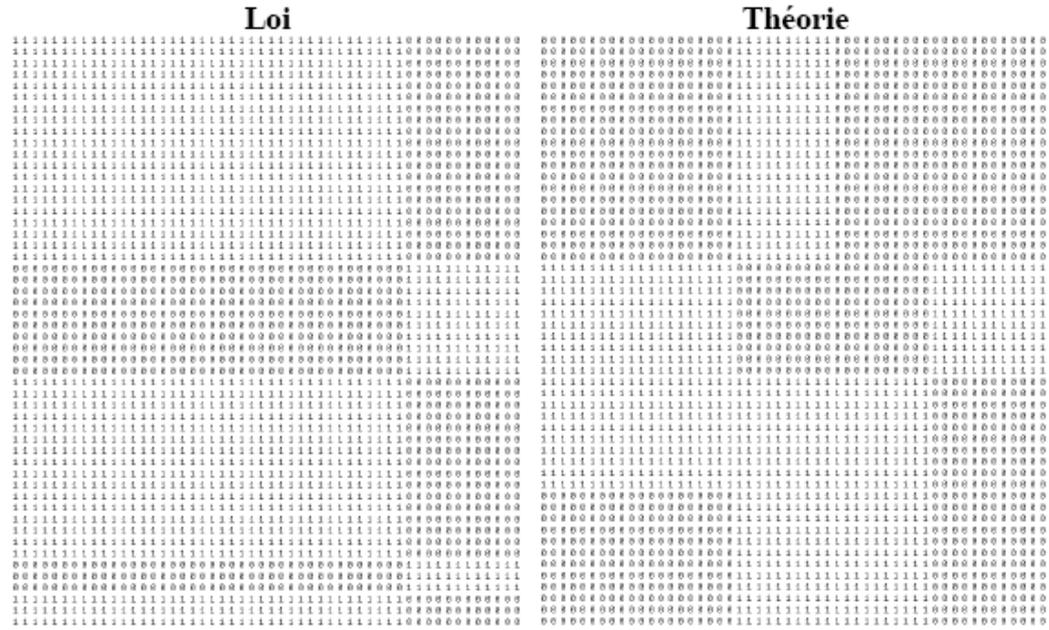


FIGURE 7.1 – Représentations Matricielles du Problème Fushia et de la Théorie Publiée par Joueur 10 (réalisation S. Charron)

Par défaut, la distance entre une théorie non publiée et un problème a été définie à  $2704=52 \times 52$ . Afin d'étudier plus précisément les différences entre les théories publiées et les lois, deux distances partielles  $d_+$  et  $d_-$  ont été calculées :

- le nombre de faux négatifs, c'est-à-dire les cas pour lesquels la théorie publiée est plus spécifique que le problème,
- le nombre de faux positifs, c'est-à-dire les cas pour lesquels la théorie est plus générale que le problème.

La distance  $d$  est donc exprimée comme un triplet ( $d = d_+ + d_-$ ).

La Figure 7.1.4.1 montre les matrices obtenues pour un problème donné, « Fushia » et une théorie publiée par un des joueurs. Le programme équivoque cible « Fushia » peut être formulé par la théorie finie suivante :

$$\begin{aligned}
 &\{\forall x, (\text{pique}(x) \rightarrow \text{figure}(s(x))), \\
 &\quad (\text{trefle}(x) \rightarrow \text{chiffre}(s(x))), \\
 &\quad (\text{coeur}(x) \rightarrow \text{chiffre}(s(x))), \\
 &\quad (\text{carreau}(x) \rightarrow \text{chiffre}(s(x)))\} \quad (7.1)
 \end{aligned}$$

La théorie publiée par le joueur dans cet exemple est visible sur la Figure 7.1.4.1.

La distance calculée entre ces deux programmes logiques est (1076,260,816).

Publication de joueur10, du groupe joueur10Group.wg pour la regle fuchsia



FIGURE 7.2 – Publication du Joueur 10 pour le problème Fushia

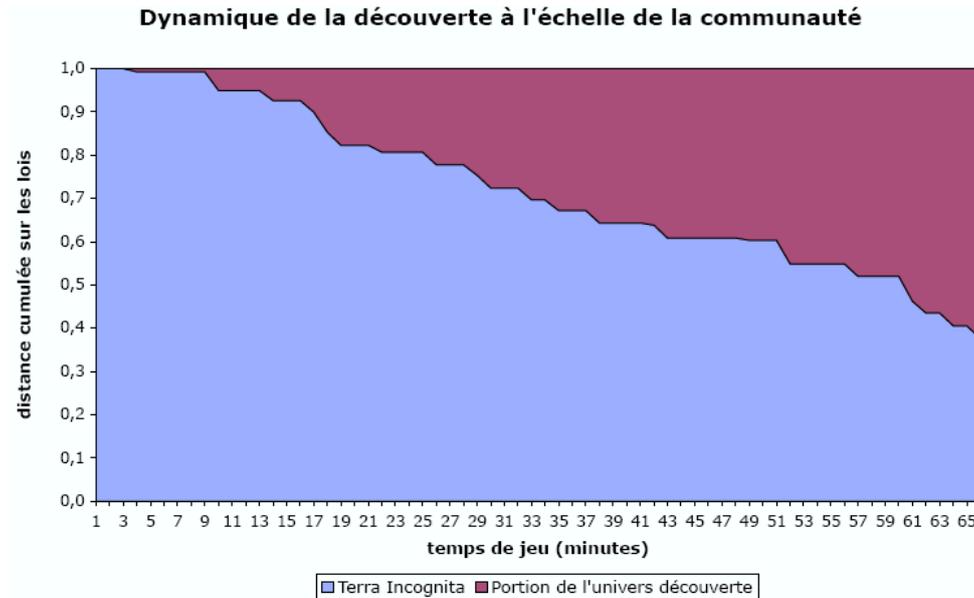


FIGURE 7.3 – Dynamique de la Découverte à l'Échelle de la Communauté (réalisation S. Charron)

Cette distance a été calculée pour toutes les publications pour tracer la distance cumulée en fonction du temps afin de caractériser le comportement de la communauté (Figure 7.1.4.1).

#### 7.1.4.2 Résultats et Cognition Sociale

Nous avons comparé les performances de joueurs isolés avec celles d'une communauté de joueurs. Dans le premier cas, un humain jouant seul met en moyenne entre 5 et 15 minutes pour publier une théorie. Celui-ci considère généralement sa théorie correcte et n'essaie pas de la réfuter. Le nombre moyen de théories publiées se situe entre 10 et 20 (les joueurs se lassaient de jouer avant d'avoir résolu l'ensemble des règles, et la séance ne durait qu'une heure

ou une heure et demie au lieu des deux heures prévues). De plus, peu des théories publiées de cette façon s'avèrent correctes (les calculs de distance n'ont malheureusement pas été faits lors du premier test). Pour contraster ces résultats, comparons-les à ceux des expériences impliquant plusieurs joueurs en interaction. Le temps moyen pour une publication est sensiblement le même, (entre 5 et 10 minutes), et nous avons observé une période d'à peu près une demi-heure durant laquelle les joueurs se sont contentés de publier des théories. Cette tendance s'est estompée avec l'épuisement des théories non publiées (qui peuvent être considérées comme des ressources pour l'activité de publication), et les joueurs ont commencé à se réfuter les uns les autres, les théories étant alors modifiées et re-publiées.

Nous avons observé (lors d'expériences ultérieures décrites en 7.2) qu'il faut entre une heure-et-demie et deux heures à une communauté de 10 à 13 joueurs humains pour arriver à un consensus représentant l'horizon du jeu. Ce chiffre est à mettre en correspondance avec les cinq heures-et-demie qu'il faudrait en théorie, sur les bases du premier test, à un humain isolé pour arriver au même résultat (en théorie car le premier test a montré qu'un joueur seul se lasse bien avant d'en arriver là). Le nombre de théories correctes parmi celles publiées par une communauté de joueurs est également supérieur.

Une des conclusions de ce rapport de stage était la conjecture suivante : l'effet des requêtes d'équivalence sur la dynamique de la découverte est d'empêcher l'apparition de paliers de connaissances (des optimums locaux) liés aux biais de la découverte individuelle ce qui, en plus d'accélérer la convergence, devrait avoir un effet sur la distance à laquelle s'établit le consensus à la fin du jeu. L'apport de cette interaction dans le processus de découverte peut ainsi être interprété comme un moyen de contourner les différents biais inhérents à l'activité de découverte individuelle :

- biais liés à l'histoire personnelle du joueur dans le jeu. Si on interprète le jeu comme un Q-learning<sup>21</sup>, on ne peut pas assurer que le résultat de l'apprentissage, temps d'arrêt et heuristiques de découverte, soient optimaux.
- biais liés à la séquence des cartes posées, les joueurs peuvent explorer différemment l'espace des solutions en fonction des catégorisations faites sur les premières cartes. Nous pouvons alors voir le rôle d'une réfutation comme un coup de projecteur sur une partie de cet espace délaissée par le parcours d'un joueur.

---

21. Le Q-learning est une technique qui se place dans le cadre de l'apprentissage par renforcement et qui produit une matrice  $Q$  dans laquelle chaque élément  $Q(s,a)$  mesure l'intérêt d'effectuer l'action  $a$  lorsque l'on se trouve dans l'état  $s$ . Par ailleurs, des résultats théoriques garantissent, dans des cas précis, la convergence de l'algorithme vers des valeurs optimales de  $Q$ .

- biais cognitifs tels que la difficulté qu'a un joueur à quitter une loi difficile pour laquelle il s'est investi, ou encore la sur-spécification consistant à décrire de façon unique la séquence de carte observée, comme dans l'identification à la limite.
- biais psychologiques comme la difficulté de se remettre en cause tout seul.

La constitution d'une communauté, à laquelle il est possible de faire appel dans les situations où l'individu est limité par ses capacités (manque de temps, incapacité à imaginer de nouvelles expériences), est donc une solution simple qui a l'avantage de se fonder sur la diversité des individus qui la composent sans leur demander plus.

### 7.1.5 Dynamique Individuelle

Un des résultats les plus importants de la simulation sur le protocole Nobel, issus de [Chavalarias, 1997], est la prédiction que le processus de découverte passe par différentes phases :

- une phase d'exploration de l'espace des solutions durant laquelle les agents publient beaucoup sans réellement s'assurer de la validité de leurs théories, et sans réellement s'intéresser à la vérification des théories publiées par leurs pairs.
- peu à peu, les temps d'expérimentation augmentent et les agents s'attachent de plus en plus à la vérification des théories publiées,
- enfin, une phase de consolidation de l'ensemble des connaissances jusqu'à l'horizon du jeu.

Bien que le jeu n'ait pas été mené à son terme, nous avons pu observer les comportements des agents et les interpréter au regard de ce résultat.

À tout moment, un *Learner* peut mener une expérience, changer de problème, ou encore publier ou réfuter une théorie. Tous ces événements ont été archivés dans des fichiers de logs pour analyser les comportements individuels et tâcher de les mettre en relation avec les événements publics.

La première constatation qui doit être faite est la diversité des comportements observés chez les joueurs. La typologie ci-dessous et le graphe représenté sur la Figure 7.1.5 présentent les principaux comportements que nous avons isolés grâce à ces événements. Cette typologie est assez cohérente avec un schéma de temps d'arrêts relatifs à la publication et à la réfutation conformément aux prévisions de [Chavalarias, 1997]. Les comportements des joueurs s'expliquent effectivement en termes de compromis exploitation/exploration marqués par des temps d'arrêt (Figure 7.1.5).

Voici la typologie proposée par [Charron, 2004], renommée pour l'occasion, classée dans l'ordre décroissant des scores des joueurs les plus caractéristiques des types.

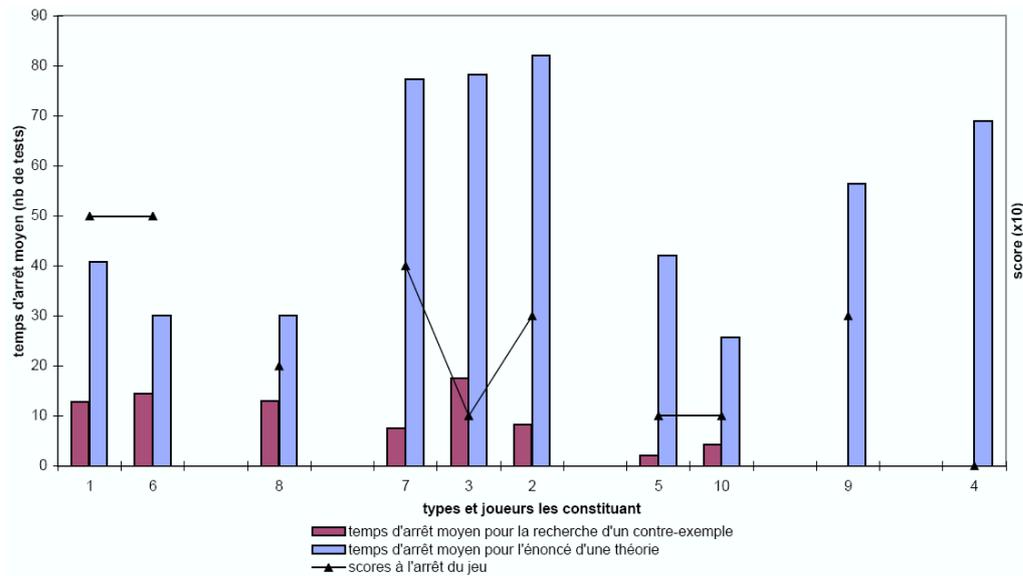


FIGURE 7.4 – Typologie des Joueurs sur 66 Minutes de Jeu (réalisation S. Charron)

**Définition 7.1 (Opportuniste :  $x=1$  et  $x=6$  sur la Figure 7.1.5)** *Le joueur teste assez systématiquement toute nouvelle publication, et cherche entre temps à publier des théories concernant les problèmes non-publiés, sans tester à fond ses théories.*

**Définition 7.2 (Prudent :  $x=7$ ,  $x=3$ ,  $x=2$ )** *Le joueur teste très longtemps les théories qu'il publie. Après publication, il effectue quelques expériences sur toutes les théories publiées pour tenter de les réfuter, puis choisit un nouveau problème à étudier.*

**Définition 7.3 (Optimiste :  $x=5$ ,  $x=10$ )** *Le joueur effectue peu d'expériences avant de publier, mais va tenter de re-publier une théorie lorsqu'il se fait réfuter. Il cherche à réfuter les théories de ses pairs principalement lorsqu'il les a déjà étudiées.*

**Définition 7.4 (Ermite :  $x=4$ )** *Le joueur teste à fond une loi sans se préoccuper des actions du reste des joueurs, puis la publie lorsqu'il est sûr de lui.*

**Définition 7.5 (Classiques :  $x=8$ ,  $x=9$ )** *typiquement des cas intermédiaires entre deux types.*

Deux interprétations à cette diversité de comportements ont été avancées :

- l'interprétation en termes de théorie des jeux pourrait être, outre le fait que les joueurs sont novices et n'ont pas encore déterminé les stratégies optimales, qu'il y ait pour ce jeu des stratégies mixtes qui se renforcent. Par exemple, les types *Prudent* et *Opportuniste* semblent suffisants pour garantir la publication de conjectures ainsi que leur réfutation à la limite.
- L'interprétation en termes de psychologie expérimentale est que les joueurs auraient des buts et des préférences diverses face au jeu et que celles-ci ne seraient pas forcément liées à l'aspect stratégique du jeu. Cette interprétation est étayée par l'enthousiasme des joueurs face au jeu, ce qui a également été observé lors de l'expérience rapportée en Section 7.2. Nous pouvons avancer l'hypothèse que pour certains joueurs, la satisfaction d'avoir trouvé une théorie pour une loi prime sur l'espérance de gain à l'horizon du jeu. Par exemple, le comportement *Optimiste* (que l'on pourrait également qualifier d' *Altruiste*) est caractéristique d'un joueur qui exploite le principe de réfutation pour tendre vers une théorie correcte, sans chercher à optimiser ses gains.

La deuxième constatation que l'on peut faire après cette expérience est que la qualité des publications s'accroît au cours du jeu. En effet, nous avons constaté que la qualité (au sens de la distance la plus faible) de la première publication concernant chaque problème augmente drastiquement au cours du jeu (la distance diminue). Au bout de 35 minutes de jeu, nous ne voyons plus de première théorie dont la distance serait supérieure à 10% de la distance maximale.

Plusieurs interprétations ont été apportées :

- Conformément au résultat énoncé par [Chavalarias, 1997], nous observons typiquement une transition dans les phases de la découverte : pendant la première demi-heure, les joueurs cherchent à publier des théories et essaient peu d'en réfuter (puisque'il n'y a pas -ou peu- de théories publiées à réfuter en début de partie). Au fur et à mesure que des théories sont publiées, elles se font réfuter et les joueurs prendraient conscience du coût de la publication d'une théorie peu étayée. Les joueurs seraient donc capables en une demi-heure, c'est-à-dire en ayant constaté l'effet de 5 réfutations sur 13 théories publiées, d'ajuster leurs temps d'arrêt.
- une autre interprétation est qu'au début du jeu, il y a peu de mise en commun d'information au niveau de la communauté. Cette interprétation diffère de la précédente sur le mode de l'apprentissage. Il ne s'agirait plus d'un simple Q-learning, mais plutôt de méta-mimétisme : en ayant connaissance des théories publiées, les joueurs pourraient apprendre sur les différents types de problèmes qui existent et inférer sur les heuristiques de découverte des autres joueurs au regard du degré de finesse

des partitions formulées par les théories. Il y aurait ainsi une augmentation progressive de la mémoire collective relative au jeu, une prise de conscience de l'existence de types de problèmes (restrictifs, permissifs, etc...), et une propagation des heuristiques de découverte.

- Une dernière interprétation qui a été proposée serait qu'au fur et à mesure que le temps passe, les lois qui restent à publier seraient les plus difficiles. Les joueurs pourraient anticiper sur ce fait et les testeraient donc plus longuement.

Cette expérience a donc confirmé de façon expérimentale que la distribution des requêtes sur une communauté d'apprentis peut rendre opérationnel un protocole d'apprentissage, et que l'utilisation de cette communauté pour confronter des jugements et partager des expériences sur un problème donné est à la fois réaliste et efficace. Notons que [Okada & Simon, 1995] a obtenu des résultats comparables (lors d'une expérience utilisant le micro-monde Simulated Molecular Genetics Laboratory [Dunbar, 1993]), en mettant en évidence le fait que des chercheurs travaillant par paires obtiennent de meilleures performances que des chercheurs isolés, notamment pour les tâches d'explication.

Nous avons également observé l'émergence de différents comportements types, tels que les *Opportunistes*, les *Prudents* ou les *Optimistes*, qui constitue une piste à explorer pour définir quelles combinaisons de comportements peut être implantée chez des *Leaner* artificiels pour assurer une dynamique de réfutation à la limite et obtenir les meilleures performances pour la communauté.

## 7.2 Impact Épistémologique et Intérêt Pédagogique

La découverte scientifique est un processus complexe possédant des dimensions psychologiques, sociales, et historiques. En ce qui concerne la dimension psychologique, la recherche fait un bond depuis qu'à la fois les produits (formation de concepts ou acquisition de connaissances) et les processus (mises au point d'expériences et évaluation de preuves) ont été intégrés au cadre de référence sous la forme de recherche duale (Dual Search, [Klahr & Dunbar, 1988]). Les simulations de découvertes se sont focalisées sur le processus scientifique en simulant des découvertes spécifiques à un domaine [Zimmerman, 2000], et peu de simulations ont pris en compte la dimension sociale qui est pourtant considérée comme centrale par les épistémologues [Kuhn, 1962]. L'expérience précédente soulève donc la question de l'utilité sociale d'*Éleusis+Nobel* : qu'est-ce que les gens qui participent à de telles simulations apprennent vraiment ? Au delà des problèmes résolus pendant la simulation, sont-ils capables de transférer ce qu'ils ont appris à leur conception de la science ; changent-ils d'épistémolo-

gie? Une présentation détaillée de l'expérience dont cette section fait l'objet a été publiée dans [Hagège et al, 2007].

### 7.2.1 Problématique

L'objectif de cette expérience est d'évaluer dans quelle mesure des simulations comme *Éleusis+Nobel* peuvent être utilisées afin d'enseigner l'épistémologie, enseignement capital pour les futurs professeurs du secondaire en science. En effet, en France, les professeurs entrent en poste très rapidement sans avoir fréquenté de laboratoire de recherche et sans avoir réellement pratiqué la science. Les récents textes officiels français contiennent à ce sujet une obligation explicite pour les enseignants de collège et lycée de connaître et d'enseigner la méthode scientifique [Bulletin-Officiel, 2007], même si l'éducation traditionnelle a toujours favorisé la transmission de contenu. Étant donné le peu de place accordée à l'épistémologie dans la formation des enseignants, ainsi que la contradiction que représente un enseignement dogmatique de l'épistémologie, nous avons envisagé l'utilisation d'*Éleusis+Nobel* comme méthode alternative pour l'enseignement de l'épistémologie. Afin d'évaluer le potentiel didactique d'*Éleusis+Nobel*, la suite de cette section relate les expérimentations réalisées sur ce jeu avec une population d'étudiants en troisième année universitaire de Biologie, souhaitant devenir professeurs de biologie dans le secondaire.

### 7.2.2 Cadre théorique de l'étude

En ce qui concerne la didactique des sciences et l'épistémologie, une vision constructiviste de l'élaboration des connaissances s'est développée ([Fourez *et al.*, 1997], [Kuhn, 1962], [Strike & Posner, 1992]) à laquelle une majorité des chercheurs de ce domaine adhèrent [Lederman *et al.*, 2002]. Pourtant d'après [Boulton-Lewis *et al.*, 2001], [Lemberger *et al.*, 1999], et [Waeytens *et al.*, 2002], cette vision ne correspondrait pas aux conceptions initiales (spontanées) des enseignants et des étudiants. Selon le point de vue constructiviste, tout savoir est lié à un sujet qui *connaît* [Fourez *et al.*, 1997], donc sa nature profonde est subjective. Les opinions, convictions, points de vue, et croyances, font partie de la science et de l'apprentissage ([Bachelard, 1938], [Kuhn, 1962]).

D'un autre côté, tout savoir est issu d'un processus de construction qui consiste en une réorganisation qualitative de la structure initiale des connaissances [Lonka *et al.*, 1996], et qui peut être assimilée à un changement de conceptions [Strike & Posner, 1992]. Ces conceptions jouent un rôle dans le raisonnement et dans l'apprentissage [Strike & Posner, 1992], mais les affects et les valeurs jouent également un rôle [Hagège, 2007]. [Hagège et al, 2007] fait

référence à « l'épistémologie personnelle » en tant que système d'attitudes, chacune correspondant à un objet épistémique (tel que l'erreur, la science,...). Ces attitudes sont formées de deux composantes, cognitive et affective (*i.e.*, conception d'un objet, et relation affective à cet objet [Hagège *et al.* , 2007a]). Elles interagissent et les normes des valeurs associées émergent de ce système d'attitudes [Hagège, 2007]. Les normes sont des règles guidant les comportements qu'un sujet adopte dans une situation particulière et les valeurs sont des principes généraux qui justifient les normes correspondantes.

La plupart des travaux sur l'enseignement et l'apprentissage de l'épistémologie concernent les conceptions (*i.e.*, ce que nous avons appelé la composante cognitive des attitudes). Schommer propose que les épistémologies personnelles peuvent varier selon un axe continu de conceptions allant d'une épistémologie naïve à une épistémologie sophistiquée ([Howard *et al.* , 2000, Schommer, 1994]). Dans le cas d'une épistémologie naïve, le savoir est simple, clair, spécifique, certain, stable, et détenu par une autorité. Mais dans le cas d'une épistémologie plus sophistiquée, le savoir devient complexe, incertain, et temporaire. Selon [Chan & Elliott, 2004], les conceptions traditionnelles et constructivistes de l'enseignement et de l'apprentissage peuvent toutes deux être opposées. Dans le premier cas, l'enseignement est considéré comme un transfert non problématique d'un savoir inchangé, d'un expert vers un novice. L'apprentissage correspond alors à l'absorption directe d'un tel savoir. Dans le second cas, l'apprentissage est considéré comme une création et une acquisition de connaissances à travers des raisonnements et des justifications. L'enseignement permet et facilite alors cet apprentissage, et ne peut pas être simplement réduit à une transmission de savoir.

Des futurs enseignants en Biologie détiennent une conception positiviste<sup>22</sup> de la science et de la connaissance [Lemberger *et al.* , 1999]. Étant donnée cette conception, les notions et les lois scientifiques existent en elles-mêmes, indépendamment de tout sujet, et sont une image exacte du monde [Fourez *et al.* , 1997]. Les sciences ne seraient pas produites pour les humains ; elles refléteraient plutôt le monde en lui-même. Cette vision peut-être qualifiée de réaliste - en opposition avec un point de vue idéaliste selon lequel le monde réel ne serait pas entièrement connaissable. La vision positiviste est cohérente avec les épistémologies *naïves* et *traditionnelles*, dans le sens où un savoir serait composé d'unités d'information progressivement ajoutées et permettraient par là même le progrès du savoir. En effet, une majorité d'enseignants du secondaire définissent l'enseignement comme un « transfert d'information maximal », et l'apprentis-

---

22. Positivismisme : « Toute conception prétendant s'en tenir aux faits du monde extérieur perceptibles par les sens et tendant à rejeter comme sans valeur toute connaissance ayant une autre origine. » (Dictionnaire de la langue philosophique)

sage comme une « absorption de toute information » ([Boulton-Lewis *et al.* , 2001, Waeytens *et al.* , 2002]).

Dans la prochaine section, nous évaluons l'impact d'*Éleusis+Nobel* sur les conceptions scientifiques, les valeurs, et dans une moindre mesure, les affects des joueurs, en utilisant une procédure standard de pré-test/post-test.

### 7.2.3 Méthodologie

Le test défini par Hélène Hagège pour cette étude est principalement composé d'une échelle de Likert et de différenciateurs sémantiques d'Osgood. Les conceptions sont définies comme allant d'une extrémité positiviste/réaliste à une extrémité idéaliste/constructiviste. Les deux pôles recouvrent grossièrement les pôles naïf/traditionnel et sophistiqué/constructiviste trouvés dans [Chan & Elliott, 2004, Schommer, 1994]. Les valeurs sont considérées comme implicites dans tous les adjectifs, mais certains, tels que « bon » ou « beau », s'y réfèrent explicitement. Enfin, les affects correspondent au domaine du plaisir et de la douleur.

#### 7.2.3.1 Procédure et Sujets de l'étude

Cette étude a été réalisée en Janvier 2007 à l'Université Montpellier II, dans le sud de la France. 43 étudiants en licence 3 de Biologie ont rempli un premier questionnaire appelé « expérience initiale ». Tous ces étudiants souhaitaient à-priori devenir enseignants ou professeurs, et suivaient donc le même cours d'épistémologie et d'éducation scientifique. Un mois et demi après, 14 d'entre eux (*PI*) ont joué à *Éleusis+Nobel* avant de remplir le second questionnaire (6 jours plus tard), alors que 14 autres (*NC*) ont servi de contrôle négatif en remplissant ce second questionnaire sans avoir participé au jeu. Le second questionnaire était constitué des questions du pré-test auxquelles ont été ajoutées de nouvelles questions (voir Annexe C). Pour chacune des deux populations *PI* et *NC*, le premier questionnaire est appelé « pré-test » tandis que le second est appelé le « post-test ». Les joueurs ont été informés que le jeu simule une situation de découverte scientifique en communauté. Durant le jeu, la population *PI* a été mélangée avec 24 autres étudiants et la population ainsi formée a été répartie en 16 équipes de 2 ou 3 joueurs. Après 2 heures de jeu, l'équipe gagnante remportait un sachet de bonbons d'1kg (le prix Nobel). Le serveur de jeu ayant alors été adapté sous forme de plug-in MADKIT , les instabilités mises en évidence par les premières expériences ne se sont pas reproduites et le jeu a pu être mené à terme. Malheureusement, une erreur dans l'implantation de la fonction de gain permettait aux joueurs de se réfuter eux-mêmes sans perdre les points associés à la réfutation. Certains joueurs (2

équipes en particulier) ont donc exploité cette faille pour gagner les points associés à la réfutation de leurs propres théories sans en perdre, ce qui provoqua une atmosphère de révolte à la fin de la simulation. Il est amusant de constater que quel que soit le système considéré, il y a toujours quelqu'un qui trouve une faille et l'exploite. Toutefois, l'expérience doit être considérée comme biaisée puisque seules deux équipes ont découvert ce *bug*.

### 7.2.3.2 Outils de mesure

Les outils de socio-psychologie classiques ont été utilisés pour cette étude. Une échelle de Likert est un groupe de propositions qui mesure la même variable psychométrique, et les sujets doivent indiquer leur degré d'accord avec chaque proposition. Pour les différenciateurs sémantiques d'Osgood relatifs à un terme, les sujets doivent choisir une position entre deux adjectifs opposés, indiquant ainsi celle qui définit le mieux le terme considéré de leur propre point de vue. Pour chacun de ces outils et pour chaque item du test, nous avons proposé cinq choix de valeurs intermédiaires possibles. Comme notre test est prospectif, et étant donnés les petits échantillons dont nous disposons, nous avons calculé des scores pour chaque outil. L'Analyse en Composante Principale (ACP) a été réalisée sur le pré-test afin de repérer les items qui semblaient mesurer la même dimension, et de les écarter le cas échéant. En nous basant sur ces résultats, nous avons regroupé les items co-variants par échelles, et vérifié à nouveau la consistance interne de ces échelles en calculant l'alpha de Cronbach. Pour chaque échelle, nous avons calculé un score, qui correspond à la valeur moyenne des réponses sur l'ensemble des items.

Le pré-test, composé de deux parties, a été initialement conçu pour évaluer les attitudes et les valeurs associées à la science et à la découverte scientifique. La première partie, une échelle de Likert à cinq points, est composée de 39 propositions. Il a pour but de différencier les positions des sujets entre deux pôles : constructivisme et idéalisme d'un côté, positivisme et réalisme de l'autre. En se basant sur l'expérience initiale, nous avons choisi 19 items qui constituent une échelle robuste (alpha de Cronbach = 0.823). L'ACP nous a permis de faire une distinction entre deux sous-échelles : étant donné la spécificité de la répartition des items observés pour chaque sous-échelle, nous avons nommé celles-ci *RTS* pour « Realisme and Truth Status » (« Statut de Vérité et Réalisme ») et *RWS* pour « Research Worker Status » (« Statut du Chercheur »), bien que chaque sous-échelle contienne des items concernant des thèmes comparables (tels que la croyance positiviste selon laquelle « le savoir s'accumule graduellement ». La seconde partie était constituée de différenciateurs sémantiques d'Osgood établis pour évaluer les conceptions, les valeurs, et dans une moindre mesure, les affects liés à cinq termes considérés comme objets épistémiques. Le test

final contenait de nouveaux différenciateurs sémantiques d'Osgood relatifs à la conception de « preuve » et de « réfutation » et deux questions ouvertes :

- « donnez trois termes que vous associez à la communication de résultats au sein d'une communauté scientifique »,
- « donnez 3 termes que vous associez à la découverte scientifique ».

Nous nous attendions à voir apparaître les termes « publication » et « réfutation », respectivement, surtout au sein de la population *PI*. Puisque ces parties ne concernent que le post-test, nous n'avons pas pu observer de changement dans les conceptions des joueurs, et les résultats ne sont donnés qu'à titre indicatif.

### 7.2.3.3 Analyse des résultats

Les données ont été collectées, saisies dans un tableur Excel©, puis analysées à l'aide du logiciel *SPSS 9.0*. Des tests non paramétriques ont été utilisés pour des comparaisons item par item (les tests de Wilcoxon sur des items appariés et des tests de Mann-Whitney sur les items non-appariés). Des tests de Student sur les items appariés ou non ont permis de comparer les scores.

## 7.2.4 Résultats et discussion

### Homogénéité des populations *PI* et *NC*

Pour vérifier que *PI* et *NC* étaient comparables, nous avons tout d'abord étudié les variables sociales (tableau 7.2.4). Les deux populations avaient la même moyenne d'âge et contenaient la même proportion d'hommes et de femmes. En ce qui concerne la catégorie socio-professionnelle des parents, nous ne pouvons rien dire puisque la plupart des sujets a répondu « autre », bien que nos échantillons ne semblent pas avoir été biaisés par des professions parentales en lien avec l'enseignement ou la recherche scientifique.

score	âge	sexe	prof. des parents
t (PI vs NC)	-1.381	0.000 <sup>a</sup>	-0.801 <sup>a</sup>
Mean (PI+NC)	20.86	1.29	5.11
ecT	1.11	0.46	1.64

TABLE 7.1 – Variables Sociales

- $t$  est obtenu après un T-test sur des items indépendants (dl = 26)
- <sup>a</sup> : Concernant ces variables ordinales, le test de Wilcoxon mène aussi à la conclusion que la population est homogène
- sexe : 1=femme, 2= homme
- prof. des parents : 1=enseignement scientifique, 2=recherche scientifique, 3=agriculture, 4=industrie, 5=santé, 6=autre
- ecT : écart type

score	RTS	RWS	V1	V2	V3	V4	A	C1	C2	C3
t (Pl vs NC)	-0.245	0.217	2.664*	0.137	0.113	1.168	0.437	-0.303	-0.625	0.077
Mean (Pl+NC)	0.18	-0.20	1.23	1.00	1.33	1.28	-1.09	0.31	-0.08	0.23
ecT	0.60	0.77	0.65	0.54	0.55	0.62	0.64	0.62	0.57	0.73

TABLE 7.2 – scores du Pré-test

– pertinence statistique :  $*p < 0.05$

Nous avons ensuite comparé les scores initiaux d'épistémologie de chaque population à travers un T-test (tableau 7.2.4). Nous pouvons noter qu'à une exception près, tous les scores peuvent être considérés comme similaires. L'exception concerne le score de la valeur esthétique *V1*, qui est significativement supérieur dans la population *PI* (voir tableau 7.2.4.1). Pourtant, nous avons pu constater que tous les sujets *PI* et *NC* ont un score positif ou nul en ce qui concerne cette variable *V1* (non montré), ce qui suggère que si la quantité de cette valeur diffère entre *PI* et *NC*, la qualité est la même : elle est positive. L'analyse item par item du test de Mann-Whitney indique que seulement 4 items sur les 106 formant l'expérience initiale sont statistiquement différents entre *PI* et *NC* (non montré). Deux de ces items font partie du score *V1*, un fait partie du score *V4*, et le dernier a été exclu de l'analyse. Nous en concluons que pour tous les scores considérés, sauf *V1*, la population *NC* constitue un contrôle négatif satisfaisant de la population *PI*.

### Épistémologie de la population initiale

Certaines moyennes ont une valeur absolue supérieure à 1, alors que d'autres sont proches de 0. Ces dernières, qui représentent des scores épistémiques à tendance marquée, concernent les valeurs positives associées à la connaissance scientifique, à la science, à l'erreur, à l'enseignement et au savoir scientifique (*V2*, *V3*, et *V4*), et les affects négatifs associés à l'erreur (*A*). Ces valeurs positives étaient attendues de la part de futurs enseignants. La dimension affective négative envers l'erreur a déjà été décrite ailleurs [Favre, 1995] et est expliquée dans une autre étude, comme étant liée à la norme socio-culturelle et institutionnelle du « donner la bonne réponse », qui est sanctionnée par l'attribution d'une bonne ou mauvaise note (et non élevée ou basse) [Hagège, 2007]. Pour les autres moyennes, qui ne reflètent pas une tendance partagée, la population est plus hétérogène. Nous soulignons donc grossièrement les tendances générales qui sont analysées pour l'expérience initiale dans son ensemble, avec plus de détails dans [Hagège, 2007]. Les moyennes légèrement négatives (*RWS* et *C2*) - indiquant des conceptions constructivistes- sont interprétées comme concernant le savoir et les connaissances : les étudiants concernés sont conscients que l'erreur a sa place en science et que les chercheurs peuvent se tromper. Mais ils pensent qu'une fois que l'erreur est détectée et supprimée, la connaissance qui reste est vraie. Ce dernier point expliquerait les moyennes légèrement positives (*RTS*, *C1*, *C3*) qui indiqueraient un point de vue réaliste et positiviste.

Score (ecT)	NC			PI		
	Moyenne au pré-test	Moyenne au post-test	t	Moyenne au pré-test	Moyenne au post-test	t
RTS	0.21 (0.68)	0.04 (0.65)	1.201	0.15 (0.55)	-0.16 (0.52)	1.967
RWS	-0.23 (0.78)	-0.56 (0.48)	1.469	-0.17 (0.79)	-0.72 (0.54)	3.016**
V1	0.93 (0.66)	0.81 (0.69)	0.563	1.52 (0.52)	1.31 (0.59)	1.188
V2	0.99 (0.54)	0.83 (0.38)	1.230	1.01 (0.56)	0.96 (0.53)	0.328
V3	1.32 (0.60)	1.25 (0.54)	0.479	1.35 (0.51)	1.29 (0.40)	0.359
V4	1.14 (0.73)	1.10 (0.64)	0.268	1.41 (0.47)	1.59 (0.45)	-1.075
A	-1.14 (0.53)	-0.79 (0.64)	-1.859	-1.04 (0.75)	-0.86 (0.77)	-1.439
C1	0.34 (0.76)	0.50 (0.50)	-0.962	0.29 (0.47)	0.09 (0.61)	1.129
C2	-0.01 (0.68)	-0.11 (0.56)	0.490	-0.15 (0.46)	-0.25 (0.33)	0.766
C3	0.22 (0.67)	0.00 (0.49)	1.223	0.25 (0.80)	-0.32 (0.56)	2.543*

TABLE 7.3 – Évaluation de l'impact d'*Éleusis+Nobel* sur les scores

- ecT : *écart type*
- *t* obtenu d'après un test de Student sur échantillons appariés comparant les moyennes des scores au pré-test et au post-test ( $dl = 13$ )
- seuil statistique de significativité : \* $p < 0.05$ , \*\* $p < 0.01$

#### 7.2.4.1 Changements dans les réponses du post-test spécifiques à la population PI

Le tableau 7.2.4.1 montre que seulement deux scores (*RWS* et *C3*) ont changé en moyenne dans le post-test de *PI*. Pour cette population, la moyenne du score *RWS* est plus négative dans le post-test que dans le pré-test, et le score de *C3* devient négatif en moyenne dans le post-test. Parmi neuf items de la sous-échelle *RWS*, six concernent le rôle de la subjectivité du chercheur en science. De plus, tous les différenciateurs sémantiques de *C3* concernent la subjectivité (ou créativité et imagination) relativement aux différents objets épistémiques. Il semble donc que les changements majeurs intervenant dans les conceptions des sujets ayant joué à *Éleusis+Nobel* concernent le rôle central de la subjectivité dans le processus de construction des connaissances.

L'analyse item par item révèle seulement quelques différences entre les réponses de *PI* au post-test et au pré-test (voir tableau 7.2.4.1). Nous pouvons noter que parmi sept items dont les réponses varient de façon significative, quatre concernent la subjectivité (*Q2*, *Q4*, *D1*, *D3*), et ce toujours dans le sens

d'une augmentation de l'intégration de la subjectivité dans leurs conceptions. Le fait que *Q2* et *Q4* entrent en jeu dans le calcul du score *RTS* renforce le résultat précédemment obtenu avec *RWS* (tableau 7.2.4.1). Un résultat intéressant est obtenu en observant l'item *Q1* ; il semblerait que le jeu ait convaincu un tiers des joueurs (non montré) qu'un chercheur isolé ne peut pas faire de science. Un autre résultat prometteur concerne *D2* : l'erreur est devenue plus « belle » pour les joueurs. C'est le seul résultat de notre étude qui concerne un changement axiologique lié à la pratique d'*Éleusis+Nobel*. Finalement, un résultat inattendu provient des réponses à *Q3*. En effet, les joueurs semblent moins croire que plusieurs interprétations sont possibles face à une observation. Peut-être ont-ils assimilé « possible », dans le sens de ce qu'un chercheur peut proposer, à « juste », dans le sens de ce qui est acceptable étant donnée une théorie. Ceci pourrait être dû au formalisme strict du jeu, dans lequel les théories sont prédéterminées et peuvent être décrites parfaitement.

Question	$Z^a$	$NC$			$Pl$		
		Moyenne au pré-test	Moyenne au post-test	$Z^b$ (pré-test vs post-test)	Moyenne au pré-test	Moyenne au post-test	$Z^b$ (pré-test vs post-test)
Q1	-0.026	0.71 (1.49)	0.29 (1.59)	-1.540	0.69 (1.70)	-0.57 (1.40)	-2.401*
Q2	-0.951	0.31 (1.38)	0.07 (1.27)	-0.666	0.79 (1.58)	-0.38 (1.04)	-2.476*
Q3	-0.171	-1.46 (0.88)	-1.21 (0.70)	-1.000	-1.57 (0.64)	-0.86 (1.17)	-2.309*
Q4	-1.278	-1.08 (1.32)	-1.14 (1.03)	-0.520	-0.29 (1.64)	-1.36 (0.84)	-2.324*
D1	-0.025	0.36 (1.08)	0.14 (0.53)	-0.918	0.38 (1.04)	-0.57 (1.09)	-2.220*
D2	-1.524	0.36 (0.84)	0.07 (0.73)	-1.265	-0.29 (1.20)	0.64 (1.01)	-2.804**
D3	-1.135	-0.21 (1.12)	0.14 (1.03)	-0.905	0.29 (0.99)	-0.50 (1.02)	-1.995*

TABLE 7.4 – Évaluation de l'impact d'*Éleusis+Nobel* sur les réponses au questionnaire

- $a$  : la variable du test de Mann-Whitney est issue de la comparaison des réponses de *Pl* et *NC* au pré-test.
- $b$  :  $Z$  est issue du test de rang de Wilcoxon sur les items appariés en comparant les réponses moyennes du pré-test et du post-test.

Voici listés ci-dessous l'ensemble des items dont les réponses au post-test varient significativement entre *PI* et *NC* (voir tableau 7.2.4.1. Un « + » indique que le fait d'adhérer totalement à cet item est compté comme +2, et un « - » indique le fait d'adhérer totalement à cet item est compté comme -2, c'est à dire que la question est inversée. Quand l'item appartient à une échelle, ceci est mentionné; *Q1*, *Q4*, *D2*, et *D3* n'ont pas été retenus dans les échelles présentées) :

- Q1 : Un chercheur isolé peut faire de la science. +
- Q2 : Les théories scientifiques sont des inventions. - (RTS)
- Q3 : Il y a toujours plusieurs interprétations possibles à un résultat d'expérience. - (RTS)
- Q4 : Les chercheurs n'utilisent pas leurs croyances pour faire de la science. +
- D1 : La connaissance scientifique : subjectif/objectif (C3)
- D2 : L'erreur : laid/beau
- D3 : Apprentissage : subjectif/objectif

#### 7.2.4.2 Changements putatifs de conception spécifiques à *PI*

Les réponses aux questions ouvertes (tableau 7.5) indiquent que nos attentes concernant l'apparition du terme « publication »- qui correspond à la nomenclature d'*Éleusis+Nobel* - chez *PI* ne sont pas satisfaites : les sujets *PI* mentionnent « article » autant que les sujets *NC*. Le terme « réfutation » n'a pas été mentionné non plus. Les deux seules nettes différences, qui n'étaient pas prédites, concernent l'utilisation des termes « discussion » et « subjectif » par les sujets *PI*. Puisque ces questions ne faisaient pas partie du pré-test, nous ne pouvons pas assurer que cette spécificité soit apparue en relation avec *Éleusis+Nobel* . Quoi qu'il en soit, ces résultats contribuent à renforcer les précédents concernant la subjectivité et le rôle de la communauté en science.

« Discussion » est le terme cité le plus grand nombre de fois. Les autres termes, qui ne sont pas cités ici, sont très disparates et semblent se regrouper dans des catégories hétérogènes.

Les différenciateurs sémantiques d'Osgood spécifiques du post-test relatifs à la *preuve* et à la *réfutation* (tableau 7.2.4.2) indiquent que la *preuve* (Pr) est significativement plus « relative », « temporelle », « statistique » et « collective » pour les sujets *PI* que pour les *NC*. À nouveau, pour les sujets *PI*, à la fois la *preuve* et la *réfutation* (RePr) sont plus « expérimentales » et « complexes ». Il est tentant de penser que ceci est lié à la pratique du jeu. L'analyse item par item (non montrée) révèle que les changements concernent l'aspect complexe de la preuve et de la réfutation, et la propriété expérimentale de la réfutation (Re).

mots	<i>NC</i>		<i>PI</i>	
	N (occurrence)	N (subjects)	N (occurrence)	N (subjects)
article	4	3	5	5
eurêka <sup>b</sup>	1	1	1	1
experimentation <sup>a,b</sup>	3	3	3	3
discussion	1	1	6	5
subjectif <sup>a</sup>	0	0	4	4

- <sup>a</sup> ou termes assimilés : « expérience », « subjectivité », ...
- <sup>b</sup> terme choisi comme contrôle négatif et n'était pas attendu

TABLE 7.5 – Nombre d'occurrences de 5 termes choisis et nombre de sujets concernés par ces occurrences dans les réponses aux 2 questions ouvertes

	RePr	Re	Pr
NC score mean (ecT)	-0.21 (0.30)	-0.11 (0.49)	0.11 (0.57)
PI score mean (ecT)	-0.91 (0.55)	-0.07 (0.87)	-0.33 (0.48)
t (NC vs PI)	-4.201***	0.134	-2.124*

TABLE 7.6 – Comparaison des scores moyens de *PI* et *NC* au différenciateurs additionnels du post-test

- *ecT* : écart type
- 't' est issu d'un T-test sur échantillons indépendants.
- pertinence statistique : \*\*\*  $p < 0.0005$ , \*  $p < 0.05$

### 7.2.5 Discussion

Étant donnés nos résultats convergents, et le changement contextuel de conceptions que nous avons observés dans la sous-population *PI*, nous avons conclu que les joueurs ont spécifiquement assimilé les rôles de la subjectivité et de la communauté dans les composantes cognitives de leurs attitudes envers la construction des objets de la connaissance. Nous avons interprété ces résultats comme un changement de leur épistémologie initiale vers une conception plus constructiviste. Étant donné que ces changements ne concernent pas uniquement des thèmes explicitement évoqués dans le jeu, mais simplement pratiqués, nous avons inféré que cette conception constructiviste a été assimilée de façon subconsciente, au sens Piagétien du terme, pendant le jeu. Nous ne pouvons pas exclure que cet effet soit apparu en synergie avec le suivi de cours traditionnels d'épistémologie. Cependant, même si tel était le cas, les changements observés sont extrêmement encourageants puisqu'ils apparaissent après seulement deux heures de jeu.

Des questions ouvertes dans le post-test concernaient le thème des sensations et d'amusement durant le jeu. Nous avons remarqué que les réponses étaient très contrastées : soit les joueurs ont beaucoup apprécié, soit ils ont été « très frustrés à cause des fausses réfutations ». Ces faits mis en évidence par les questions ouvertes ont également pu être observés durant le jeu : les joueurs se font réellement prendre au jeu. Des expérimentations antérieures réalisées sur des étudiants de 13 ou 20 ans, ainsi que l'expérience relatée en Section 7.1.4 ont conduit aux mêmes observations. À la fin du temps réglementaire, la majorité des joueurs voulaient continuer, ce qui est rarement le cas avec les cours traditionnels !

Toutes ces observations, qui n'ont pu voir le jour que grâce à la collaboration d'Hélène Hagège et à son enthousiasme, indiquent que *Éleusis+Nobel* peut constituer un outil très intéressant pour l'enseignement de l'épistémologie. Certains enseignants encadrant les expérimentations sur les plus jeunes ont également été séduits par les possibilités du jeu et proposaient déjà des adaptations à leur matière, en particulier aux mathématiques.

Toutefois, cette étude n'a pas donné de résultat convainquant en ce qui concerne la capacité des joueurs à apprendre au travers du jeu. Dans le pré-test et le post-test, nous avons également évalué leur capacité à appliquer le principe de réfutation, à manipuler des hypothèses, et à proposer des expériences pour résoudre un problème imaginatif, mais les résultats n'ont pas été concluants. Peut-être l'unique test n'était pas adapté. Cette question mérite donc de plus amples études.

## 7.3 Conclusion

J'ai relaté deux expériences menées dans des contextes différents et avec 3 ans d'écart. La première expérience a été réalisée sur une version *beta* de la plate-forme développée en partie durant mon DEA. Elle a ainsi mis en évidence une gestion multi-joueurs approximative, et la modélisation *AGR* présentée au Chapitre 6 a contribué au choix du système multi-agents MADKIT pour remédier à ce problème. La seconde expérience a permis de vérifier que notre prototype est désormais stable jusqu'à 15 joueurs, durant deux heures. Un test de charge supplémentaire est prévu pour évaluer le nombre d'utilisateurs que cette plate-forme peut désormais supporter.

Ces expériences sont représentatives du caractère pluri-disciplinaire de ce projet. J'ai développé un système rendant opérationnelle une résolution distribuée de problèmes, en me basant sur les paradigmes d'apprentissage automatique et sur des considérations logiques.

Des chercheurs en sciences cognitives membres du CREA (Sylvain Charon et David Chavalarias, qui étaient déjà présents à l'origine du projet *Éleusis+Nobel*) s'en sont servi pour conforter leurs théories concernant la cognition sociale, et j'y ai gagné une validation empirique de la pertinence de la distribution des requêtes d'équivalence sur une communauté de joueurs. Cette étude a également permis de mettre en évidence une typologie des comportements des joueurs humains dont je pourrais m'inspirer pour développer une simulation complètement automatisée afin d'observer l'influence de la répartition de ces comportements sur les performances d'une communauté d'agents artificiels :

- des agents *Optimistes* (ou *Altruistes*, ou encore *Naiifs*) qui publieraient souvent, sans se soucier des risques associés à la réfutation et sans tester à fond leurs théories,
- des agents *Opportunistes* qui se contenteraient d'attendre que d'autres publient pour tenter de les réfuter, et ne publieraient que très rarement (lorsqu'il n'y a aucune publication à réfuter),
- des agents *Prudents* qui tâcheraient de définir une durée d'expérimentation raisonnable (peut être paramétrable) avant de considérer qu'une conjecture (la leur ou celle d'un autre agent) est vraie.

Clairement, ce dernier comportement est le plus riche en terme de stratégie et de gestion du risque, mais les deux précédents assurent un flux constant de publication et de réfutation. Une interaction entre de tels agents (optimistes et opportunistes) semble suffisante pour assurer un processus de réfutation à la limite. On peut également imaginer la possibilité des rôles (au sens *AGR*) correspondant à ces comportements, afin de créer des sous-compétitions entre des agents ayant les mêmes stratégies, et observer l'effet sur les performances de la communauté.

Une chercheuse en didactique membre du LIRD (Hélène Hagège) s'en est servi pour en évaluer le potentiel didactique, et j'y ai gagné une validation empirique des fondements épistémologique du protocole d'apprentissage. Cette étude n'a pas montré que les joueurs apprenaient les problèmes auxquels ils étaient confronté (ce qui n'est pas surprenant puisque l'intérêt en est limité), mais que leurs conceptions de la science change en utilisant *Éleusis+Nobel*, après seulement deux heures de jeu. Si les utilisateurs sont capables d'acquérir une conception constructiviste de la science, alors l'implantation de notre protocole ne peut être que cohérente avec ses fondements épistémologiques. Ce résultat somme toute inattendu offre d'excellentes perspectives de travail, notamment le développement de jeux sérieux (appliqués aux mathématiques, aux sciences de la vie ou aux sciences physiques) durant lesquels les joueurs pourraient effectivement apprendre, via l'expérimentation et la publication de conjectures, les lois de la nature enseignées par la matière d'enseignement concernée.

# Chapitre 8

## Conclusion et Perspectives

*Entre ce que je pense, ce que je veux dire, ce que je crois dire, ce que je dis, ce que vous avez envie d'entendre, ce que vous croyez entendre, ce que vous entendez, ce que vous avez envie de comprendre, ce que vous comprenez, il y a dix possibilités qu'on ait des difficultés à communiquer.*

Bernard Werber.

**P**OUR conclure ce document, je vais tâcher de résumer de façon claire les différents problèmes concernés par ce travail de thèse et les solutions que j'ai proposé d'y apporter. Je vais également dégager des axes de recherche offrant selon moi des perspectives dans différents domaines.

### Conclusion

Ce travail s'inscrit dans le cadre de la découverte scientifique computationnelle, et plus précisément dans le domaine de l'assistance à la découverte par des machines capables d'apprendre. J'ai décrit ce contexte au Chapitre 1 en mettant en avant l'importance de la *dialectique* dans de tels systèmes. En m'appuyant sur un état de l'art dressé par [Langley, 2002b], j'ai mis en évidence la nécessité d'utiliser un langage logique proche de celui du chercheur pour manipuler les produits de l'apprentissage automatisé et permettre l'exploitation d'un modèle initial fourni par celui-ci. Enfin, en plus de la capacité d'apprentissage, j'ai identifié le besoin de rendre rationnel l'assistant en le dotant de capacités cognitives lui permettant de raisonner dans une situation de connaissances *insuffisantes* ou *contradictaires* qui est caractéristique du problème de découverte. Plutôt que d'apporter une solution spécifique à un problème donné,

et en insistant sur la distinction qui existe entre une méthode d'apprentissage et le paradigme dans lequel elle se place, j'ai proposé de définir un cadre général d'interaction pour la découverte scientifique computationnelle qui soit compatible à la fois avec l'apprentissage *automatique* et avec l'apprentissage *humain*.

Dans ce but, j'ai présenté au Chapitre 2 les principaux paradigmes d'apprentissage automatique en insistant sur les différences qui existent entre les formes d'interaction qu'ils définissent et la démarche scientifique. J'ai ainsi écarté l'apprentissage *PAC* [Valiant, 1984] qui pose des *restrictions arbitraires* sur les formes de réalité envisagées ainsi que sur la durée du processus de découverte, et j'ai sélectionné l'identification à la limite [Gold, 1967] et l'apprentissage par requêtes [Angluin, 1988] comme étant les mieux adaptés au contexte de découverte, en mettant en correspondance l'évolution de ces paradigmes et celle de notre conception de l'apprentissage naturel, qui tendent tous deux à attribuer à l'apprenti un *rôle actif* durant le processus d'apprentissage.

Afin de préciser ces différences, j'ai proposé au Chapitre 4 une formalisation logique du problème de découverte, en illustrant sur un jeu de cartes comment un changement de paradigme influence les règles du jeu. J'ai fait en sorte que cette formalisation reste la plus simple possible, tout en permettant de refléter les problèmes rencontrés durant un processus de découverte, telle que la mise au point d'un langage traduisant des observations, ou l'augmentation de ce langage par l'utilisation de *prédicats théoriques*, comme cela a été proposé par [Muggleton & Buntine, 1988], [Muggleton, 1994] ou encore [Banerji, 1997]. En m'inspirant de la programmation logique, j'ai insisté sur la différence que l'on peut établir entre un programme logique *univoque*, qui décrit de façon finie et unique une seule réalité infinie, et un programme logique *équivoque*, qui définit de façon finie un ensemble de réalités infinies partageant certaines caractéristiques. En m'appuyant sur ce formalisme, j'ai mis en évidence les limites des paradigmes actuels, et en particulier le fait que supposer l'existence d'un oracle capable de répondre aux requêtes de l'apprenti dans ce contexte requiert de la part de cet oracle des capacités calculatoires supérieures à celle d'une machine de Turing. J'ai alors proposé une extension *multi-apprenti* du protocole d'Angluin inspirée par l'épistémologie de Popper [Popper, 1963]. Ce modèle d'interaction fondé sur la publication et la réfutation de conjectures dans un système de *peer-reviewing* est tout à fait compatible avec la démarche scientifique et peut être appliqué à différents niveaux d'interaction : entre un chercheur et son assistant, entre une communauté de chercheur, entre une communauté d'assistants rationnels, ou encore au sein d'un groupe hétérogène. En m'inspirant de la théorie des jeux [Neumann & Morgenstern, 1944, Chavalarias, 1997], j'ai associé une fonction de gain à cette interaction afin de motiver la résolution des

---

requêtes ainsi formulées, et la *réfutation à la limite* des conjectures erronées.

Ce protocole d'apprentissage distribué, limité à l'utilisation de conjectures peut être étendu à un protocole de débat permettant de confronter des *jugements scientifiques*. La définition de ces jugements a fait l'objet du Chapitre 3 dans lequel je me suis inspiré du *carré des oppositions* et de certaines de ses extensions pour définir un *cube logique* caractérisant les oppositions qui existent entre différentes modalités. J'ai également proposé une interprétation de ces modalités qui semble naturelle et intuitive, de façon à caractériser un ensemble de jugements scientifiques suffisamment diversifié pour permettre un débat et la mise en oeuvre d'un processus dialectique. Des travaux récents [Luzeaux et al, 2008] ont montré que cet ensemble de jugements modaux, clos par négation, est en réalité une sous-structure *forte* d'une structure *faible* plus complète construite à partir des axiomes de la logique  $S_5$ , ce qui me conforte dans mon approche.

Le Chapitre 6 présente la *modélisation* et l'*implantation* d'un *prototype* mettant en oeuvre la formalisation décrite au Chapitre 4. Puisque les notions d'apprenti, d'environnement, et d'interaction sont naturellement adaptées au paradigme de modélisation *agent*, je me suis tourné vers le formalisme *AGR* [Ferber & Gutknecht, 1998] afin de décrire l'organisation de cette plate-forme d'apprentissage distribué. La définition des différents messages échangés par les agents prévoit l'utilisation des jugements définis au Chapitre 3, et l'implantation sous forme d'application web permet d'utiliser facilement l'outil pour nos expériences, puisqu'il suffit d'un navigateur internet pour y accéder. Toutefois, le serveur reste compatible avec d'autres clients éventuels respectant ce protocole, et la redéfinition d'agents MADKIT permet à un développeur de greffer simplement une méthode d'apprentissage ou une stratégie d'expérimentation dans les apprentis, afin de définir des assistants ou des agents autonomes qui pourront être utilisés dans une *simulation automatisée*.

Ce prototype nous a servi à mener une série d'expériences rapportées au Chapitre 7. Notre collaboration avec des membres du *CREA* nous a ainsi permis de valider empiriquement la *pertinence de la distribution des requêtes* sur une communauté d'apprentis, et a permis de mettre en évidence des comportements types qui pourront inspirer des simulations automatisées. Notre collaboration avec des membres du *LIRD* nous a permis de constater que *Éleusis+Nobel* représente un outil pédagogique prometteur, validant par là-même les *fondements épistémologiques* de notre protocole.

## Perspectives

Il est difficile de dire de quoi demain sera fait, mais si je suis amené à poursuivre ce travail, voici les grands axes de travail que je mettrai en place :

- Exploitation des Jugements Modaux
- Diversification des Problèmes d'Apprentissage
- Simulation Automatisée
- Application Concrète

### Exploitation des Jugements Modaux

Le protocole que j'ai défini se limite à la publication et la réfutation de conjectures. De plus j'impose que chaque conjecture (supposée décrire complètement le problème) soit réfutée avant de pouvoir en publier une nouvelle. Cette *politique* de publication correspond donc aux requêtes d'équivalence définies par [Angluin, 1988] : une conjecture est vraie ou alors elle peut-être réfutée. Afin de se rapprocher encore de l'épistémologie poppérienne, ce protocole devrait prévoir que plusieurs conjectures puissent co-exister, par exemple des conjectures décrivant chacune un aspect du problème. Cet objectif peut-être atteint en introduisant les requêtes de sous-ensemble et de sur-ensemble et en laissant la possibilité aux apprentis de publier des conjectures trop spécifiques, ou trop générales, ce qui favoriserait le partage des données expérimentales et des pistes de recherche. L'utilisation des jugements modaux définis au Chapitre 3 semble être un moyen approprié d'arriver à cet objectif : en précisant la notion de conjecture grâce à ces jugements, et en mettant en place la gestion logique des contradictions décrite par le cube logique, il me semble que l'on pourrait transformer ce protocole de publication et réfutation en cadre logique pour un débat constructif.

La fonction de gain devra elle aussi être modifier afin de permettre aux théories de gagner des points tant qu'elles ne sont pas réfutées, et de tenir compte du fait qu'une théorie trop spécifique puisse être réutilisée dans une théorie plus générale. Une théorie pourrait par exemple gagner  $f(P)$  points à chaque fois qu'un apprenti expérimente le problème concerné sans trouver de contre-exemple, et l'apprenti la réfutant gagnerait la totalité des points acquis par cette théorie, de façon à récompenser les apprentis essayant de réfuter des théories qui semblent établies. De même, une théorie qui exploiterait une théorie existante pourrait partager les points acquis avec celle-ci.

Dans cette perspective de travail il me semble également de me rapprocher de la théorie des actes de langage et de travaux tels que [Cerri, 2002] , [Castro *et al.* , 2002], ou encore [Lemoisson *et al.* , 2005], que je n'ai pas eu le temps d'explorer.

---

## Diversification des Problèmes d'Apprentissage

Comme je l'ai décrit, les paradigmes classiques d'apprentissage peuvent être reformulé dans le formalisme que j'ai proposé. Il me semble intéressant de modéliser et d'implanter ces paradigmes comme je l'ai fait au Chapitre 6 afin de pouvoir placer les apprentis dans l'un ou l'autre de ces contextes d'apprentissage. *Éleusis+Nobel* pourrait alors être exploité comme benchmark pour des méthodes particulières d'apprentissage adaptées à l'un ou l'autre de ces paradigmes. En effet, nous avons constaté que la plupart des techniques développées sont justifiées par les résultats obtenus sur des données privées, ce qui rend difficile leur comparaison objective. L'utilisation du domaine des cartes, qui bien que simple, permet de formuler une grande quantité de problèmes variés (en particulier des classes de problèmes logiques telles que les *CNF*, *DNF*, avec ontologie fixe ou libre), permettrait de comparer différentes approches sur des données publiques sans trahir le secret professionnel.

## Simulation Automatisée

Bien que le protocole présenté ait été motivé par la problématique posée par l'interaction entre un humain et un assistant logiciel, il me semble crucial de l'exploiter dans des simulations totalement automatisées pour évaluer dans quelle mesure il permet à une communauté d'agents autonomes de converger vers un consensus lors d'une résolution distribuée de problèmes. Le plug-in Warbot de MADKIT pourrait être un candidat pour ces simulations. Il met en situation deux équipes de robots artificiels formées chacune d'agents de type *Base* (immobiles, résistants, inoffensifs, et dotés d'un grand rayon de perception), de type *RocketLauncher* (mobiles mais lents, relativement résistants, offensifs, et doté d'un rayon de perception restreint), et de type *Explorer* (mobiles et rapides, peu résistants, inoffensifs, et doté d'un rayon de perception intermédiaire). L'équipe gagnant est celle qui détruit les bases adverses en premier. Ce plugin offre la possibilité à un développeur de définir un comportement pour chacun de ces types de robots, et il a un grand succès auprès des étudiants de DEA qui sont amenés à travailler dessus lors du module dédié à l'étude des systèmes multi-agents. En se plaçant dans ce contexte, on peut imaginer y appliquer notre protocole de résolution de problèmes, en laissant la possibilité aux agents de publier les problèmes qu'ils rencontrent pour demander l'aide de la communauté. Lorsqu'un agent est dans une situation problématique, il a ainsi le choix entre exploiter la solution qui fait référence dans la communauté, ou au contraire de chercher à la réfuter (si aucun danger immédiat ne le menace, par exemple).

L'exploitation de la formalisation proposée par [Angluin & Krikis, 2003] et

résumée en Section 2.4 offre des pistes de recherche intéressantes. En plus de la définition de systèmes hybrides, comme je l'ai proposé en Section 2.4.4, la perspective de laisser les agents choisir eux-mêmes d'autres agents comme des tuteurs pose le problème de l'apprentissage par imitation. L'un des problèmes posés par cette perspective de recherche est la capacité d'un agent à se *mettre à la place* d'un autre pour comparer sa façon de réagir à une situation à celle d'un autre agent. Cette capacité d'*empathie* permettrait à la fois ce choix d'un tuteur, mais également des inférences intéressantes. Par exemple, un agent pourrait, en se projetant à la place d'un autre, se *rendre compte* qu'il n'agirait lui-même ainsi que dans une situation de danger immédiat et décider de lui porter secours.

Cette application d'un cadre logique de concertation et de confrontation de jugements à une simulation automatisée pose également le problème de l'émergence d'une culture artificielle au sein d'une communauté d'agents autonomes adaptatifs. Quels seraient les concepts créés alors pour représenter leur environnement ? Ces concepts auraient-ils des ressemblances avec les concepts que pourrait utiliser un humain ? observeraient-on l'apparition de dogmes ou d'écoles de pensée chez des agents artificiels ? Cette problématique est au centre du projet « The Emergence of Artificial Culture in Robot Societies » mené par l'*EPSRC*<sup>23</sup>.

## Application Concrète

Le protocole que j'ai défini pourrait être appliqué directement à une plateforme de travail collaboratif dans des sociétés humaines, par exemple les « labos sans murs », ou une plateforme de débat public en ligne qui permettrait de résumer *visuellement* les oppositions qui existent entre les jugements modaux énoncés par les utilisateurs et les points de conflit majeurs de la communauté sur une thématique donnée, de façon à améliorer la cohérence des forums de discussion actuels et à rendre leur utilisation plus constructive, comme cela est fait sur le site *béta-politique*<sup>24</sup>, par exemple, ou dans le projet *Intermed* dont un objectif est de doter les régions d'un outil adapté au débat concernant les règles de gestion du littoral par les localités concernées.

De plus, les expériences relatées au Chapitre 7 nous incitent à appliquer ce protocole d'interaction à des outils pédagogiques. Il semble qu'il soit déjà approprié à l'enseignement de l'épistémologie, mais de nombreux enseignants encadrant les élèves qui ont participé aux expériences ont également été séduits et nous ont suggéré de l'appliquer à un domaine spécifique tel que les

---

23. Engineering and Physical Sciences Research Council. Détails du projets visibles sur <http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=EP/E062083/1>

24. [www.betapolitique.fr](http://www.betapolitique.fr)

---

mathématiques, les sciences physiques ou la biologie. Cet objectif semble réalisable même si la description des expériences proposées aux élèves soulèverait certainement de nouveaux problèmes passés inaperçus avec l'utilisation des séquences de cartes.



# Bibliographie

- [Afshar *et al.* , 2007] AFSHAR, M., DARTNELL, C., LUZEAUX, D., SALLANTIN, J., & TOGNETTI, Y. 2007. Aristotle's square revisited to frame discovery science. *Journal of computers*.
- [Afshar *et al.* , 2006] AFSHAR, MOHAMAD, LANOUE, ALIX, & SALLANTIN, JEAN. 2006. New directions : multidimensionnal optimization in drug discovery. *Comprehensive medicinal chemistry* 2, 4.
- [Angluin, 1988] ANGLUIN, DANA. 1988. Queries and concept learning. *Machine learning*, 2, 319–342.
- [Angluin & Krikis, 2003] ANGLUIN, DANA, & KRIKIS. 2003. Learning from different teachers. *Machine learning*, 51, 137–163.
- [Bachelard, 1938] BACHELARD, GASTON. 1938. *La formation de l'esprit scientifique*. Paris : Vrin.
- [Banerji, 1997] BANERJI, RANAN B. 1997. Learning theoretical terms. *Pages 93–112 of* : MUGGLETON, STEPHEN (ed), *Inductive logic programming*.
- [Beller, 2007] BELLER, SIEGHARD. 2007. Human deontic reasoning and the deontic square of opposition. *Page 3 of* : JEAN-YVES BÉZIAU, GILLMAN PAYETTE (ed), *Handbook of the first world congress on the square of oppositions*.
- [Bennett, 1988] BENNETT, C.H. 1988. Logical depth and physical complexity. *Pages 227–257 of* : HERKEN, R. (ed), *The universal turing machine : A half-century survey*. Oxford Univ. Press.
- [Bessiere *et al.* , 2007] BESSIERE, C., COLETTA, R., O'SULLIVAN, B., & PAULIN, M. 2007 (January). Query-driven Constraint Acquisition. *Pages 50–55 of* : *Proceedings of the 20th international joint conference on artificial intelligence (ijcai'07)*.
- [Beurier, 2007] BEURIER, GRÉGORY. 2007. *Codage indirect de la forme dans les systèmes multi-agents : Emergence multi-niveaux, évolution et morphogénèse*. Ph.D. thesis, Université Montpellier II, école I2S.
- [Béziau, 2002] BÉZIAU, J.-Y. 2002. S5 is a paraconsistent logic and so is first-order logic. *Logical investigations*, 9, 301–309.

- [Béziau, 2003] BÉZIAU, J.-Y. 2003. New light on the square of oppositions and its nameless corner. *Logical investigations*, **10**, 218–233.
- [Béziau, 2005] BÉZIAU, J.-Y. 2005. Paraconsistent logic from a modal viewpoint. *Journal of applied logic*, **3**, 7–14.
- [Béziau, 2007] BÉZIAU, J.-Y. 2007. Adventures in the paraconsistent jungle. *In : Handbook of paraconsistency, king's college*.
- [Blanché, 1966] BLANCHÉ, R. 1966. *Structures intellectuelles : essai sur l'organisation systématique des concepts*. Vrin, Paris.
- [Blum & Blum, 1975] BLUM, L., & BLUM. 1975. Toward a mathematical theory of inductive inference. *Inform. control*, **28 :2**, 125–155.
- [Boulton-Lewis et al. , 2001] BOULTON-LEWIS, G.M., SMITH, D.J.H., MCCRINDLE, A.R., BURNETT, P.C., & CAMPBELL, K.J. 2001. Secondary teachers' conceptions of teaching and learning. *Learning and instruction*, **11**, 35–51.
- [Brenner, 2007] BRENNER, JOSEPH. 2007. Applying the square of opposition to reality. *Page 20 of : JEAN-YVES BÉZIAU, GILLMAN PAYETTE (ed), Handbook of the first world congress on the square of oppositions*.
- [Bridewell et al. , 2006] BRIDEWELL, WILL, SANCHEZ, JAVIER NICOLAS, LANGLEY, PAT, & BILLMAN, DORRIT. 2006. An interactive environment for the modeling and discovery of scientific knowledge. *International journal of human-computer studies*.
- [Bulletin-Officiel, 2007] BULLETIN-OFFICIEL. 2007. *Les compétences professionnelles des maîtres*.
- [Castro et al. , 2000] CASTRO, E., SALLANTIN, J., & CERRI, S. 2000. Misunderstanding detection using a constraint based mediator. *In : Alcaa*.
- [Castro et al. , 2002] CASTRO, E., SALLANTIN, J., & CERRI, S. A. 2002. Crat - vers une programmation conversationnelle. *Pages 111–122 of : ET S. GOUARDERES, P. ANIORTE (ed), Cognitiques*.
- [Cavaliere & Donnarumma, 2007] CAVALIERE, FERDINANDO, & DONNARUMMA, ANTONIO. 2007. Fuzzy syllogisms, numerical square, triangle of contraries, inter-bivalence - with an historical appendix on the quantification of the predicates. *Page 24 of : JEAN-YVES BÉZIAU, GILLMAN PAYETTE (ed), Handbook of the first world congress on the square of oppositions*.
- [Cerri, 2002] CERRI, S. A. 2002. Human and artificial agent's conversations on the grid. *1st international workshop on educational models for grid based services*.
- [Cerri & Maraschi, 2001] CERRI, S. A., & MARASCHI, D. 2001. The relations between technologies for human learning and agents. *Atelier : Mé-*

- 
- thodologies et environnements pour les systèmes multi-agents - afia 2001*, 25–29.
- [Chalmers, 1993] CHALMERS, A. F. 1993. *Qu'est-ce que la science finalement?* Editora Brasiliense, Sao Paulo.
- [Chan & Elliott, 2004] CHAN, K-W, & ELLIOTT, RG. 2004. Relational analysis of personal epistemology and conceptions about teaching and learning. *Teaching and teacher education*, 817–31.
- [Charron, 2004] CHARRON, SYLVAIN. 2004. *Un protocole pour l'étude d'un processus social de découverte : le jeu nobel-eleusis*. Rapport de DEA, CREA - CNRS/Ecole Polytechnique.
- [Chaumon *et al.* , 2003] CHAUMON, LOIC, MAZLIAK, LAURENT, & YOR, MARC. 2003. *A.n. kolmogorov, quelques aspects de l'oeuvre probabiliste*.
- [Chavalarias, 1997] CHAVALARIAS, DAVID. 1997. *La thèse de popper est-elle réfutable?* Ph.D. thesis, CREA - CNRS/Ecole Polytechnique.
- [Chavalarias *et al.* , 2005] CHAVALARIAS, DAVID, CHARRON, SYLVAIN, GARDELLE, VINCENT ROGER DE, & BOURGINE, PAUL. 2005. *Nobel, le jeu de la découverte scientifique*. Tech. rept. halshs-00005009<sub>v1</sub>. HAL.
- [Cheeseman *et al.* , 1988] CHEESEMAN, P., GOEBEL, J., SELF, M., STUTZ, M., VOLK, K., TAYLOR, W., & WALKER, H. 1988. *Automatic classification of the spectra from the infrared astronomical satellite (iras)*. Tech. rept. Washington DC : National Aeronautics and Space Administration.
- [Clément *et al.* , 1994] CLÉMENT, E., DEMONQUE, C., HANSEN-LOVE, L., & KAHN, P. 1994. *La philosophie de a à z*. Hatier, Paris.
- [DaCosta & Beziau, 1997] DACOSTA, N. C. A., & BEZIAU, JEAN-YVES. 1997. La logique paraconsistante. *Logiques classiques et non classiques : essai sur les fondements de la logique*.
- [Dartnell & Martin, 2008] DARTNELL, CHRISTOPHER, & MARTIN, ÉRIC. 2008. Learning from each other. *In : Unpublished : Submitted to discovery science*.
- [Dartnell & Sallantin, 2005] DARTNELL, CHRISTOPHER, & SALLANTIN, JEAN. 2005. Assisting scientific discovery with an adaptive problem solver. *Pages 99–112 of : Discovery science*.
- [Delahaye, 1997] DELAHAYE, JEAN-PAUL. 1997. Statut mathématique des contradictions. *Pour la science*, Novembre.
- [Delahaye, 1999] DELAHAYE, JEAN-PAUL. 1999. *Information, complexité et hasard*. Paris : Hermès Publishing.
- [Demazeau, 2001] DEMAZEAU, YVES. 2001 (avril). *Voyelles*. Habilitation à diriger des recherches, Institut National Polytechnique de Grenoble INPG, Grenoble.

- [Dosen, 1984] DOSEN, K. 1984. Negative modal operators in intuitionistic logic. *Publications de l'institut mathématique*, **35(49)**, 3–14.
- [Došen, 1984] DOŠEN, K. 1984. Intuitionistic double negation as a necessity operator. *Publications de l'institut mathématique*, **35(49)**, 15–20.
- [Dubucs, 1992] DUBUCS, JACQUES. 1992. Pourquoi les mathématiques sont-elles difficiles? *Pages 210–217 of : First european congress of mathematics*.
- [Dunbar, 1993] DUNBAR, K. 1993. Concept discovery in a scientific domain. *Cognitive science*, **13**, 397–434.
- [Engel, 2007] ENGEL, PASCAL. 2007. Can there be an epistemic square of opposition? *Page 5 of : JEAN-YVES BÉZIAU, GILLMAN PAYETTE (ed), Handbook of the first world congress on the square of oppositions*.
- [Favre, 1995] FAVRE, D. 1995. Conception de l'erreur et rupture épistémologique. *Revue française de pédagogie*, 85–94.
- [Ferber & Gutknecht, 1998] FERBER, JACQUES, & GUTKNECHT, OLIVIER. 1998. A meta-model for the analysis and design of organizations in multi-agent systems. *Pages 128–135 of : Third international conference on multi-agent systems (icmas98)*.
- [Finin et al. , 1994] FININ, T., MCKAY, D., FRITZSON, R., & MCENTIRE, R. 1994. Kqml : An information and knowledge exchange protocol. *In : K. FUCHI, T. YOKOI (ed), Knowledge building and knowledge sharing*.
- [FIPA, 2002] FIPA. 2002. *Fipa agent management specification*. Tech. rept. xc00037. FIPA (Communicative Act Library Specification).
- [Fourez et al. , 1997] FOUREZ, G., ENGLEBERT-LECOMTE, V., & MATHY, P. 1997. *Nos savoirs sont nos savoirs*. DeBoeck Université.
- [Gardner, 1959] GARDNER, M. 1959. Mathematical games. *Scientific american*, June.
- [Gold, 1967] GOLD, E. MARK. 1967. Language identification in the limit. *Information and control*, **10(5)**, 447–474.
- [Goldblatt, 2005] GOLDBLATT, R. 2005. Mathematical modal logic : a view of its evolution. *In : D.M. GABBAY, J. WOODS (ed), Handbook of the history of logic, volume 6*. Elsevier BV.
- [Greimas & Courtès, 1993] GREIMAS, ALGIRDAS JULIEN, & COURTÈS, JOSEPH. 1993. *Sémiotique. dictionnaire raisonné de la théorie du langage*. Hachette, Paris.
- [Gutknecht & Ferber, 1997] GUTKNECHT, O., & FERBER, J. 1997 (Dec.). *MadKit : Organizing heterogeneity with groups in a platform for multiple multi-agent systems*. Tech. rept. 97188. LIRMM, 161, rue Ada - Montpellier - France.

- 
- [Hagège, 2007] HAGÈGE, HÉLÈNE. 2007. Jugement de valeurs, affects et conceptions sur l'élaboration du savoir scientifique : à la recherche d'obstacles à l'enseignement des questions vives. *Xviii<sup>èmes</sup> journées internationales sur la communication, l'éducation et la culture scientifiques, techniques et industrielles*. "école, culture et actualité des sciences techniques".
- [Hagège et al. , 2007a] HAGÈGE, HÉLÈNE, REYNAUD, C., & FAVRE, D. 2007a. *New conceptualisation of environmental attitudes : cut, relatedness and fusion towards the non-human environment -preliminary measure*. Environment and Behaviour.
- [Hagège et al. , 2007b] HAGÈGE, HÉLÈNE, DARTNELL, CHRISTOPHER, & SALLANTIN, JEAN. 2007b. Positivism against constructivism : A network game to learn epistemology. *In : Discovery science*.
- [Harthong, 1990] HARTHONG, J. 1990. Le continu ou le discret, un problème indécidable. *In : SPRINGER-VERLAG (ed), Le labyrinthe du continu*.
- [Harthong, 1992] HARTHONG, J. 1992. *L'idéalisme platonicien et la science*. [http ://moire4.u-strasbg.fr/phil/platon.htm](http://moire4.u-strasbg.fr/phil/platon.htm).
- [Horn, 2007] HORN, LAWRENCE R. 2007. Lexical pragmatics and the geometry of opposition. *Pages 5–6 of : JEAN-YVES BÉZIAU, GILLMAN PAYETTE (ed), Handbook of the first world congress on the square of oppositions*.
- [Howard et al. , 2000] HOWARD, B.C., MCGEE, S., SCHWARTZ, N., & PURCELL, S. 2000. The experience of constructivism : Transforming teacher epistemology. *Journal of research on computing in education*, **32**, 455–465.
- [Kant, 1845] KANT, E. M. M. 1845. *Critique de la raison pure (2<sup>de</sup> édition en français)*. Librairie Philosophique de Ladrangé, Paris.
- [Kauffmann, 1999] KAUFFMANN, THOMAS. 1999. *Eleusis, le jeu qui simule la recherche de la vérité*. Web : [http ://membres.lycos.fr/eleusis/](http://membres.lycos.fr/eleusis/).
- [King & Srinivasan, 1996] KING, R. D., & SRINIVASAN, A. 1996. Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. *Environmental health perspectives*, 1031–1040.
- [Klahr & Dunbar, 1988] KLAHR, D., & DUNBAR, K. 1988. Dual search space during scientific reasoning. *Cognitive science*, 1–48.
- [Kolmogorov, 1930] KOLMOGOROV, ANDREI NIKOLAEVITCH. 1930. *Grundbegriffe der wahrscheinlichkeitrechnung (fondements de la théorie des probabilités)*. Springer.
- [Kruskal, 1979] KRUSKAL, MARTIN. 1979. *Delphi : un jeu de raisonnement par induction*. Boston, Little Brown.

- [Kuhn, 1962] KUHN, T.S. 1962. *La structure des révolutions scientifiques*. Champs Flammarion.
- [Lacan, 1966 (réed. 1999)] LACAN, JACQUES. 1966 (réed. 1999). *écrits (séminaire xxii)*. éditions du Seuil, Paris.
- [Lakatos, 1976] LAKATOS, IMRE. 1976. *Proofs and refutations*. Cambridge University Press.
- [Langley *et al.* , 1981] LANGLEY, P., BRADSHAW, G. L., & SIMON, H. A. 1981. Bacon5 : The discovery of conservation laws. *Ijcai'81*.
- [Langley, 2000] LANGLEY, PAT. 2000. The computational support of scientific discovery. *Pages 393–410 of : Internationnal journal of human-computer studies*, vol. 53.
- [Langley, 2002a] LANGLEY, PAT. 2002a. Computational discovery of communicable scientific knowledge. *Logical and computational aspects of model-based reasoning*.
- [Langley, 2002b] LANGLEY, PAT. 2002b. Lessons for the computational discovery of scientific knowledge. *In : First international workshop on data mining lessons learned*. Sidney.
- [Laughlin, 2005] LAUGHLIN, ROBERT B. 2005. *Un univers différent*. Fayard.
- [Lecourt, 1999] LECOURT, D. 1999. *Dictionnaire d'histoire et philosophie des sciences*. Puf, Paris.
- [Lederman *et al.* , 2002] LEDERMAN, N.G., ABD-EL-KHADICK, F., BELL, R.L., & SCHWARTZ, R.S. 2002. Views of nature of science questionnaire : Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of research in science teaching*, **39**.
- [Lee *et al.* , 1998] LEE, Y., BUCHANAN, B. G., & ARONIS, J. M. 1998. Knowledge-based learning in explanatory science : Learning rules to predict rodent carcinogenicity. *Machine learning*, **30**, 217–240.
- [Lemberger *et al.* , 1999] LEMBERGER, N.G., HEWSON, P.W., & PARK, H.-J. 1999. Relationships between prospective secondary teachers' classroom practice and their conceptions of biology and of teaching science. *Science education*, **83**, 347–371.
- [Lemoisson *et al.* , 2005] LEMOISSON, P., CERRI, S. A., & SALLANTIN, J. 2005. Conversational interactions among rational agents. *Towards the learning grid : advances in human learning services*.
- [Lenat, 1977] LENAT, D. B. 1977. Automated theory formation in mathematics. *Pages 833–842 of : CAMBRIDGE, MA : MORGAN KAUFMANN (ed), Fifth international joint conference on artificial intelligence*.

- 
- [Liquière, 1998] LIQUIÈRE, M. 1998. Structural machine learning with galois lattice and graphs. *International conference on machine learning - icml*.
- [Lonka *et al.* , 1996] LONKA, K., JORAM, E., & BRYVIN, M. 1996. Conceptions of learning and knowledge : Does training make a difference. *Contemporary educational psychology*, **21**, 347–371.
- [Luzeaux *et al.* , 2008] LUZEAUX, DOMINIQUE, SALLANTIN, JEAN, & DARTNELL, CHRISTOPHER. 2008. Logical extensions of aristotle’s square. *In : Logica universalis*. Springer.
- [Machtey & Young, 1978] MACHTEY, M., & YOUNG, P. 1978. An introduction to the general theory of algorithms. *Theory of computation series*.
- [Maraschi *et al.* , 2001] MARASCHI, D., CERRI, S. A., MARTINENGO, G., DA NÓBREGA, G. M., SALLANTIN, J., DIVOL, J., & LIQUIÈRE, M. 2001. *A conversational, constructive view of web knowledge : the symbol level*. Tech. rept. 01221. Laboratoire d’Informatique, de Robotique et de Micro-électronique de Montpellier (LIRMM).
- [McCarthy, 1992] MCCARTHY, J. 1992. *Elephant 2000 : A programming language based on speech acts*. <http://www-formal.stanford.edu/jmc/elephant.html>.
- [Michel, 2004] MICHEL, FABIEN. 2004. *Formalisme, outils et éléments méthodologiques pour la modélisation et la simulation multi-agents*. Ph.D. thesis, Université Montpellier II.
- [Mitchell, 1997] MITCHELL, TOM. 1997. *Machine learning*. McGraw-Hill.
- [Moretti, 2004] MORETTI, A. 2004. Geometry for modalities ? Yes : through *n*-opposition theory. *Pages 102–145 of : J.-Y. BÉZIAU, A. COSTA-LEITE, A. FACCHINI (ed), Aspects of universal logic*. Travaux de logique 17, Neuchâtel.
- [Muggleton, 1994] MUGGLETON, S. 1994. Predicate invention and utilisation. *Theoretical artificial intelligence*, 127–130.
- [Muggleton & Buntine, 1988] MUGGLETON, S., & BUNTINE, W. 1988. Machine invention of firstorder predicates by inverting resolution. *Pages 339–352 of : Fifth international conference on machine learning*.
- [Nakamatsu *et al.* , 2003] NAKAMATSU, KASUMI, KATO, TAKUYA, & SUZUKI, ATSUYUKI. 2003. Basic ideas of defeasible deontic traffic signal control based on a paraconsistent logic program evalpsn. *Advances in intelligent systems and robotics*.
- [Neumann & Morgenstern, 1944] NEUMANN, J. VON, & MORGENSTERN, O. 1944. *Theory of games and economic behavior*. Princeton University Press, Princeton, New Jersey.

- [Nilsson, 1996] NILSSON, N. J. 1996. *Introduction to machine learning*. <http://robotics.stanford.edu/> - unpublished.
- [Nobrega *et al.*, 2003] NOBREGA, G. M. DA, CERRI, & SALLANTIN. 2003. A contradiction driven approach to theory information : Conceptual issues pragmatics in human learning, potentialities. *Journal of the brazilian computer society*, **9**, 37–55.
- [Okada & Simon, 1995] OKADA, TAKESHI, & SIMON, HERBERT A. 1995. Collaborative discovery in a scientific domain. *Page 340 of : Cognitive science society*.
- [Paulin, 2008] PAULIN, MATHIAS. 2008. *Contributions à l'apprentissage automatique de réseau de contraintes et à la constitution automatique de comportements sensorimoteurs en robotique*. Ph.D. thesis, Université Montpellier II.
- [Pellissier, 2006] PELLISSIER, R. 2006. “Setting” *n*-opposition. *In : Unilog05*.
- [Piaget, 1924] PIAGET, J. 1924. *Le jugement et le raisonnement chez l'enfant*. Neuchâtel, Paris : Delachaux et Niestlé.
- [Piaget, 1972a] PIAGET, JEAN. 1972a. *essai de logique opératoire*. Dunod, Paris (2<sup>e</sup> éd.).
- [Piaget, 1972b] PIAGET, JEAN. 1972b. *L'épistémologie génétique*. PUF, Paris.
- [Popper, 1963] POPPER, KARL RAIMUND. 1963. *Conjectures and refutations : The growth of scientific knowledge*. Harper and Row.
- [Sallantin, 2004] SALLANTIN, JEAN. 2004. Que peut-on apprendre en interagissant ? *troisièmes journées “complexités”, perception et apprentissage*.
- [Sammut & Banerji, 1986] SAMMUT, C., & BANERJI, R. 1986. *Learning concepts by asking questions*. Vol. 2. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.
- [Schommer, 1994] SCHOMMER, M. 1994. Synthesizing epistemological belief of research : tentative understandings and provocative confusions. *Educational psychology review*, 193–319.
- [Searle, 1969] SEARLE, J. 1969. *Speech acts*. Cambridge University Press.
- [Shapiro, 1981] SHAPIRO, EHUD Y. 1981. An algorithm that infers theories from facts. *Pages 446–451 of : HAYES, PATRICK J. (ed), Ijcai*. William Kaufmann.
- [Simon, 1966] SIMON, H. A. 1966. Scientific discovery and the psychology of problem solving. *In : COLODNY, R. (ed), Mind and cosmos*. University of Pittsburgh Press.
- [Stevenson, 1999] STEVENSON, I. 1999. What are the irreducible components of the scientific enterprise ? *Journal of scientific exploration*, **13**(2), 257–270.

- 
- [Strike & Posner, 1992] STRIKE, K.A., & POSNER, G.J. 1992. *A revisionist theory of conceptual change*.
- [Valiant, 1984] VALIANT, L.G. 1984. A theory of the learnable. *Commun. acm*, **27**, 1134–1142.
- [Waeytens *et al.* , 2002] WAEYTENS, K., LENS, W., & VANDENBERGHE, R. 2002. Learning to learn : Teachers conceptions of their supporting role. *Learning and instruction*, **12**, 305–322.
- [Witten *et al.* , 1996] WITTEN, IAN H., NEVILL-MANNING, CRAIG, & MAULSBY, DAVID. 1996. *Interacting with learning agents : implications for ml from hci*.
- [Zimmerman, 2000] ZIMMERMAN, C. 2000. The development of scientific reasoning skills. *Developmental review*, **20**, 99–149.
- [Zvonkin & Levin, 1970] ZVONKIN, A K, & LEVIN, L A. 1970. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russ math surv*, **25**(6), 83–124.
- [Zytkow & Simon, 1986] ZYTKOW, J. M., & SIMON, H. A. 1986. A theory of historical discovery : The construction of componential models. *Machine learning*, **1**, 107–137.



# Annexe A

## Publications

- [Sallantin et al, 2003] Jean SALLANTIN, Christopher DARTNELL, Jacques DIVOL et Patrice DUROUX. « Conceptual Framework for Interactive Ontology Building ». Dans les actes de *Second IEEE International Conference on Cognitive Informatics (ICCI'03)*, page 179, 2003.
- [Sallantin et al, 2004] Jean SALLANTIN, Christopher DARTNELL, Jacques DIVOL, Michel LIQUIÈRE, Paul BOURGINE, Sylvain CHARRON, David CHAVALARIAS, Denis PIERRE, Christophe FAGOT et Clément Durand DAUBIN. « E+N : une plateforme pour stimuler l'activité collective de découverte scientifique ». *ALCAA (Agents Logiciels - Coopération - Apprentissage - Activités Humaines)*, 2004.
- [Dartnell et Sallantin, 2005] Christopher DARTNELL et Jean SALLANTIN. « Assisting Scientific Discovery with an Adaptive Problem Solver. ». Dans les actes de *Discovery Science*, pp 99–112, 2005.
- [Sallantin et al, 2006] Jean SALLANTIN, Christopher DARTNELL et Mohammad AFSHAR. « A Pragmatic Logic of Scientific Discovery ». Dans les actes de *Discovery Science*, volume 4265 de *Lecture Notes in Computer Science*, pp 231–242. Springer, 2006.
- [Paulin et al, 2006] M. PAULIN, E. BOURREAU, C. DARTNELL et S. KRUT. « Modélisation et Planification d'Actions Élémentaires Robotiques par Apprentissage de Réseaux de Contraintes ». Dans les actes de *Deuxièmes Journées Francophones de Programmation Par Contraintes (JFPC'2006)*, 2006.
- [Sallantin et al, 2006] J. SALLANTIN, C. DARTNELL et M. AFSHAR. Interactive Learning from Contradictions in a Paraconsistent Logic. Dans les actes de *16th international Conference on Inductive Logic Programming*, 2006.
- [Dartnell et al, 2007] Christopher DARTNELL, Dominique LUZEAUX et Jean SALLANTIN. « Aristotle's Square in a Logic of Scientific Discovery. ».

Dans les actes de *Handbook of the First World Congress on the Square of Opposition*, 2007.

[Afshar et al, 2007] M. AFSHAR, C. DARTNELL, D. LUZEAUX, Y. Tognetti et J. SALLANTIN. « Aristotle's Square Revisited to Frame Discovery Science ». *Journal of Computers*, 2007.

[Hagège et al, 2007] Hélène HAGÈGE, Christopher DARTNELL et Jean SALLANTIN. « Positivism Against Constructivism : A Network Game to Learn Epistemology. ». Dans les actes de *Discovery Science*, 2007.

[Luzeaux et al, 2008] Dominique LUZEAUX, Jean SALLANTIN et Christopher DARTNELL. « Logical Extensions of Aristotle's Square ». Dans les actes de *Logica Universalis*. Springer, 2008.

[Dartnell et al, 2008] Christopher DARTNELL, Éric MARTIN , Hélène HAGÈGE et Jean SALLANTIN. Human Discovery and Machine Learning. Dans les actes de *International Journal of Cognitive Informatics and Natural Intelligence*, 2008.

[Dartnell et Martin, 2008] Christopher DARTNELL, Éric MARTIN et Jean SALLANTIN. « Learning From Each Other ». Dans les actes de *Discovery Science*, 2008.

# Annexe B

## Deux Définitions de la Complexité : Complexité Aléatoire et Complexité Organisée

CETTE annexe est extraite de [Beurier, 2007], chapitre 2, avec l'accord de l'auteur bien entendu. Je ne référence ici que les discussions concernant deux mesures de complexité, introduites par Kolmogorov (Section B.1) et Bennett (Section B.2). Pour une discussion plus complète sur les systèmes complexes, je renvoie le lecteur à [Beurier, 2007] qui dresse un état de l'art complet du problème.

### B.1 Théorie de la Complexité de Kolmogorov

La théorie algorithmique de l'information est apparue suite à un constat : les données du monde réel ainsi que les données artificielles<sup>25</sup> ne suivent pas des lois de probabilités uniformes. Cela peut être illustré en regardant la mémoire d'un ordinateur ; si l'on observe de façon aléatoire les données stockées dans cette mémoire, il y a de forte chance de tomber sur une séquence régulière de bits (une suite de 0 étant la plus probable). Pour tenter d'expliquer cet état de fait, Solomonof, Kolmogorov, Chaitin, Martin-Löf et Levin, ont proposés [Kolmogorov, 1930, Chaumon *et al.*, 2003], d'analyser l'information contenue dans un objet/système en fonction de la plus courte description que l'on peut en faire (implicitement du plus court programme capable de générer cette description) ; cette description étant alors ramenée à sa plus simple expression binaire (une suite de 0 et de 1).

---

25. Les données informatiques par exemple.

La complexité de Kolmogorov d'une suite  $s$ , notée  $K(s)$ , est la taille (en bits) du plus petit programme, noté  $s^*$ , capable de produire  $s$ . On voit immédiatement que cette complexité est étroitement liée aux méthodes de compression de données sans pertes. Une suite de bits pouvant être grandement compressée possède ainsi une complexité inférieure à une suite de bits de même taille ne pouvant pas l'être ou ne pouvant l'être que moins. De même, une suite de bits de faible complexité (au sens de Kolmogorov) pourra être compressée de meilleure façon qu'une suite de forte complexité.

Les techniques de compressions sans pertes sont une bonne approche pour estimer la complexité d'une suite de bits néanmoins, elles ne permettent pas d'obtenir de façon certaine la meilleure compression. Gödel a en effet montré en 1931, en développant son théorème d'incomplétude, qu'il n'existe aucune méthode mathématique pour calculer exactement  $K(s)$  pour chaque  $s$  donné. En effet, par le théorème d'incomplétude, Gödel a montré que certains énoncés exprimables dans le langage d'une théorie  $T$ , et dont on est certain qu'ils sont vrais (parce qu'on les démontre à l'aide d'autres théories), ne sont pas démontrables avec les moyens de  $T$ . Une théorie mathématique ne connaît donc  $K(s)$  que pour quelques suites  $s$ , et au-delà d'une certaine taille pour  $s$ , elle ne connaît plus aucune valeur de  $K(s)$ . Cette indétermination a une raison profonde : des structures, invisibles au premier abord, peuvent se trouver cachées dans les objets, qui apparaissent alors, à tort, aléatoires ou chaotiques.

Bien qu'introduisant des concepts forts attractifs, la théorie de la complexité de Kolmogorov est par nature difficile à mettre en oeuvre pour l'analyse de systèmes complexes. Elle est donc présentée ici à but informatif. Nous verrons cependant par la suite qu'elle peut éclairer de façon formelle la vision que l'on peut avoir de l'organisation d'un système complexe (voir section suivante).

La mesure de complexité de [Zvonkin & Levin, 1970], directement inspirée de la théorie de la complexité de Kolmogorov, a été développée suite à un constat : les systèmes/objets à l'état naturel présentent des symétries/structures qui rendent leur compression plus facile. Levin part de cet état de fait et définit une mesure de complexité liée à la probabilité qu'un programme à de générer la suite  $s$ . Cette mesure,  $m(s)$ , directement liée à la complexité de Kolmogorov, est défini pour une suite  $s$ , par :

$$m(s) = \left(\frac{1}{2}\right)^{K(s)}$$

Cette équation exprime qu'un objet résultant d'un programme court à une probabilité plus forte ( $m$  plus grand) qu'un objet résultant d'un programme long. Il est intéressant de voir que partant de ce postulat, on peut énoncer que les explications les plus simples sont plus probables que les explications les plus longues ou les plus complexes (adage ô combien primordial en science!).

## B.2 La profondeur logique de Bennett : complexité aléatoire et complexité organisée

Comme nous avons pu le voir, la complexité de Kolmogorov d'un système mesure son contenu d'information incompressible, son désordre, sa composante aléatoire. Grâce à cette notion mathématique parfaitement précise la notion de complexité prend un sens formel dépourvu de toute ambiguïté. Cependant la complexité qui est approchée via cette méthode n'est que celle donnée par le désordre d'un mélange qui ne produit rien de nouveau : c'est par exemple la complexité du détail d'un tas de sable ou d'une suite de tirages de piles ou faces. Cette méthode ne permet en effet pas d'aborder correctement la complexité d'un organisme vivant, ni celle d'une ville par exemple ; complexité qui présente des structures organisées.

Ainsi, on peut distinguer deux types de complexité : la complexité aléatoire, qui est la complexité d'un désordre sans règle, et la complexité organisée, la richesse en structures et sous-structures que l'on peut trouver dans les systèmes tels que les organismes vivants, les organisations sociales et les machines artificielles, etc. Comment alors formuler une définition mathématique qui soit pour la complexité organisée ce que la complexité de Kolmogorov est pour la complexité aléatoire ? (Question posée par [Delahaye, 1999])

A cet effet, le physicien [Bennett, 1988] a proposé une définition complémentaire de celle de Kolmogorov qui aborde la complexité non plus seulement en terme de taille de programme générateur ou de description du système mais également en temps de calcul nécessaire à ce même programme pour produire la description du système. Pour une suite binaire  $s$ , ce temps de calcul est alors appelé profondeur logique de  $s$  et noté  $P(s)$ .

Les implications de la profondeur logique sont immédiatement visibles lorsque l'on s'intéresse aux systèmes complexes. En effet, les profondeurs logiques d'un système simple et d'un système aléatoire sont toutes deux très faibles : ce sont des objets sans richesse de structure, sans contenu authentique en information. Dans le cas du système simple, un programme court permettra de générer la description en mettant peu de temps à s'exécuter. Dans le cas d'un système aléatoire, un programme long (au minimum de la taille du système dans le cas d'un système complètement aléatoire), mettra également peu de temps à s'exécuter (*print* de la description du système).

En revanche, la profondeur logique d'un objet/système finement organisé (comme un système complexe biologique) sera sensiblement plus élevée : en effet, le programme sera certes plus court que dans le cas d'un système purement aléatoire mais son exécution exploitera les propriétés et régularités particulières du système décrit, ce qui demandera nécessairement un grand temps de cal-

cul. Dans de tels cas, le passage du plus court programme à la description du système n'est plus la simple exécution d'un *print* mais le parcours d'un chemin computationnel, riche en boucles récursives et appel de sous-procédures.

Les systèmes complexes biologiques portent généralement des traces de ces deux types de complexité. La structure d'un être vivant, par exemple, dont l'essentiel est tiré du génome, est intuitivement de l'ordre de la complexité organisée. Le "calcul" qui lui a donné naissance est celui effectué de manière indirecte pendant des centaines de millions d'années par l'évolution. Quant à elle, la complexité aléatoire d'un être vivant évolué est celle qui fixe, par exemple, l'emplacement précis des cheveux ou des vaisseaux sanguins, dont les positions dans le détail ne sont pas le simple résultat de l'interprétation de l'information génétique. Comme le note d'ailleurs [Delahaye, 1999] : en négligeant la complexité organisée contenue dans le cerveau et provenant de l'apprentissage, on pourrait dire par exemple que deux vrais jumeaux portent en eux le même contenu en complexité organisée, et diffèrent très fortement pour ce qui est de leur complexité aléatoire.

La profondeur logique de Bennett a l'avantage d'aborder des problématiques généralement délaissées par la théorie de la complexité. En effet, en proposant la loi de croissance lente, l'approche de Bennett permet ainsi d'aborder le concept d'émergence du point de vue de la complexité.

La loi de croissance lente peut brièvement s'expliquer comme suit : lors du déroulement d'une dynamique, l'évolution de la complexité de Kolmogorov dans un système peut être brusque. Par exemple, en jetant un verre de cristal au sol, on passe d'une complexité de Kolmogorov faible (le verre est un objet simple) à une complexité assez forte : le détail des morceaux brisés a soudainement accru la complexité aléatoire de l'objet. En revanche, la profondeur logique d'un système donné ne peut quasiment jamais croître brusquement ; c'est cela que Bennett nomme la loi de croissance lente. Ainsi, un système richement structuré et organisé ne peut pas apparaître de rien, instantanément, mais demande un long processus d'interactions entre ses divers éléments, c'est-à-dire une sorte de calcul prolongé et cumulatif. La profondeur logique de Bennett permet ainsi de mesurer la quantité de calcul fixée dans un système ; c'est une mesure du contenu computationnel du système et une mesure de la longueur de la dynamique qui y a donné naissance.

**Annexe C**

**Questionnaire Présenté lors des  
Expériences**

Jeudi 15 mars 2007

J'ai joué au jeu E+N :  Oui  Non

NUMERO D'ANONYMAT

Jour de naissance de votre maman : |\_|\_| Jour de naissance de votre papa : |\_|\_|

Début du prénom de votre maman : |\_|\_| Début du prénom de votre papa : |\_|\_|

RENSEIGNEMENTS PERSONNELS

P1. Age : \_\_\_\_\_ ans

P2. Sexe :  Féminin  Masculin

P3. Quel est, ou était le domaine d'activité professionnelle de vos parents ? Cochez la ou les case(s) pertinente(s) :

Education scientifique  Recherche scientifique  Agriculture  Industrie  Santé  Autre

**Première partie : questionnaire**

A propos de la science						
1	Le savoir scientifique se construit.	D'accord				Pas d'accord
2	L'objectivité est intrinsèque à l'activité scientifique.	D'accord				Pas d'accord
3	La subjectivité est intrinsèque à l'activité scientifique.	D'accord				Pas d'accord
4	La science produit des connaissances qui s'amoncellent progressivement.	D'accord				Pas d'accord
5	Le savoir produit par la science repose aussi sur les préjugés et les opinions des scientifiques.	D'accord				Pas d'accord
6	Toute observation scientifique est chargée de théorie.	D'accord				Pas d'accord
7	Le progrès de la science consiste en une accumulation graduelle de connaissances.	D'accord				Pas d'accord
8	Tous les chercheurs d'un domaine sont égaux face à une expérience.	D'accord				Pas d'accord
9	Tous les chercheurs d'un domaine sont égaux face à une théorie.	D'accord				Pas d'accord
10	Quelque chose est dit « scientifique » uniquement parce que les scientifiques en ont décidé ainsi.	D'accord				Pas d'accord
11	Toute observation scientifique est neutre.	D'accord				Pas d'accord
12	Toute observation scientifique est objective.	D'accord				Pas d'accord
13	Toute observation scientifique est subjective.	D'accord				Pas d'accord
14	Toute connaissance est subjective.	D'accord				Pas d'accord
15	Un chercheur isolé peut faire de la science.	D'accord				Pas d'accord
16	Du point de vue scientifique, il y a une seule bonne interprétation des faits.	D'accord				Pas d'accord
17	Un observateur scientifique n'interprète en rien ce qu'il voit.	D'accord				Pas d'accord
18	Les théories scientifiques sont des inventions.	D'accord				Pas d'accord

19	La notion d'atome est une invention.	D'accord							Pas d'accord
20	Les atomes existent vraiment dans la nature.	D'accord							Pas d'accord
21	La notion d'atome est une découverte.	D'accord							Pas d'accord
22	Le résultat d'une expérience impose une conclusion.	D'accord							Pas d'accord
23	Les théories scientifiques guident les conclusions issues des expériences.	D'accord							Pas d'accord
24	Il y a toujours plusieurs interprétations possibles à un résultat d'expérience.	D'accord							Pas d'accord
25	Observer, c'est recevoir passivement de l'information.	D'accord							Pas d'accord
26	Observer, c'est construire une interprétation de ce que l'on a vu.	D'accord							Pas d'accord
27	Il y a des connaissances scientifiques qui sont considérées comme acquises et sur lesquelles on ne reviendra jamais.	D'accord							Pas d'accord
28	Toute théorie scientifique est susceptible d'être remise en cause dans le futur.	D'accord							Pas d'accord
29	L'élaboration des théories scientifiques implique de la créativité.	D'accord							Pas d'accord
30	On peut dire de certaines connaissances scientifiques qu'elles sont vraies.	D'accord							Pas d'accord
31	Avant, il y avait des théories qui étaient fausses mais maintenant, on tend de plus en plus vers la vérité.	D'accord							Pas d'accord
32	Les chercheurs n'utilisent pas leurs croyances pour faire de la science.	D'accord							Pas d'accord
33	Lorsqu'un chercheur travaille, il procède selon des étapes bien déterminées – la démarche scientifique.	D'accord							Pas d'accord
34	Si le résultat d'une expérience est incompatible avec une théorie scientifique, la théorie sera nécessairement remise en question.	D'accord							Pas d'accord
35	L'erreur est toujours évitée en science.	D'accord							Pas d'accord
36	Parfois, je ne crois pas les faits relatés dans des livres écrits par des experts.	D'accord							Pas d'accord
37	Même les conseils prodigués par des experts devraient être soumis au questionnement.	D'accord							Pas d'accord
38	Les scientifiques parviendront ultimement à trouver la vérité s'ils continuent de la chercher.	D'accord							Pas d'accord
39	La connaissance scientifique est certaine et ne change pas.	D'accord							Pas d'accord
40	Pour formuler des lois scientifiques, on étudie plusieurs cas particuliers puis on généralise.	D'accord							Pas d'accord
41	D'abord on formule des lois scientifiques générale puis on essaie de les vérifier par des études de cas.	D'accord							Pas d'accord
42	On peut dire qu'une loi est vraie quand on a montré qu'elle s'appliquait partout où a essayé de l'appliquer.	D'accord							Pas d'accord
43	En fait quand on dit qu'une loi est « vraie », ça veut seulement dire qu'on n'a pas (encore) trouvé de cas où elle ne marchait pas.	D'accord							Pas d'accord
44	C'est l'addition des travaux individuels des chercheurs qui augmente le savoir scientifique.	D'accord							Pas d'accord
45	C'est à partir des confrontations au sein de la communauté scientifique que naît le savoir scientifique.	D'accord							Pas d'accord
46	Même un chercheur isolé peut décider si une connaissance donnée peut être ajoutée au savoir scientifique.	D'accord							Pas d'accord
47	Seule la communauté scientifique peut décider si une connaissance donnée peut être ajoutée au savoir scientifique.	D'accord							Pas d'accord

## Deuxième partie : adjectifs

Comment percevez-vous le « savoir scientifique », la « science », « l'erreur », « l'apprentissage », ... ?

Cochez 1 case sur 5 par ligne (pour chaque paire d'adjectif). Cochez la case la plus proche de l'adjectif qui vous semble le mieux caractériser le sujet indiqué en gras :

<b>savoir scientifique</b>	
beau	laid
exact	approximatif
faux	vrai
universel	contextuel
subjectif	objectif
bon	mauvais
issu de la raison	issu de l'imagination
donné	créé
définitif	temporaire
relatif	absolu
réaliste	idéaliste
mal	bien
précis	imprécis

<b>science</b>	
belle	laide
exacte	approximative
fausse	vraie
universelle	contextuelle
subjective	objective
bonne	mauvaise
précise	imprécise
abstraite	concrète

<b>erreur</b>	
belle	laide
utile	inutile
négative	positive
agréable	douloureuse
subjective	objective
bonne	mauvaise
intéressante	inintéressante
relative	absolue
fait peur	fait envie
mauvaise	bonne
scientifique	non scientifique

<b>enseignement</b>	
beau	laid
faux	vrai
subjectif	objectif
bon	mauvais
dévolutif	directif
subjectif	objectif
transmissif	accompagnateur
actif	passif

### apprentissage

beau					laid
faux					vrai
subjectif					objectif
bon					mauvais
universel					contextuel
subjectif					objectif
cumulant					réorganisant
définitif					temporaire
actif					passif

### connaissance

belle					laide
utile					inutile
négative					positive
agréable					douloureuse
subjective					objective
bonne					mauvaise
intéressante					inintéressante
relative					absolue
fait peur					fait envie
mauvaise					bonne
additionnée					organisée

### preuve

Relative					Absolue
Expérimentale					Théorique
Statistique					Logique
Définitive					Temporaire
Complexe					Simple
Individuelle					Collective

### réfutation

Relative					Absolue
Expérimentale					Théorique
Statistique					Logique
Définitive					Temporaire
Complexe					Simple
Individuelle					Collective

## Troisième partie : problème non routinier

Cette situation a lieu dans une société imaginaire. Vous êtes chercheur dans une communauté scientifique. La proposition A est admise par la communauté scientifique, car elle a été confirmée de nombreuses fois. Un début de recherche vous a apporté les résultats B et C.

*NB* : La mistonite est un animal.

A. Toutes les mistonites sont noires.

B1. Des mistonites ont été isolées à Trumus-les-Eaux.

B2. Elles sont marron clair.

B3. Une découverte fortuite a mis en évidence qu'elles sont infectées par le virus CP (qui ne semble pas affecter leur longévité).

- C1. Les mistonites de Rocheblais sont noires.
- C2. Elles ne sont pas infectées par le virus CP

**1/** Discutez vos résultats

**2/** Maintenant (d'après vos résultats), qu'envisagez précisément vous de faire pour faire progresser la science ? Vous pouvez discuter les différents cas de figure.

## **Quatrième partie : mots**

a) **Donnez 3 mots ou expression que vous associez à la communication des résultats au sein de la communauté scientifique :**

- 
- 
- 

b) **Donnez 3 mots ou expression que vous associez à la découverte scientifique :**

- 
- 
- 

c) **Cochez les TROIS mots qui vous semblent le mieux évoquer le processus de la découverte scientifique :**

hypothèse  résultat  expérience  difficile  réfutation  réflexion  long  collectif

## **Cinquième partie : impressions sur le jeu E+N**

C1. **Nom de l'équipe dans laquelle vous avez joué :** \_\_\_\_\_

C2. **Nombre de joueurs dans l'équipe :** \_\_\_\_

C3. **Est-ce que le jeu vous a plu ?** pas du tout  beaucoup

**Pourquoi ?**

C4. **Quelles émotions avez-vous ressenties en jouant ?**

C5. **Avez-vous des améliorations à proposer ? Si oui, lesquelles ?**



## Annexe D

# Évolution des formalismes logiques

# Aristotle's Square Revisited to Frame Discovery Science

Mohammad Afshar, Christopher Dartnell, Dominique Luzeaux, Jean Sallantin, Yannick Tognetti

Ariana Pharmaceuticals, Pasteur Biotop, 28 rue Dr Roux Paris 75724 - France

Email: {m.afshar, y.tognetti}@arianapharma.com

Euriware, 44 Rue des Vindits, 50130 Cherbourg-Octeville - France

Email: christopher.dartnell@gmail.com

Centre Technique des Systèmes d'Information, 24 av. Prieur de la Côte d'Or 94117 Arcueil Cedex France

Email: dominique.luzeaux@polytechnique.org

LIRMM, CNRS UM2, 161 rue Ada, 34392 Montpellier Cedex 5 - France

Email: js@lirmm.fr

**Abstract**—The paper attempts to give a formal framework to capture the entire process of scientific discovery including hypothesis formation, reasoning, identifying contradictions, peer reviewing, reformulating and so on. Data mining can be seen as one step in this complex process of interactive learning of an empirical theory. This paper uses the terminology from paraconsistent logic and paracomplete logic that extends Aristotle square in a hypercube of oppositions which defines or substantiates any step of the discovery process.

The central formal notions are validated on a mathematical scientific discovery game, and an industrial application in the field of Drug Discovery illustrates how the presented framework combines different learning processes to predict pharmaco-kinetic properties (ADME-T) and adverse side effects of therapeutic drug molecules.

**Index Terms**—Machine Learning, Scientific Method, Logical Reasoning Framework, Aristotle's Square of Oppositions

## I. INTRODUCTION

Our objective is to propose a logical framework to assist scientists in supervising the entire process of theories formation while studying a phenomenon, as it is done by chemists studying properties of a new molecule, such as absorption or toxicity.

The first piece of the framework's logical structure is known since medieval logic as Aristotle's square [1], and has appeared in logical texts ever since: Aristotle defines syllogistic reasoning by differentiating universal and particular statements, and linking them by deduction and negation. The detection of contradictions then occupies the diagonal of such a square and can warn the scientist that a revision of knowledge is necessary (Figure 1).

The framework is formulated in a logical form such that contradictions can occur: in logic, a contradiction is produced by the conjunction of a formula and its negation. Each logic defines its own negation by an axiomatic schema expressing relations between universal and existential modalities, as "Necessary" (positive universal) and "Contingent" (negative existential). In the case of scientific discovery, events can be declared contingent when they sometimes occur by chance, or can be declared necessary if they have to exist. A theory then presents a

contradiction when it allows to deduce, for instance, that a contingent event occurs by necessity.

Since Platon and Aristotle, logicians created different logics by selecting axioms that impose for instance that *a necessary action is a possible action*. But all these logics exhibit paradoxes provoked by their axioms, which often lead to their triviality. However, we admit that contradictions occur during a causal reasoning, and they are used in this framework to alert about problems concerning the consistency or the completeness of the ongoing theory.

Let us sketch such a formation process (a definition of unintuitive terms is provided in sections IV-A and V-A). Since Popper, it is admitted that a scientific theory must be refutable by experimentation and empirical data. A scientific experimentation, designed to enable empirical proof or refutation, requires the use of an accurate and efficient instrumentation to determine the existence of positive observations used to formulate postulates and conjectures. This experimentation design is a tool to ensure the progression of the ongoing theory by revealing facts. A fact is a piece of information (data) having an undeniable empirical evidence for scientists, and techniques such as data mining consist in the induction of a model from these facts. Such a model is built to match with positive information describing facts and to prevent the prediction of non observed facts. Predictions are made using a theory completed by a model, and the consideration of the overfitting and underfitting of these predictions informs the scientists on the bias conditions making it possible to decide correctly with the ongoing theory.

The logical framework presented in this paper is designed to take into account the two dynamics of scientific discovery [2] [3]. The first dynamic, which we refer to as the *personal dynamic*, embraces the supervision of a computer assistant by a scientist. This dynamic is centered on individual behaviors, and depends strongly on the research strategy of each individual and on its use of computer assistants. In fact, more and more learning techniques and data mining tools are used to find correlations in data and propose models to explain a studied phenomenon

and make predictions. We will see how the framework presented in this paper takes into account this personal dynamic by giving logical modalities to the statements occurring during the construction of a theory, and by placing the interpretation of contradictions in the heart of the interactive process leading to the construction of a model which can be discussed, justified, and proven in pure logic. The second dynamic we are concerned with, which we refer to as the *social dynamic*, comes from the collective behavior emerging from the social game during which scientists publish their theories and confront them to understand how models in a given domain match or influence models concerning other domains. In fact, scientists are experts in the different scientific fields involved in the understanding of a phenomenon, and we suppose that it is not always possible to merge all the scientific domains in a only one. In other words, scientists can all use their own models and data to compute and predict (personal dynamic). But by confronting their works, they may realize that a good solution for a given model can be a bad one for another model, and therefore point out the need to improve their understanding of the phenomenon. For instance, a model of Absorption may not take into account the toxicity of the molecule. To enable this dynamic, the framework proposed in this paper takes into account the process of crisis, conflict and transactions which continuously rhythms Science and contributes to the formation of theories admitted by the society of scientists as being scientifically valid.

We illustrate in section II the recurrent problems posed by the personal and social dynamics, by describing the domain of Drug discovery and the prediction of ADME-T properties. The pharmaceutical industry is indeed confronted to a pressing need to analyze ever growing quantities of collected data and convert them into relevant decisions, using cheminformatics methods. Most often, reliable predictions are only possible on molecules very similar to the learning set, and these predictions use descriptions which are not easy to be translated in better molecular structures. This domain concentrates many challenges for inductive learning : the bias problem, the underfitting/overfitting problem, the constraint satisfaction problem, the multiparametric decision problem, the empirical testing and the interactive problem solving by scientists assisted by machine learning. As there are hundreds of descriptors used to describe the surface of the molecules and thousands related to the computation of properties for a single molecule, the risk of overfitting is permanent. When we try to limit the number of descriptors to reduce this overfitting, we create a risk of falling in the adverse case of underfitting. So the balance between completeness and consistency corresponds to the balance between underfitting and overfitting. As learned rules in drug design can conjugate more than three terms, the constraint satisfaction problem is divided in two domains: the domain where deciding with learned rules is easy but produce errors, and the domain where finding rules to decide is so difficult that the system learns

by heart and can only decide for molecules which are very similar to the examples used during learning. Finally the multiparametric decision, required to deal with a distributed set of constraints that is not convex, often leads to an antagonism between optimized decisions. Furthermore, examples from disparate domains such as medical, chemical, legal, ... have different theoretical basis. Therefore, the regularities learned from these examples cannot be justified only by causal arguments.

After this introductory example, we present step by step the four levels of our formalism. Section III describes Aristotle's square of oppositions, which defines the oppositions between universal and particular modalities, therefore enabling mathematical and logical reasoning on simple problems. In section IV, this square is used to formalize the notions of postulate and conjecture, which are necessary to define a theory. The cubic structure presented in this section is obtained by linking the various squares together. section IV-B offers an intuitive illustration of these notions. Section V then presents a hypercubic structure obtained by linking the cube of oppositions to two new Aristotle's squares introduced to define the modalities of model, experimentation, prediction and facts. We emphasize the fact that the cube of oppositions lacked of temporality, and that this hypercubic structure links reasoning on facts and postulates to decision making and action, which are sufficient to define an agent. Finally, section VI establishes a link to deontic, autepistemic, and defeasible modalities which are produced when agents using this hypercube to reason interact by implementing a conversational process of discovery and learning from each other. In conclusion we discuss how the objective to capture the entire process of scientific discovery is achieved starting from Aristotle's square to frame computer agents assisting humans during a problem resolution.

## II. THE PROBLEM OF DRUG DISCOVERY, ADME-T

Schematically, the pharmaceutical activity can be divided into three sectors: drug discovery (i.e. going from a target to a molecule that is ready to be tested in man), drug development (i.e. the proof of concept in man and the clinical trials) and finally the marketing and monitoring of the product.

It is widely accepted that out of one hundred drug discovery projects that are started, less than one would eventually reach the market ten to fifteen years later.

Despite over a decade of massive investment by the pharmaceutical industry into high throughput methods (Genomics, High Throughput Screening and combinatorial chemistry), efficient identification and optimization of potent and quality lead molecules is still the highest and riskiest hurdle in current drug discovery and development. The only clear outcome of high throughput methods has been an unparalleled production of large quantities of data that need to be analyzed.

In order to reduce risks in the clinical stages of development, in a typical lead optimization process, 40 to 60

assays are run in parallel or in a cascade to evaluate the potential of each candidate molecule, its specificity, its good Absorption and Distribution, good Metabolism and Excretion profiles and limited Toxicity (ADME-T). In this multi-parametric space, identifying “quality” molecules which display desirable properties is a true challenge.

The use of computational tools (data mining, predictive modeling etc) has been seen as the potential solution to this dramatic inefficiency.

QSAR (Quantitative Structure Activity Relation) equations are standard examples in predictive modeling for drug discovery where an overall fitness score is developed as a weighted sum of numerous descriptors. In Docking, the score includes ligand internal energy, interaction energy and entropic considerations in the form of a weighted sum of terms [4]. Typically the score is developed empirically by analyzing a set of examples and deriving a weighted sum. The weights are fitted to the learning set and may not necessarily be relevant or precise for other complexes.

QSAR attempts to relate a numerical description of a molecular structure to a known biological activity. Large numbers of readily computable descriptors are available, in combination to sophisticated techniques that improve the initial linear regression analysis methods used in deriving QSAR equations (PCA, PLS, NN, GA, SVM etc). In general, QSAR equations relate one objective (such as activity for example) with a number of descriptors. QSAR equations are constructed by the combination of a number of weighted terms (descriptors).

These methods rely on the choices of (1) the descriptors for “generalization” and (2) the examples in the learning set to avoid overfitting.

An inadequate choice of either parameters will generally lead to useless models that do not generalize or are not interpretable. This is also true for initial ILP approaches [5]. In addition, search strategies can be compromised when confronted to non-convex solution fronts, i.e. when a solution “between” two valid solutions might be invalid. Furthermore scale invariance is not always true, i.e. even for a continuous property such as molecular weight, its use and therefore significance is distinct for different ranges (for example 200-600 range correspond to small molecules, a molecular weight greater than 2000 does not). This is to say that some relations are sensitive to scale. More generally, qualities can be converted into quantities (binning) but the reverse is not always true. This requires defining domains of validity for all parameters, in both the search and the objective spaces. In turn the notion of domain is linked to boundaries and hence allows characterization of paradoxical combinations or conflicts. Here, conflicts are real mutual exclusions rather than a competition between several continuous parameters.

All in all, it is fair to say that the current state of the art in cheminformatics is insufficient: “In general, reliable predictions are only possible for molecules similar to those in the training set” [6] hence undermining their predictive use, and “most models [...] use descriptors that

are not easily understood by the chemist and not easy to translate into better molecular structures”, and hence have little impact in drug discovery.

This discussion about the limits of QSAR shows how contradictions occur each time a numerical description of a molecular structure is related to biological activities. These contradictions are logical events and have to be framed in such a context by a logical framework placing them at the core of the user/assistant interaction in order to enable their understanding and to control the proof process of conjectures generated by learning from examples. Section III recalls some definitions about Aristotle’s square, the building brick of this logical framework. This article presents how this classical logical structure is extended to frame all the facets of scientific discovery illustrated by the ADME-T problem.

### III. ARISTOTLE’S SQUARE OF OPPOSITIONS

In this section, we recall what the classical square of oppositions is, and we provide a common sense interpretation to introduce the modalities which take sense in the context of theory formation and scientific discovery.

#### A. Classical square of oppositions

The doctrine of the square of opposition originated with Aristotle in the fourth century B.C. and has occurred in logic texts ever since. It connects various quantified propositions and their negations by introducing various notions of oppositions: contradiction, contrariety and sub-contrariety (Figure 1).

*Definition 1 (Contradiction):* Contradiction for two terms is defined as the impossibility for them to be both true or both false at the same time.

*Definition 2 (Contrariety):* Contrariety is defined as the impossibility for two terms to be both true, but the possibility to be both false.

*Definition 3 (Sub-contrariety):* Sub-contrariety is defined as the impossibility for two terms to be both false, but the possibility for them to be both true.

According to these definitions, opposition is based on various degrees of truth difference. A last useful notion is sub-alternation between two terms, also better known as implication:

*Definition 4 (Sub-alternation / Implication):* Sub-alternation is defined as the impossibility of having the first term true without having also the second true.

The square of oppositions is represented by the following geometrical relations (Figure 1).

The column with *A* and *I* corresponds to affirmative propositions, while the column with *E* and *O* corresponds to negative propositions. The line with *A* and *E* corresponds to universal propositions, while the line with *I* and *O* corresponds to existential (also called particular) propositions. Several extensions have been proposed in order to palliate the logical drawbacks and develop the inference capabilities of the traditional Aristotelian square. Various modal decorations on the vertices can be found in [7].

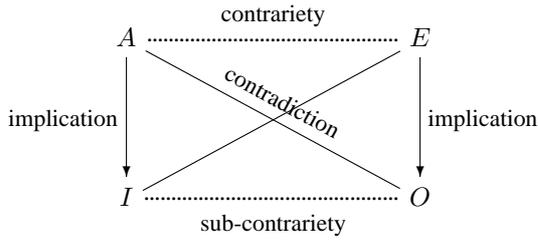


Figure 1. Aristotle's square of oppositions

To express our formalism, we need to define, starting from this square, a closed set of modalities to distinguish the logical status of the statements occurring during the process of theory formation.

### B. An interpretation of Aristotle's square

It is very important to understand that the modalities are the key of the communication between the scientist and his computer assistant: they are used by the assistant to reason logically, and therefore must be interpretable easily by the human to enable a control and a supervision of the theory formation. Several interpretations may be possible, and the ones presented here were chosen because their definitions as found in the dictionary are related in the same way as the modalities they represent.

*Definition 5 (Proof:  $\Box\alpha$ ):* A *proof* is a process which enables the verification of a computation's exactitude, or of the pertinence of problem's solution. To prove  $\alpha$  is to establish with reasoning the truth of  $\alpha$ , and  $\Box\alpha$  denotes the fact that  $\alpha$  has been demonstrated as being necessarily true (or necessary).

*Definition 6 (Refutation:  $\Box\neg\alpha$ ):* A *refutation* is a process which enables to demonstrate the falsity of an affirmation by contrary proofs.  $\Box\neg\alpha$  denotes the fact that  $\alpha$  has been demonstrated as being impossible, or necessarily false.

*Definition 7 (Contingent:  $\neg\Box\alpha$ ):* Contingent denotes the fact that  $\alpha$  has not been proven.  $\alpha$  may then occur, without any explanation or "by chance".

*Definition 8 (Possible:  $\neg\Box\neg\alpha$ ):* Possible denotes the fact that the formula has not been refuted.

If we gave a probabilist interpretation of these four modalities, we would say that  $\alpha$  has a high risk of occurring, whereas  $\neg\Box\alpha$  will most likely not occur. Of course,  $\Box\alpha$  has 100% chances of happening, and  $\Box\neg\alpha$  has none.

Figure 2 represents Aristotle's square decorated with these modalities.

Aristotle's square is the building block of our logical structure. This formalism closes the set of modalities used to type statements treated during logical reasoning, instead of closing the world on which one has to reason. In the following sections, we explain why at least three of these squares are needed to formalize the formation of theories in the context of incomplete and inconsistent knowledge (section IV), and five of them to take into

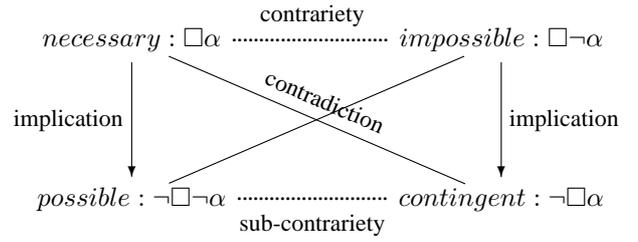


Figure 2. A modal interpretation of Aristotle's square of oppositions

account the scientific aspect of the discovery process (section V).

## IV. THE CUBE OF OPPOSITIONS

The process which supplies evidence for the validity, or for the invalidity, of certain inferences and conversions (of a proposition into its negative) is based on the square presented on Figure 2, which opposes Proof ( $\Box$ ) and its negation ( $\neg\Box$ ). Therefore the square of opposition appears as a geometrization of the inference process. But as we mentioned in the introduction, each logic defines its own negation, and this square actually corresponds to the definition of proof in classical logic, which could be sufficient to help solving problems in a closed and perfectly described world. But to deal with scientific discovery, learning, and theory formation, we need to take into account the inconsistency and the incompleteness of the knowledge used to reason, which justifies the use of paraconsistent logic (in a paraconsistent logic there can exist a proposition which is true and the negation of which is true, without implying the triviality of the theory, i.e. the truth of any proposition) and paracomplete logic (in a paracomplete logic there can exist a proposition which is false and the negation of which is false, without implying triviality of the theory, i.e. the truth of any proposition).

Extending the classical square by introducing conjunctions and disjunctions of the terms, and working on the geometrical aspects of the so formed hexagons as well as on its various modal decorations, [7] introduces new modalities with two other hexagons corresponding to a paraconsistent and to a paracomplete definition of negation (Figure 3).

Each of these hexagons shows how the bottom modality is opposed to the top modality. Looking at the three hexagons, we have thus the most general sub-alternation (or implication) relation between these various negative terms:  $\Box\neg\alpha \longrightarrow \neg\alpha \longrightarrow \neg\Box\alpha$  [8]. This is no surprise, since these terms are known as expressing various kinds of negation in classical and modal proposition logics with the corresponding weakening relations: [9], [10] show that  $\Box\neg$  is an intuitionistic paracomplete negation, and [11] shows that  $\neg\Box$  is a paraconsistent negation.

### A. Modalities attached to a theory formation

We derive from these negations the formal definitions of a Postulate ( $\alpha \wedge \neg\Box\alpha$ ) and Conjecture ( $\neg\alpha \wedge \neg\Box\neg\alpha$ )

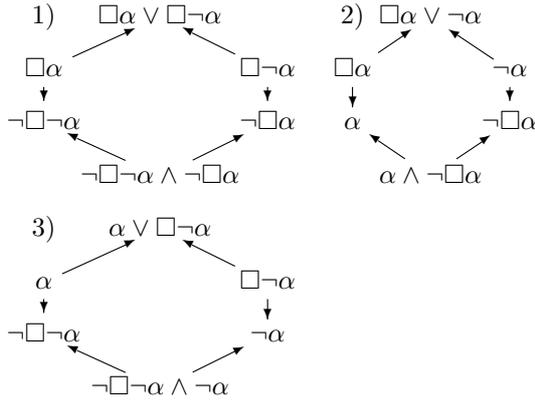


Figure 3. 1) “classical” negation, 2) “paraconsistent” negation, and “paracomplete” negation

(Figure 4), introduced to deal respectively with the inconsistency and the incompleteness of knowledge during the formation of a theory. An ongoing theory is then determined by a set of postulates and conjectures that should finally be instituted as being its principles and its theorems to be demonstrated.

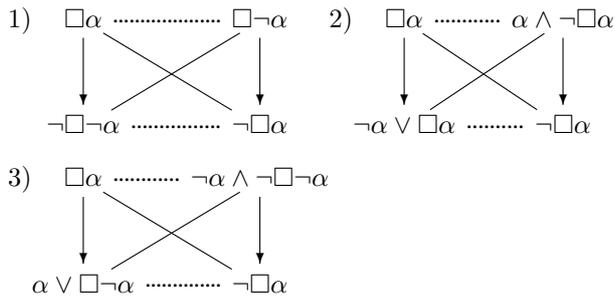


Figure 4. 1) classical definition of a refutation (as on Figure 2), 2) paraconsistent definition of a conjecture, and 3) paracomplete definition of a postulate

**Definition 9 (Observed:  $\alpha$ ):**  $\alpha$  is often interpreted as “true” (and  $\neg\alpha$  as “false”), but here, the interpretation of  $\alpha$  as “is observed” (and  $\neg\alpha$  as “not observed”) is preferred. This notion of observation is refined in section V in which observations are related to experimentation in a context of modeling: there is a glimpse of temporality under this notion, which should be also refined by the process of prediction and factualisation. In the cubic structure presented in the current section, the notion of observation “forgets”, in a certain sense, all the meanings required to design the modeling actions of an agent.

We introduce the use of postulates to fix some limits to what is arguable or not, and conjectures to restrain the objectives, to fix some limits to what is provable or not (for example, one could state that the conjecture  $P = NP$  is not to be proven)

**Definition 10 (Postulate:  $\alpha \wedge \neg\Box\alpha$ ):** A postulate is a primary principle, indemonstrable or undemonstrated. Formally, a postulate is a statement supposed on the basis of incomplete observation;

( $\neg\alpha \vee \Box\alpha$ : the formula is not a postulate).

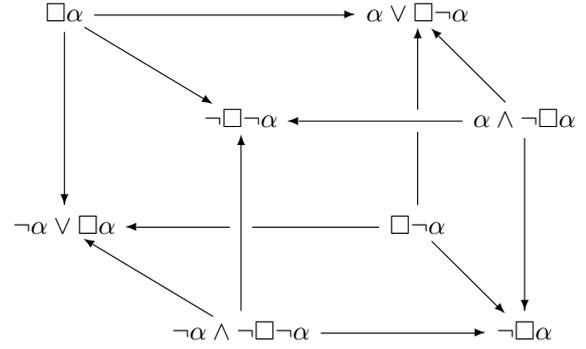


Figure 5. The cube of oppositions

**Definition 11 (Conjecture:  $\neg\alpha \wedge \neg\Box\neg\alpha$ ):** A conjecture is a simple supposition founded on apparent possibility or probabilities, a hypothesis which has not received any confirmation. Formally, a conjecture is a statement which existence is assumed without being observed, since it is not refuted. For instance, *tomorrow is a new day* is a conjecture.

( $\alpha \vee \Box\neg\alpha$ : the formula is not a conjecture)

**Definition 12 (Theory):** A theory includes a set of postulates and a set of conjectures: an ongoing theory is not supposed to be complete and consistent. The use of paraconsistent and paracomplete logics to formulate theories simply comes from the fact that postulates and conjectures are expressing contradictions when using respectively a paraconsistent and a paracomplete negation.

As [12] and [13], we combine the three squares presented on Figure 4 to form a higher order cubic structure, the *Cube of Oppositions* (Figure 5), relating all the modalities presented so far.

The squares are not visible on the faces, but on the diagonals of the cube. Indeed, since the relations of contradiction are visible on the diagonals of Aristotle’s square, it is natural that the different squares form the diagonal planes of the cube of oppositions.

The latter is built from two distinct tetrahedrons. The one of contrariety, which vertices are those from which the implication arrows start, opposes the proof  $\Box$  to the modalities that can imply its contrary  $\Box\neg$ , and the one of sub-contrariety which vertices are those to which the implication arrows lead. The latter opposes the contradiction of a proof  $\neg\Box$  to the modalities that can be derived from a proof. Any vertex of the cube is then contradictory to the furthest lying opposite vertex (easily obtained by central symmetry).

A proof is the result of a dynamic process of constant revision: a new proof is interesting when it proves some conjectures or eliminate surnumerous postulates, and is reciprocally suspected when it proves some conjecture that are reputed unsolvable or false. Logicians appreciate that solvers reason with consistent and complete theories, this is why theories which take into account inconsistency and incompleteness model them by believes, intentions, and defaults. However, during the interactive phases of learning, it is illusory to try being consistent and complete

by considering that errors are the defaults of some known consistent and complete theories, which would lead to making a theory of the whole. A discovery is triggered by the resolution of a contradiction. For instance, it is contradictory to consider a statement as a proof and as a postulate, since proof is related to postulate by a contrariety relation. So when a postulate is proven then either it has to be removed, either the proof is false. This shows how this set of modalities is closed by negation, and why paracomplete and paraconsistent contradictions are needed to reason with incomplete and inconsistent knowledge.

By chance, this cube of oppositions is not an exotic logical structure, but simply a fragment of a classical logic which is used to raise an alert when a theory needs to be revised, i.e. when a postulate or a conjecture is proven. How can a postulate or a conjecture be empirically proven? This question shows the limits of this formalism, as illustrated in section IV-B, and leads to the hypercubic structure presented in section V.

### B. $E+N$ , a game of scientific discovery

The application chosen to illustrate this first part of the formalism is  $E+N$ , a toy game inspired from Abott's mathematical game "Eleusis" [14], and an interaction protocol "Nobel" designed by cognitive scientists to study collective behaviors in a controlled and parametrized environment [15]. As described in [2],  $E+N$  game was designed to simulate the discovery phenomenon and the formation of theories by implementing an Angluin like machine learning interaction protocol [16]. It is currently experimented in with PhD students, to gather enough data to enable the comparison with their results with those of computer assisted players, but also with children to validate the didactic impact of the game on their learning dialectics.

$E+N$  is a card game in which players have to discover hidden rules determining the valid card sequences that can be formed during the game. Players have access to personal experimentation spaces in which they test sequences to observe their validity for the selected hidden rule. They also have access to a public environment in which they can publish their own theory explaining a hidden rule, read the ones submitted by other players, and possibly refute them when they find a sequence which is irrelevant with what was published. The game ends when the flow between published and refuted theories stops, and the winner is the player with the higher score. A player scores by publishing and refuting theories, and loses points when his own theories are refuted.

The choice to use cards was made to make this game accessible to the largest backgrounds as possible (including children), but this simple game is actually a constructive and collective process of theory construction during which players have to observe the sequences at their disposal, and formulate postulates and conjectures to build a publishable theory. The social game confronting players who do not have tested the same sequences

create cases of contingency: when publishing a theory without testing sequences to prove or refute possible and contingent statements, there is a risk that another player has tested them. We insist on the fact that this game has been simplified for experimentation reasons: the length of the sequences is fixed to two cards, and the description of the cards is also fixed and known by every player, we will discuss in section IV-B.2 the consequences of these simplifications, after illustrating with the game's concepts the definitions given so far.

#### 1) Illustration of Definitions:

- Observed ( $\alpha$ ): By creating a new card sequence, one can observe if it is accepted or not. On the contrary,  $\neg\alpha$  denotes that a sequence has not been played. For instance, the sequence  $S=[\text{King of hearts}][\text{Ace of spades}]$  is either played, or not.
- Proof ( $\Box\alpha$ ): When a sequence is valid, then the statement describing it is proven. Considering that the way to describe cards is fixed and known, and supposing that the card sequence  $S$  given above is valid (accepted by the hidden rule), then the following statement  $St$ , is proven: *A card which form is "hearts", which color is "red", which strength is "King", and which is a court card, can be followed by a card which form is "spades", which color is "black", which strength is "Ace", and which is not a court card.*
- Refutation ( $\Box\neg\alpha$ ): Supposing that the card sequence  $S$  given above is not valid (rejected by the hidden rule), then the  $St$  is refuted. If the statement in question, or part of it, has been published, then the player who observe the counter example can refute it and score points. For instance, a theory as "A red card is followed by a black card" can be refuted if  $S$  is not valid.
- Possible ( $\neg\Box\neg\alpha$ ): Possible denotes the fact that a sequence has not been played, and therefore a statement describing it is not refuted or contradicted.
- Contingent ( $\neg\Box\alpha$ ): Contingent denotes the fact that a sequence has not been played, so a statement describing it is not proven. Surely, there is a temporal notion underlying these two last modalities, since what is possible or contingent has not been played *yet*. Contingency is strongly linked to the notion of action, and in this game, every card sequence is contingent, as the result of the player's choice. It is not as if cards were played continuously, following a specific rule, as Earth turning around the sun in a bit less than 365 days, which is a phenomenon strictly observed, and which can not be influence by the observer. Moreover a sequence, until it is played, can be at the same time possible and contingent since these modalities are linked by a contrariety relation.
- Postulate ( $\alpha \wedge \neg\Box\alpha$ ): A postulate can be seen as a statement describing only partially an observed and contingent card sequence. Indeed, the description may not be accurate or really relevant. Since the language to describe cards is given in this simplified

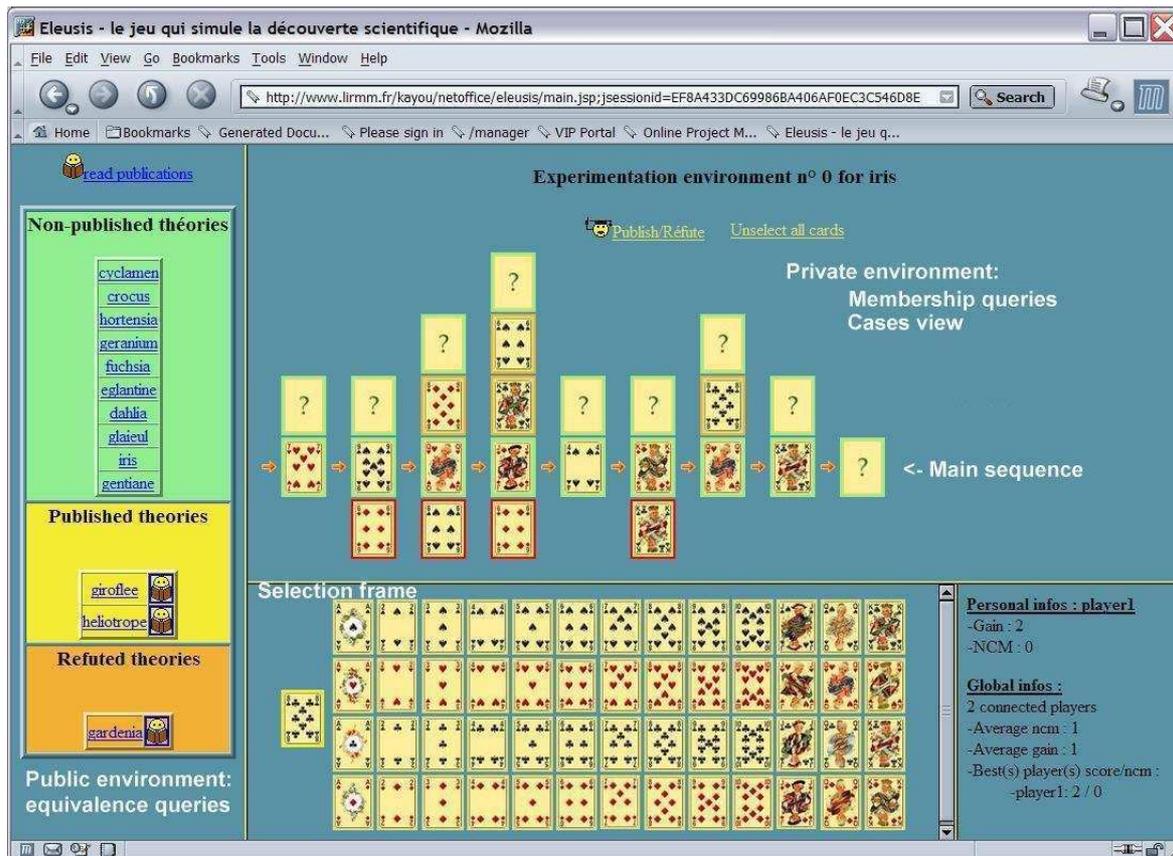


Figure 6. Eleusis + Nobel Game display

version of the game, there can be no postulates, these statements are already proven.

- Conjecture ( $\neg\alpha \wedge \neg\Box\neg\alpha$ ): A conjecture is a set of logical rules that produce predictions (statements) which are not observed and possible. For instance, if there is at least one sequence of a red card followed by a black card which has not been observed as valid, then the statement “a black card will be accepted after a red one” is a conjecture, but can be used to predict, associated to a risk of producing an error.

The dynamic of the interactive process is given by the “Nobel” part of the game during which each player, motivated by the acquisition of credits, publishes his own theories: each player has his own bias, and individual theories are improved by being confronted to others.

2) *Discussion*: As it was emphasized in the presentation of this game, some simplifications were made to make this game accessible to the children who experimented it at first: hidden rules are consistent and complete on the set of sequences of two cards, and the cards description is also fixed and known by each players. As we mentioned in the illustration of postulates (section IV-B.1), the first consequence of these simplifications is to eliminate the use of postulates. A harder game could make players doubt about the most pertinent vocabulary to describe the properties of cards involved in the hidden rule, as well as about the relations between cards (hidden rule concerning  $n$  consecutive cards, or a card could be

related to the one placed in the  $n^{th}$  position after it). In this harder game using postulates would make sense, since the observation of a sequence would only proof its existence. Moreover, in the context of scientific discovery, a publication not only contains postulates and conjectures, but also a model together with a device for experimental validation and facts confirming or invalidating the model’s predictions, which are not taken into account in the Cube. The state of the art and the credibility of an author (credit) are not dealt with neither. However, some of these are already present in E+N and visible on Figure 6: the state of the art, for instance, is represented by a journal of publications, listing every theory published, as well as the eventual refutations (left hand side). On the right hand side is shown information as personal credits (score), the credits of the best player, the average number of cards played before publishing ... The main frame shows the private environment in which each player can freely play cards to form sequences, and this whole experimentation environment could be published as a device for empirical validation or refutation.

Section V fills the gaps mentioned earlier, and presents how the cube is extended into a hypercubic structure by absorbing two other squares of oppositions defining the relations between modeling and proof or refutation on the one hand, facts and predictions one the other.

## V. THE HYPERCUBE OF OPPOSITIONS

In this section, we suppose that each scientist publishes his own model and the experimental protocol to validate it, together with the postulates and conjectures defining the theory as mentioned in section IV. A scientist also produces experimentable facts, as well as possible and contingent statements making factual some predictions coming from others scientists model. Section IV presents a set of modalities closed by negation and used to describe the formation of a theory, in a mathematical way, in a purely theoretic manner. But as Popper [17], we believe that science is pragmatic and must empirically rely on experimentation and be validated by the confrontation of a theory to reality. Every formal element has to be experimentally refutable by a scientific society, and proof and refutation must form the social accreditation process of a formalism.

In other words, the role of the computer assistant is to build, starting from facts, a model which is used to simulate a phenomenon and predict the behavior of a system, in an almost scientific manner. The modalities representing these notions are introduced in section V-A, and an illustration in drug design is provided in section V-B. Section VI then present informally how agents using such a logical framework can interact to agree on a consensual vocabulary to describe a phenomenon.

### A. Modalities in empirical Theory formation

**Definition 13 (Model:  $\sim$ ):** A modeling is a method of study and measurement consisting in replacing a studied system by a model supposed to have an analogous behavior.  $\sim$  is the universal modality associated to a formal statement representing a model.

**Definition 14 (Experimentation:  $\neg\sim$ ):** An experimentation is an effective test realized to study a phenomenon.  $\neg\sim$  is the existential modality associated to a statement expressing an experimentation.

**Definition 15 (Prediction:  $\square \vee \sim \vee \square\neg$ ):** A prediction is the result of a computation, or a principle.  $\square \vee \sim \vee \square\neg$  is the existential modality associated to a statement implied by a proof or a refutation or a model.

**Definition 16 (Fact:  $\neg\square \wedge \neg\sim \wedge \neg\square\neg$ ):** A Fact is the concrete result of an action, and has an unquestionable empirical evidence for scientists.  $\neg\square \wedge \neg\sim \wedge \neg\square\neg$  is the universal modality attached to a statement from which may be implied an experimental, possible and contingent result.

The two opposition squares relating these new modalities are visible on Figure 7 (only modalities are shown to simplify the diagrams). They can be linked to the previous cube to build a hypercubic construction of higher-order geometrical figures of oppositions as suggested by [12] (Figure 8 shows only a part of this hypercube for clarity reasons).

Both experimentation and modeling produce results, and it is the confrontation between them that puts forward an eventual contradiction between a phenomenon and the model used to simulate it, which leads to put into



Figure 7. the squares defining simulation and experimentation towards proof and refutation

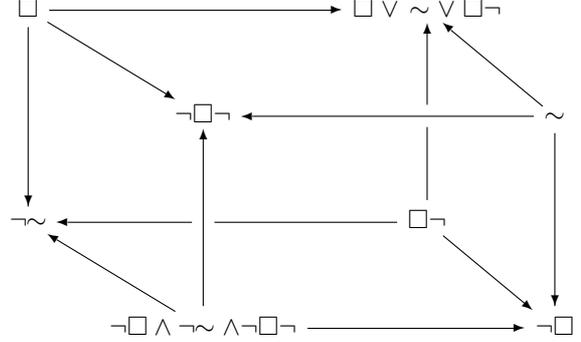


Figure 8. One facet of the hypercube of oppositions

question proofs, conjectures, and postulates to localize the theoretical error.

More generally, each face opposes two universal modalities by confronting the two particular modalities which can be derived from both of them. For instance, different models create different points of view, and modify postulates and conjectures. A first result of this methodology is to reveal the facets of different learning techniques. On the front face, the adequacy of “fact” and “model” is related to the production of possible and contingent statements which are linked by a subcontrariety relation. The corresponding learning methods are version space or Galois lattice techniques. The left face as the bottom face links respectively the “fact” to the “proof” or to the “refutation” of the model. The three other faces are related to the analysis of the prediction. The back face is used to compare a prediction, i.e. a “theoretical result” given by a “simulation”, to an “experimentation” which concretizes some “experimental results” (the fact that an object falls when it is released on earth is observable by experimenting it on a particular object in particular conditions).

In the following section, we illustrate the different facets of this supervision strategy of a learning process on an industrial application in Drug Discovery. The form of the “hidden rules” is not known anymore, but several models can exist.

### B. Application to the prediction of Absorption

In this section, we present an application of the hypercube in scientific discovery, on the absorption problem. In the example E+N presented in section IV-B, the objects were described using a fixed and known ontology and the hidden rules an error free answer to each query concerning a card sequence.

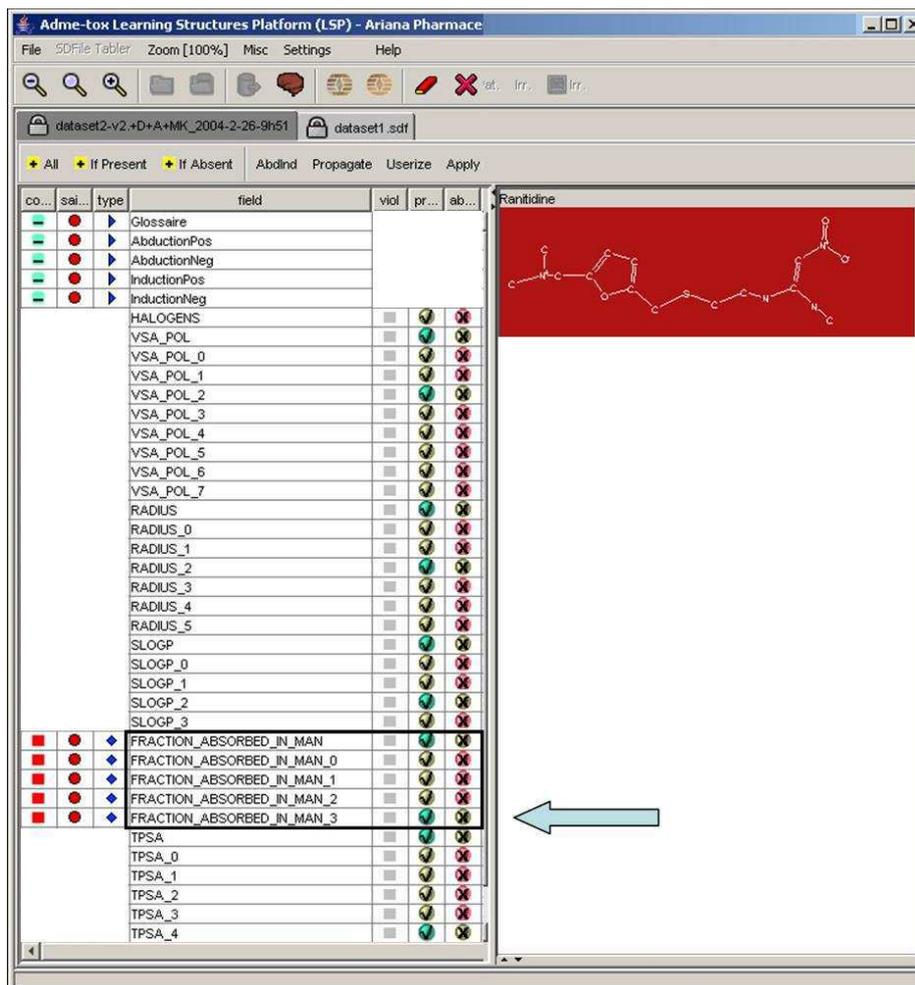


Figure 9. KEM<sup>TM</sup>

In this new problem a scientist is using a learning machine to learn a model from facts. This model is formulated as a constraint network, and each constraint concern properties of the molecular graph's description. As the model is learned from examples, since predictions may be done in a context of overfitting and underfitting, contradictions are the events which allow the supervision, by the expert, of the learning assistant modeled as a hypercube.

This real application of learning in scientific discovery is from collaboration with Ariana Pharmaceuticals in Drug design [18]. KEM<sup>TM</sup> can suggest specific molecular modifications to achieve multiple objectives, after analyzing a multi-parametric database.

In this example we focus on the prediction of absorption, a key issue in drug design since this is one of the important and early causes of failure in the drug discovery process. Indeed molecules need to be absorbed before they can perform any desired activity. Absorption is a complex process involving both passive (diffusion) and active (through transporter proteins) across cellular membranes. For passive transport, molecules need to be soluble (hydrophilic) in water and at the same time they need to be greasy (hydrophobic) to penetrate cellular

membranes that are formed of lipids. This contradicting requirement is modulated by active transport, where molecules need to be recognized (i.e. complementarity of shape and charge) by another molecule (transporter) that helps them through membranes. Although no one can for sure predict the absorption of a new molecule, a number of empirical rules are known. This is an interesting context for applying our IA since our key requirement is to capture knowledge from the experimental data and then evolve and improve this model in a consistent manner.

To illustrate our approach we focus on a set of 169 molecules for which the absorption in man has been experimentally evaluated (4 classes. 0 not absorbed, 3 highly absorbed) [19]. These molecules are described using a set of physico-chemical properties such as molecular radius, different calculated measures of their total polar surface accessible to water (*TPSA* and *VSA POL*), their hydrophobicity (*SLOGP*), presence of halogens etc.

To learn, KEM acts according the facet of the hypercube of oppositions visible on Figure 8: 1)(left face) A decision tree is used to find a good segmentation of the numerical descriptors. 2)(front face) A Galois lattice method works on these binary descriptors in order to construct a lattice of regularities. 3)(upper face) sup-

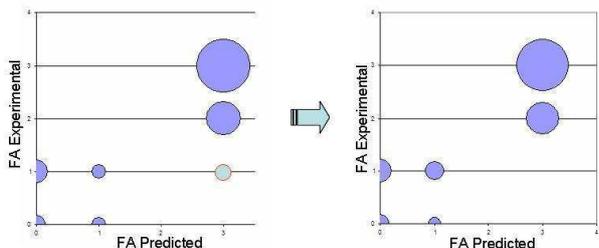


Figure 10. Predictions *A* and *B*

irreducible nodes are translated into logical constraints for prediction. 4)(back face) the prediction is confronted to the experimentation. 5)(right face) the study of the experimental error is done by a refutation of the simulation.

Initially, the system learns from the data set a set of rules linking the structure of the molecule to the absorption. The quality of the prediction is tested in a subsequent stage on a novel set of molecules. The results are shown on prediction *A* in Figure 10. Ideally the predictions should be on the diagonal. An error of one class is tolerated. However, it is clear that for one molecule, the error is larger (i.e. experimental : class 1 vs predicted: class 3). This confrontation between a theoretical result and an experimental result puts forward a contradiction in the model.

Figure 9 shows this contradiction: the molecule (Ranitidine) has been predicted with *fraction absorbed in man 3* i.e. highly absorbed. However, if the user makes a postulate and forces *fraction absorbed in man 3* to be false, the system localizes the error that induced the contradiction by showing that the postulate contradicts the conjectural learned rule  $VSA\ pol\ 2 \rightarrow fraction\ absorbed\ in\ man\ 3$ . At this stage the user realizes that indeed this conjecture was true for the learning set, however this is not generally true and it can be eliminated. The user then goes back to simulating once more the test and results are shown in Figure 10, prediction *B*. As expected, the results have been improved. The important point is that the improvement has been done in a controlled way under the user’s supervision, and this was only possible because the user and  $KEM^{TM}$  shared a common vocabulary to type statements.

In scientific discovery, there are in general no Oracles who can say a priori whether a prediction is correct or not. Experimentalists formulate a conjecture that is consistent with existing empirical data and then set about to test it. We believe that the key for a computational system is to adhere to the same process i.e. build up an explanation / reasons for predicting an outcome. If the system is able to provide enough arguments, the user will “trust” it and try the experience. This implies that the arguments are annotated with modalities which are meaningful both for the user and the machine.

## VI. THE INSTITUTION AGENT SOCIAL GAME

This section is the final step of the formalism and links the classical formalisms used to express the individual

and collective behaviors of agents. Our contribution is to describe the dynamic process of conventional formation of theories by agents (human as well as artificial agents) in respect with social norms regulating a community, in order to describe a community of agents interacting with each other to understand their environment and organize themselves to fix common objectives and chose the best actions to achieve them. The objective is then to give the intuitive idea about mathematical developments, which are our ongoing research, and to show that this is a constructive approach to build agents comparable to classical BDI agents, which use Beliefs, Desires, and Intentions to reason and decide.

The only constraint fixed on agents is to reason using a hypercube (section V) and a normative system NS, which is a logical language using the cube modalities to communicate.

*Definition 17 (Normative System):* We call a Normative System (NS) the couple  $(L, Cube)$  formed by:

- *L*: a language formed by a hierarchy of concepts and the relations between them,
- *Cube*: a cube of oppositions.

*Definition 18 (IA):* An *IA* is an agent using an *Hypercube* framework to reason and judge statements. As we emphasized, the *Hypercube* is a formalization of dialectic as it is practiced by humans, and an *IA* can then be either an artificial agent, or a human agent. The interaction protocol is compliant with both of them since the definitions of the modalities are shared by both of them.

Scientific discovery is a collective process, and needs interaction between researchers to exchange their points of view and judgments. This is how *IAs* interact (Figure 12): by exchanging judgments about statements, more exactly, by asking another *IA* if it agrees with a particular judgment: “this statement is a conjecture, is it not?”, to which the answer is “yes” or “no, it is a result”, or if a theory seems relevant or not. This type of interaction is well described in Machine Learning theories and these queries are known respectively as *Membership* and *Equivalence* queries. [16], [20] shows that the use of at least these two types of queries is required for an effective learning. Exchanging judgments creates the negation in the common frame of reference (language), and the revision of the normative system associated with one *IA* or the other. Two judgments are especially important: *judging one’s conjecture as being a paradox*, and *judging one’s postulate as already being a result*.  $KEM^{TM}$ , presented in section V-B, illustrates this control by a scientist over the artificial *IA* assisting him.

We suppose that each *IA* can be represented by a particular normative system resulting from its own experience and adaptation during an interaction with other *IAs*, and we assume that the logic used during a decision process is the same for every *IA*. We then focus on adaptation and interaction of *IAs* sharing a vocabulary and trying to build a common language or frame of reference with this vocabulary to describe a phenomenon

and provide consensual definitions, expressed logically or by constraint networks.

Three logical properties, presented in the following, are needed to qualify this interaction protocol and to add a logical control to the adaptation process:

- deontic: an *IA* must be able to attribute credits to another *IA*, to interact, and to teach another *IA*,
- defeasible: Lower *IAs* must be able to adapt their behavior to the norms imposed by the higher ones,
- autoepistemic: an *IA* can be seen as composed by at least two interacting *IAs* and can therefore learn its own hierarchy of norms and auto-adapt.

section VI-A present how a deontic and defeasible logic can be used to frame the collective dynamic during which *IAs* build a common *NS*, and section VI-B suggest how an *IA* can be itself composed by several *IA* to form an autoepistemic agent performing an inner dialog while reasoning.

### A. Collective dynamic

1) *Deontic logic*: Often used in multi-agent systems to constrain an agent's behavior, annotable deontic logic uses modalities expressing obligation, interdiction, advice, and warning. According to Frege's definition, these statements express a judgment, i.e. the recognition of the type of truth of the statement [21]. Imputations (gains or losses, risk estimation) are used to estimate the risk incurred in a given situation to decide what action to take or what behavior to adopt. A modality and an imputation have to be used to express statements of the following form: "The obligation to respect the speed limit is attached to a imputation of  $x$ ". A credit value can also be associated to *IAs*, ordering them hierarchically, to define which one is the most qualified to rule in a given context, for example by defining a social organization as a government with a parliament, a senate, . . . .

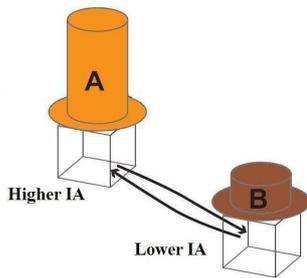


Figure 11. IA's credit

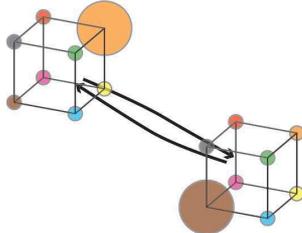


Figure 12. Exchanging judgments

2) *Defeasible logic*: It is possible to link two *IAs* by respecting a defeasible logic to take into account a hierarchy of Institution Agents. The resulting hierarchy of *IAs* has to be brought together with the transitivity axiom, that stands as follows: "What is necessary in an upper *NS* of is also necessary in a lower *NS*". In other words, no one should be unaware of the law, no one should go against a superior law. [22] gives a concrete usage of defeasible logic, that allows us to order rules and to supervise an *IA*, for example with another higher *IA*, as illustrated on Figure 13.

- Every Obligation of a lower *IA* belongs to the superior *IA*'s advice.
- Every Interdiction of a lower *IA* belongs to the superior *IA*'s warnings.

The middle line shows the conditions according to which an *IA* can be supervised by another one. The violation of this constraint ( $O_2 = \Box_2 \subset I_1 = \Box_1 \neg$  or  $I_2 = \Box_2 \neg \subset O_1 = \Box_1$ ) can put forward contradictions between the two *IA*'s normative system. Finding a contradiction, and trying to eliminate it, leads to the initiation of a transaction between the two *IAs*, during which they adapt their theory. When no contradiction remains, a new *IA* can be created, formed by the association of the two precedent *IAs*, and this process ensures the traceability of all the events leading to an *IA*'s creation.

### B. Personal dynamic

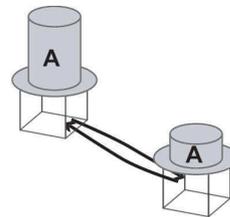


Figure 14. Autoepistemic dialog

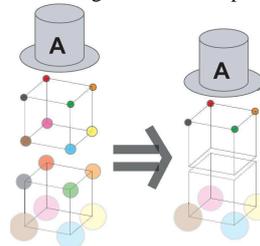


Figure 15. IA formation

1) *Autoepistemic logic*: Aristotle distinguishes en-dophasy as an inner dialog ("to think is to speak with oneself"). This is a constructive manner to build an intelligent agent as the result of an auto-adaptation. The inner *IAs* can be interpreted as managing believes, desires or intentions (BDI), for example. By applying the dialectic and deontic interaction presented in section VI-A.1, an *IA* is able to acquire its own *NS*, which prepares an efficient learning, and even enable self learning from examples.

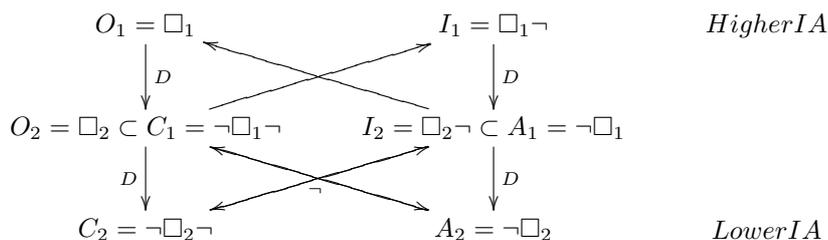


Figure 13. Normative system hierarchy

In this section, we presented how an interaction process and a hierarchical control can be used to define an agent able to adapt its defeasible deontic and autoepistemic Normative System.

## VII. CONCLUSION

We presented a logical framework in order to capture the entire process of scientific discovery as it is performed in scientific community. This framework is based on Aristotle's square, a conventional logical structure, and recent related works in logic. In this context, we focused on a geometrical representation of the notion of logical contradiction.

Our approach is progressive. We first presented with a cubic structure the formation of a theory as driven by the proof of its postulates and conjectures by scientists. Then we presented a hypercubic extension to capture the empirical modeling processes, to make the predictions of a model and the experimentations explicit. We showed how this theory construction can be done by human and by machines taught by humans in a formal and in a practical way. finally we presented in an informal way how the institution agents (agents using this hypercubic structure to build theories), educated by human, can be constrained to respect normative systems and how this approach is only a constructive alternative to BDI agents.

We illustrated on an example coming from Drug Discovery how these annotations of common sense, which are now logically defined, are used by a learning machine and a scientist to interactively build a model which is coherent and complete with observations and experimental results. We prone that this hypercube describes in a universal way a rational agent and enables the supervision of its computing process.

In perspective, to complete our experimentation, we plan on experimenting this constructive approach to teach a humanoid robot and make him gain more autonomy. A mathematical formalism of agents based on this vision is currently developed in category theory.

## REFERENCES

- [1] T. Parsons, "The traditional square of opposition," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Winter 2006. [Online]. Available: <http://plato.stanford.edu/archives/win2006/entries/square/>
- [2] C. Dartnell and J. Sallantin, "Assisting scientific discovery with an adaptive problem solver." in *Discovery Science*, ser. Lecture Notes in Computer Science, A. G. Hoffmann, H. Motoda, and T. Scheffer, Eds., vol. 3735. Springer, 2005, pp. 99–112.
- [3] J. Sallantin, C. Dartnell, and M. Afshar, "A pragmatic logic of scientific discovery." in *Discovery Science*, ser. Lecture Notes in Computer Science, L. Todorovski, N. Lavrac, and K. P. Jantke, Eds., vol. 4265. Springer, 2006, pp. 231–242.
- [4] S. Morley and M. Afshar, "Validation of an empirical rna-ligand scoring function for fast flexible docking using ribdock." *Journal of Computer Aided Molecular Design*, pp. 189–208, 2004.
- [5] M.-G. N., W. K. A. B.K., and K. R.D., "New approach to pharmacophore mapping and qsar analysis using inductive logic programming. application to thermolysin inhibitors and glycogen phosphorylase b inhibitors," *Journal of Medicinal Chemistry*, vol. 45(2), pp. 399–409, 2002.
- [6] B. Faller and F. Wohnsland, "Physicochemical parameters as tools in drug discovery and lead optimisation," in *Pharmacokinetic optimization in drug research*. Testa, Waterbeemd, Folkers and Guy editors, Wiley-VCH, Zurich, 2004, pp. 189–208.
- [7] J.-Y. Béziau, "New light on the square of oppositions and its nameless corner," *Logical Investigations*, vol. 10, pp. 218–233, 2003.
- [8] R. Blanché, *Structures intellectuelles : essai sur l'organisation systématique des concepts*. Vrin, Paris, 1966.
- [9] K. Došen, "Intuitionistic double negation as a necessity operator." *Publications de l'Institut Mathématique*, vol. 35(49), pp. 15–20, 1984.
- [10] K. Dosen, "Negative modal operators in intuitionistic logic," *Publications de l'Institut Mathématique*, vol. 35(49), pp. 3–14, 1984.
- [11] J.-Y. Béziau, "S5 is a paraconsistent logic and so is first-order logic," *Logical Investigations*, vol. 9, pp. 301–309, 2002.
- [12] A. Moretti, "Geometry for modalities? Yes: through  $n$ -opposition theory," in *Aspects of Universal Logic*, J.-Y. Béziau, A. Costa-Leite, and A. Facchini, Eds. Travaux de logique 17, Neuchâtel, 2004, pp. 102–145.
- [13] R. Pellissier, "'Setting'  $n$ -opposition," in *UNLOG05*, (to be published), 2006.
- [14] M. Gardner, "Mathematical games," *Scientific American*, June 1959.
- [15] D. Chavalarias, "La thèse de popper est-elle réfutable?" Memoire de DEA, CREA - CNRS/Ecole Polytechnique, 1997.
- [16] D. Angluin and Krikis, "Learning from different teachers," *Machine Learning*, vol. 51, pp. 137–163, 2003.
- [17] K. R. Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge*. Harper and Row, 1963.
- [18] M. Afshar, A. Lanoue, and J. Sallantin, "New directions: multidimensional optimization in drug discovery," *Comprehensive Medicinal Chemistry 2*, vol. 4, 2006.
- [19] Y. Zhao, M. Abraham, A. Hersey, and C. Luscombe, "Quantitative relationship between rat intestinal absorption and abraham descriptors," *Eur J Med Chem.*, pp. 939–947, 2003.
- [20] D. Angluin, "Queries revisited," *Theoretical Computer Science*, vol. 313, pp. 175–194, 2004.

- [21] C. D. Pozza and C. Garola, "a pragmatic interpretation of intuitionistic propositional logic," *Erkenntnis*, vol. 43, pp. 81–109, 1995.
- [22] K. Nakamatsu, T. Kato, and A. Suzuki, "Basic ideas of defeasible deontic traffic signal control based on a paraconsistent logic program evalpsn," *Advances in Intelligent Systems and Robotics*, 2003.

**Mohammad Afshar** is the CEO and co-founder of Ariana Pharmaceuticals. He was one of the founding members of the senior management group at RiboTargets, Cambridge, UK, where, as the Director of IT and Head of Drug Design, he set up and managed RiboTargets' structure based discovery platform RiboDock<sup>TM</sup> - rDock. The patented technology developed within his team (RiboDock<sup>TM</sup> - rDock) allowed the identification and validation of novel therapeutic molecules, becoming the central focus of RiboTargets. He was actively involved in the companys business development.

**Christopher Dartnell** is currently a PhD candidate at UM2, The Science University of Montpellier, France. He received his MS and degree in computer science in 2003 and is currently working in collaboration with the LIRMM (Laboratory of Computer Sciences, Robotics, and Mathematics of Montpellier) and Euriware, specialist in industrial IT and information systems for the energy sector (member of the Areva group). His research interests include machine learning, assistance for the supervision of complex systems, logical reasoning and robot control.

**Dominique Luzeaux** graduated from École Polytechnique (1987), École Nationale Supérieure des Techniques Avancées (1989), PhD from University Paris XI (1991). He held the position of an invited postdoctoral fellow at the University of Berkeley in 1991-1992. In 2001 he completed his professor thesis at the University Paris XI. He is currently director of the Technical Center for Information Systems within the Ministry of Defense. He also teaches robotics and theoretical computer science at the École Nationale Supérieure des Techniques Avancées and at graduate level.

**Jean Sallantin** is a Director of Research at the LIRMM Laboratoire d'Informatique, de Robotique et de Micro-électronique de Montpellier of the CNRS in Montpellier, France. Phd from University Paris VI in Theoretical Physic (1972) and professor thesis in Computer Science (1979). He currently heads the Rationality and machine learning team that develops and studies the applications of machine learning techniques in a scientific discovery environment. One of the pioneers of bioinformatics in France, Jean Sallantin established the bioinformatics laboratory at the Curie Institute Paris (1983-85) and was the director of the GSDIARL research consortium (1985-89).

**Yannick S. Tognetti** is a research scientist at ariana pharmaceuticals, specialist in Artificial Intelligence. He obtained a PhD in Computer Science and Artificial Intelligence working on the modeling of chemical information systems using graph's theory and algorithmics, in 2002 at LIRMM Montpellier.

# Logical extensions of Aristotle's square

Dominique Luzeaux, Jean Sallantin and Christopher Dartnell

**Abstract.** We start from the geometrical-logical extension of Aristotle's square in [Bla66], [Pel06] and [Mor04], and study them from both syntactic and semantic points of view. Recall that Aristotle's square under its modal form has the following four vertices:  $A$  is  $\Box\alpha$ ,  $E$  is  $\Box\neg\alpha$ ,  $I$  is  $\neg\Box\neg\alpha$  and  $O$  is  $\neg\Box\alpha$ , where  $\alpha$  is a logical formula and  $\Box$  is a modality which can be defined axiomatically within a particular logic known as  $S5$  (classical or intuitionistic, depending on whether  $\neg$  is involutive or not) modal logic.

[Béz03] has proposed extensions which can be interpreted respectively within paraconsistent and paracomplete logical frameworks. [Pel06] has shown that these extensions are subfigures of a tetraicosahedron whose vertices are actually obtained by closure of  $\{\alpha, \Box\alpha\}$  by the logical operations  $\{\neg, \wedge, \vee\}$ , under the assumption of classical  $S5$  modal logic. We pursue these researches on the geometrical-logical extensions of Aristotle's square: first we list all modal squares of opposition. We show that if the vertices of that geometrical figure are logical formulae and if the sub-alternation edges are interpreted as logical implication relations, then the underlying logic is none other than classical logic. Then we consider a higher-order extension introduced by [Mor04], and we show that the same tetraicosahedron plays a key role when additional modal operators are introduced. Finally we discuss the relation between the logic underlying these extensions and the resulting geometrical-logical figures.

**Mathematics Subject Classification (2000).** Primary 03B45; Secondary 06F99.

**Keywords.** Aristotle's square, modal logic.

## 1. The traditional square of opposition and its lower-order generalizations

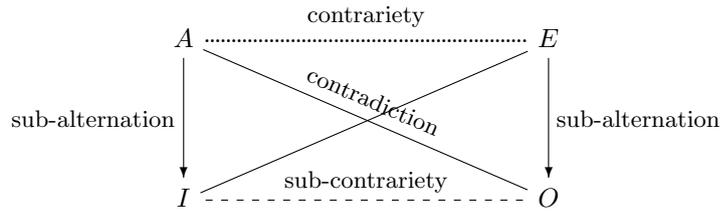
### 1.1. Aristotle's square

The doctrine of the square of opposition originated with Aristotle in the fourth century B.C. and has occurred in logic texts ever since. It relates various quantified propositions and their negation by introducing various notions of oppositions: *contradiction*, *contrariety* and *sub-contrariety*. *Contradiction* for two terms is defined

as the impossibility for them to be both true and both false. *Contrariety* for two terms is the impossibility for them to be both true, but the possibility to be both false. *Sub-contrariety* is the impossibility to be both false, but the possibility to be both true. A last useful notion is *sub-alternation* between two terms, also better known as logical implication: it is defined as the impossibility of having the first term true without having also the second true. It will be denoted by an arrow, which is consistent with the logical interpretation of implication.

We should notice that although contrariety and sub-contrariety are discussed extensively in syllogistic, they are derived logical concepts when contradiction and sub-alternation are logically defined.

The square of oppositions synthesizes these notions within the following geometrical figure:



The column with  $A$  and  $I$  corresponds to affirmative propositions, while the column with  $E$  and  $O$  corresponds to negative propositions. The line with  $A$  and  $E$  corresponds to universal propositions, while the line with  $I$  and  $O$  corresponds to existential (also called particular) propositions. Medieval syllogistic has developed a bunch of rules based on the extensive use of this square, that encode true inferences between premisses and conclusions. However we will not need them, as a standard interpretation of truth with the aforementioned notions will suffice.

This square leads however to logical ambiguities, and in the last century it has been heavily criticized with the development of formal logic. Indeed Aristotle's theory of syllogism and all the Medieval syllogistic corpus developed until the 19th century should not be seen as a formal theory of logic, but as a theory of reasoning (based on the modus ponens and modus tollens rules, which form the basis of propositional calculus), that allows to infer the truth of certain propositions. Furthermore the process which supplies evidence for the validity, or for the invalidity, of certain inferences and conversions (of a proposition into its negative) is based on simple graphics. Therefore the square of opposition appears as a geometrization of the inference process.

Several extensions have been proposed in order to palliate the logical drawbacks and develop the inference capabilities of the traditional Aristotelian square. [Bla66] decorates the vertices with modalities [Gol05], and develops a theory of reasoning based on the square of modalities:  $A$  is  $\Box\alpha$ ,  $E$  is  $\Box\neg\alpha$ ,  $I$  is  $\neg\Box\neg\alpha$  and  $O$  is  $\neg\Box\alpha$ , where  $\alpha$  is a (classical logical) proposition. The modality  $\Box$  is a universal modality, usually interpreted for instance as "necessarily" or "it is always known" or "it is proved" or "it is compulsory", depending on the modal logic used: necessity, epistemic, provability, deontic. The modality  $\neg\Box\neg$  is usually denoted by  $\Diamond$

and is interpreted in the former cases by "possibly", "it is sometimes known", "it is not refuted", "it is allowed";  $\neg$  stands for usual classical negation. In addition, [Bla66] introduces two additional vertices  $I \wedge O$  and  $A \vee E$ , yielding a contrariety triangle  $A, E, I \wedge O$  and a sub-contrariety triangle  $A \vee E, O, I$ . Furthermore there are obvious sub-alternation relations between  $A$  and  $A \vee E$ ,  $E$  and  $A \vee E$ ,  $I \wedge O$  and  $I$ ,  $I \wedge O$  and  $O$ . Thus the square turns into a hexagon, as illustrated on the left part of figure 1.

Working on the geometrical aspects of this hexagon and its various modal decorations, [Béz03] introduces other hexagons: a paraconsistent one (in a paraconsistent logic there exists a proposition which is true and the negation of which is true, without implying triviality of the theory, i.e. the truth of any proposition), and a paracomplete one (in a paracomplete logic there exists a proposition which is false and the negation of which is false, without implying triviality of the theory). Both these hexagons are illustrated in figure 1.

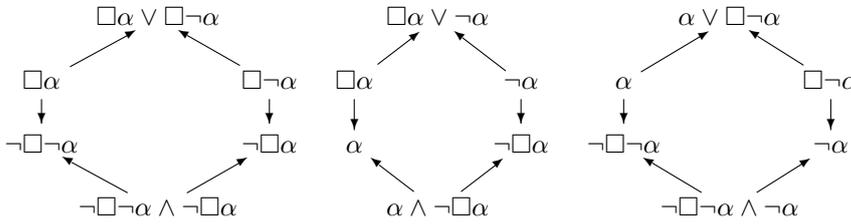


FIGURE 1. The three hexagons from left to right: classical, paraconsistent, paracomplete.

The interesting observation in these hexagons is the underlying square in each, and more specifically the column of negative terms EO. Looking at the three hexagons, we have thus the more general sub-alternation relation between these various negative terms:  $\Box\neg\alpha \rightarrow \neg\alpha \rightarrow \neg\Box\alpha$ . This is no surprise, since these terms are known as expressing various kinds of negation in classical and modal proposition logics with the corresponding weakening relations: [Béz05, Doš84a, Doš84b] show that  $\Box\neg$  is an intuitionistic paracomplete negation, and [Béz02] shows that  $\neg\Box$  is a paraconsistent negation.

It is worth noticing that the paraconsistent hexagon is obtained from the paracomplete hexagon, and conversely, whenever  $\alpha$  is changed into  $\neg\alpha$ .

Questions that arise at that point are: are there other modal decorations on hexagons? are there higher-order geometrical figures beyond squares and hexagons, that can model other possibly richer opposition theories?

## 1.2. $n$ -opposition theory

Answers to the previous questions have been provided by [Mor04] and [Pel06]. The theory of  $n$ -opposition has been developed by A. Moretti [Mor04] in order to generalize the completions of Aristotle's square into the hexagons discovered

by Blanché and Béziau. Each of these hexagons is a 2-bisimplex built from a 2-simplex (a triangle) of contrariety (two by two conjunctions of its vertices are false) and its symmetric by central symmetry, the sub-contrariety triangle (two by two disjunctions of its vertices are true).

In order to generalize the ternary contrariety relation to an  $n$ -ary opposition relation, Moretti has developed the theory of  $n$ -opposition relying on  $n$ -oppositional figures ( $(n - 1)$ -bisimplices), and has created the notion of modal  $n(m)$ -graphs. These graphs are a convenient tool to generate  $n$ -oppositional figures, as shown in [Mor04]. Furthermore all  $n$ -oppositional figures generated by a given modal graph can be listed combinatorially, and the geometrical solid which includes all these  $n$ -oppositional figures is called a  $\beta$ -structure in Moretti's terminology. The formal definitions follow.

**Definition 1.1.** An  $n$ -opposition consists of an  $(n - 1)$ -bisimplex formed by an  $(n - 1)$ -simplex, called contrariety simplex, and by its symmetric by central symmetry, called sub-contrariety simplex, each vertex of the first simplex being connected by sub-alternation to each non-symmetric vertex of the second simplex. The symmetric vertices are contradictory. The contrariety simplex has the property that two by two conjunctions of its vertices are false. Dually, the sub-contrariety simplex has the property that two by two disjunctions of its vertices are true.

In order to decorate (i.e. put the right formulas on each vertex of the bisimplex) modally such simplices, modal  $n(m)$ -graphs are introduced.

**Definition 1.2.** A modal  $n(m)$ -graph is the cartesian product in the category of graphs of an  $(n - 2)$  simplex by the oriented graph  $1 \rightarrow 2 \rightarrow \dots \rightarrow m$ , i.e.  $m$  copies of the simplex, such that for all  $j \leq n - 1$  and  $i \leq m$ :

- the  $j$ -th vertex of the  $i$ -th copy is contradictory to the disjunction of all vertices of the  $m - i + 1$ -th copy but the  $j$ -th;
- there is a sub-alternation relation between the  $j$ -th vertex of the  $i$ -th copy and the  $j$ -th vertex of the  $m - i + 1$ -th copy for  $i \leq m - i + 1$ .

The resulting graph is called modal since all vertices are labelled by modal operators, and the oriented edges correspond to logical implication.

**Definition 1.3.** A  $\beta$ -structure associated to an  $n(m)$ -graph is the geometrical solid consisting of all  $p$ -oppositions exhibited by the  $n(m)$ -graph.

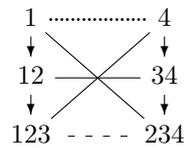
The second reference [Pel06] continues the previous work, by introducing a particular encoding. The modal  $n(m)$ -graphs – seen actually as “directed”  $n(m)$ -graphs, since the actual modal essence of the labels is not exploited – are translated into set theory, by identifying the vertices of the graph with subsets of a given set, consistently with the underlying lattice structure: an implication (resp. upper bound, resp. lower bound) between two vertices turns into an inclusion (resp. union, resp. intersection) of the corresponding subsets. With help of this translation, [Pel06] shows that an  $n(m)$ -graph is translated into the Boolean lattice

corresponding to a set of cardinality  $c_{n,m}$  equal to:  $(n - 1)\frac{m}{2} + 1$  if  $m$  is even, and  $(n - 1)(\frac{m+1}{2})$  if  $m$  is odd.

An immediate corollary [Pel06] is that an  $n(m)$ -graph and a  $3(c_{n,m} - 1)$  graph are translated into the same Boolean lattice corresponding to a set with  $c_{n,m}$  elements. Thus the general problem can be reduced, up to this translation, to the study of all modal  $3(m)$ -graphs.

In particular, as illustrated in [Pel06], the study of the modal  $3(3)$ -graph exhibits 3-bisimplices, which can be gathered into a geometrical figure, a  $\beta$ -structure in Moretti's terminology, known as a tetraicosahedron. Let us recall this, using Pellissier's encoding for sake of simplicity.

Recalling [Pel06], a  $3(3)$ -graph can be translated into a Boolean lattice constructed on a set of cardinality  $c_{3,3} = 4$ . Therefore one starts from the following modal  $3(3)$  graph, with similar graphical representations as before (dots for contrariety, dashes for subcontrariety, lines for contradiction, arrows for sub-alternation):



and obtains a geometrical figure, known as a tetraicosahedron, composed of a cube with a square-based pyramid on each face of the cube. This directed tetraicosahedron, illustrated in figure 2, represents all sub-alternation relations that exist within the Boolean lattice generated by a 4-element set under the assumption of the previous  $3(3)$ -graph.

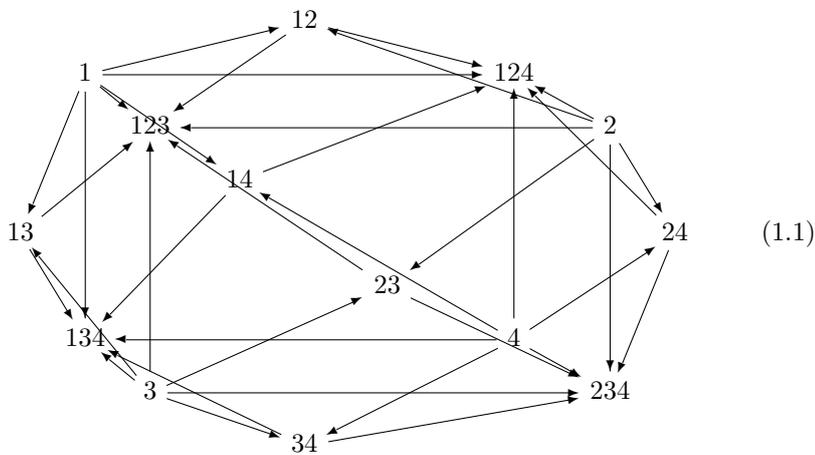


FIGURE 2. The "Boolean" tetraicosahedron.

Actually, for sake of completeness, this figure should have two more vertices, labelled  $\emptyset$  and 1234, with arrows from  $\emptyset$  to all vertices, and from all vertices to 1234; but since their interpretation, as developed in the next sections, is trivial most of the time, we will not consider them systematically. When we do consider them, we use the denomination “completed” tetraicosahedron.

This tetraicosahedron with 14 vertices and 24 faces contains four “weak” and six “strong” hexagons as well as a “strong” cube, that form all possible geometrical figures within 3-opposition theory. The denomination “strong”, introduced in [Pel06], characterizes those figures where the contrariety simplex is built from terms that are not only contrary two by two (their conjunction two by two is false), but their global disjunction is true. The denomination “weak” corresponds to situations where the disjunction of all terms within the contrariety simplex is not necessarily true. Actually one of the originalities of [Pel06] is this distinction, as all previous references had always assumed “strong” constructions. Back to our tetraicosahedron, all geometrical figures are thus (hexagons are described by starting from the left uppermost vertex and turning clockwise):

- weak hexagons  $\{1, 124, 2, 234, 3, 134\}$ ,  $\{1, 123, 2, 234, 4, 134\}$ ,  $\{1, 124, 4, 234, 3, 123\}$ ,  $\{124, 2, 123, 3, 134, 4\}$ ;
- strong hexagons  $\{12, 2, 234, 34, 134, 1\}$ ,  $\{12, 124, 4, 34, 3, 123\}$ ,  $\{1, 124, 24, 234, 3, 13\}$ ,  $\{123, 2, 24, 4, 134, 13\}$ ,  $\{1, 123, 23, 234, 4, 14\}$ ,  $\{124, 2, 23, 3, 134, 14\}$ ;
- a strong cube whose faces are  $\{1, 124, 2, 123\}$  and  $\{134, 4, 234, 3\}$ .

## 2. Logical interpretations within classical $S5$ modal logic

### 2.1. The logical tetraicosahedron and the logical cube

Let us turn now to the logical interpretations of the previous notions. From now on, we will assume some kind of a logical framework is given (for instance classical propositional logic). The modal  $n(m)$ -graphs provide additional logical constraints on this given logic. The vertices of the  $n$ -oppositional figure are then obtained by the closure, under conjunction, disjunction and contradiction (negation), of the algebra generated by the  $n(m)$ -graph, and the oriented edges are the implications induced by the  $n(m)$ -graph between the corresponding vertices.

Such a logical interpretation of the  $n$ -oppositional figures motivated Béziau’s, Moretti’s and Pellisier’s pioneering work, following Blanché’s initial research, and in the rest of this paper we will delve deeper into this connection between logical and geometrical aspects of  $n$ -opposition theory.

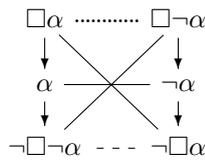
The first logic we will consider is classical propositional modal logic. Recall that the latter is not uniquely defined. However among all the diversity, there are a few commonly used systems, following Lewis’s seminal work in the 20th century.

One starts from classical propositional logic (with the usual connectives  $\wedge$ ,  $\vee$ ,  $\neg$ ,  $\rightarrow$ , and its usual theorems, such as de Morgan’s laws, contraposition and involutivity of negation). In this system there are only two modalities ( $\cdot$  the identity

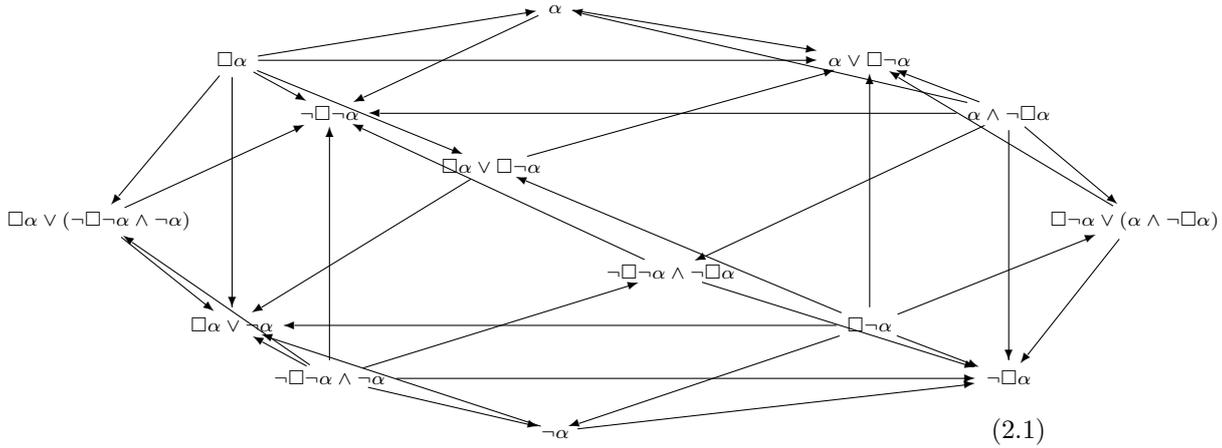
and  $\neg$  its negation) which yields in the classical interpretation: something is or is not. Then one adds a new modal operator  $\Box$  (usually interpreted as necessity); two axiom schemata ( $\alpha$  is any formula of classical propositional logic):  $\Box\alpha \rightarrow \alpha$ , known as axiom  $T$ , and  $\Box(\alpha \rightarrow \beta) \rightarrow (\Box\alpha \rightarrow \Box\beta)$ , known as axiom  $K$ ; and the inference rule of necessitation: from any theorem  $\alpha$  infer  $\Box\alpha$ . As a shortcut, the string of symbols  $\neg\Box\neg$  is written  $\Diamond$ . This yields a system known as  $KT$ , that has in fact an infinity of non equivalent affirmative and negative modalities (affirmative modalities are finite iterates of  $\Box$  and  $\neg$  with an even number of  $\neg$ , whereas negative modalities have an odd number of  $\neg$ ). If we add to  $KT$  the axiom  $\Box\alpha \rightarrow \Box\Box\alpha$  (also known as 4, or reflexivity of necessity, or positive introspection: if I know something then I know that I know it), the system obtained is called classical  $S4$  modal logic. If one adds to  $S4$  (or equivalently to  $KT$  as is shown in any textbook on modal logic) the axiom  $\Diamond\alpha \rightarrow \Box\Diamond\alpha$ , i.e.  $\neg\Box\neg\alpha \rightarrow \Box\neg\Box\neg\alpha$  (also known as 5, or negative introspection: if I do not know something then I know that I do not know it), the system obtained is called classical  $S5$  modal logic. It is easy to see that the system  $S4$  has actually 14 distinct irreducible modalities ( $\cdot, \Box, \Diamond, \Diamond\Box, \Box\Diamond, \Diamond\Box\Diamond, \Diamond\Box\Diamond\Box$ , and their negations) while the system  $S5$  has even fewer distinct irreducible modalities ( $\cdot, \Box, \Diamond$ , and their negations).

As the previous lines show,  $S5$  is a rather strong system, due to its numerous additional axioms. It is largely used, as it is the “simplest” modal extension of classical logic, in the sense that any finite combination of  $\neg, \Box, \Diamond$  in front of a formula can be reduced to a combination of at most one  $\neg$  and one  $\Box$ . Furthermore, returning to our initial motivation of dealing with Aristotle's square, if one looks carefully at the modal interpretations of that square that led to medieval syllogistic, it appears that  $S5$  was actually the logic used: although the left edge of the basic Aristotelian square ( $\Box\alpha \rightarrow \Diamond\alpha$ ) needs only a very weak modal logic (known as the  $D$  system) which is of interest to those exploring the deontic interpretations of the modal operators, there is a need for stronger logics to address the usual modal perplexities by making all modal truths necessary (axioms 4 and 5), and  $S5$  is the commonly employed system for this. Indeed, it is known as the system to capture the notion of Leibnizian possible worlds.

Back to our main thread, let us take the following modal 3(3)-graph and interpret it within classical  $S5$  modal logic:

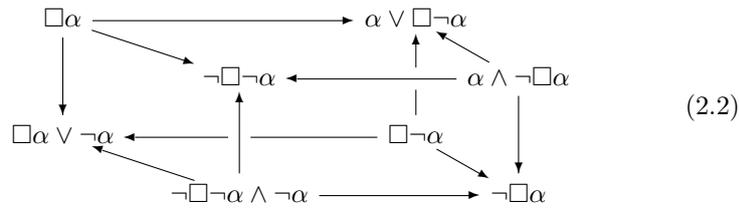


This yields the tetracosahedron illustrated in figure 3 (in (1.1) replace 1 by  $\Box\alpha$ , 2 by  $\alpha \wedge \neg\Box\alpha$ , 3 by  $\neg\Box\neg\alpha \wedge \neg\alpha$  and 4 by  $\Box\neg\alpha$ , and any combination of digits by the disjunction of the corresponding replacements).

FIGURE 3. The  $S5$  modal logical tetraicosahedron.

Since a sub-alternation relationship is the same as logical implication, it is easy to see that the vertices of the tetraicosahedron (2.1) are none else than all possible formulas (except  $\top$  and  $\perp$  which are the two additional vertices of the “completed” tetraicosahedron) one can build from arbitrary recursive application of  $\wedge$ ,  $\vee$ ,  $\neg$ ,  $\Box$  on  $\alpha$ , using classical  $S5$  modal logic axioms.

The various “strong” and “weak” subfigures exhibited before can be deduced easily. Among them, the strong logical cube, illustrated in figure 4, is a logical bisimplex of dimension 3, built from two distinct tetrahedra: one of contrariety (any two of its vertices are contrary), one of sub-contrariety (any two of its vertices are sub-contrary). The vertices of the tetrahedron of contrariety are those from which the sub-alternation arrows start. Furthermore any vertex of the cube is contradictory to the furthest lying opposite vertex (i.e. obtained by central symmetry when the cube is considered as a geometrical figure).

FIGURE 4. The  $S5$  modal logical cube.

Thus the logical cube is a three-dimensional generalization of the square of oppositions (where the contrariety simplex is a segment) and of the paracomplete

and paraconsistent hexagons (where the contrariety simplex is a triangle). As such it can be used as a model of formal reasoning [ADL<sup>+</sup>07].

## 2.2. Aristotelian-like modal squares

Due to its construction, the logical cube contains six squares of opposition, and features several operators of contingency ( $\alpha \wedge \neg \Box \alpha$  and  $\neg \Box \neg \alpha \wedge \neg \alpha$ ) which refine the case of pure contingency  $\neg \Box \alpha \wedge \neg \Box \neg \alpha$  of Aristotle's modal square, introduced by [Bla66]. These operators are those introduced in the paracomplete and paraconsistent hexagons discussed previously. Figure 5 lists these 6 squares of opposition.

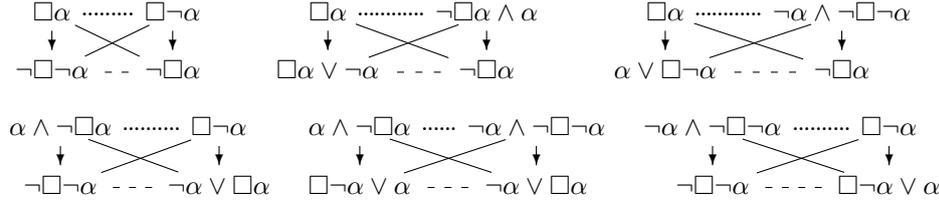


FIGURE 5. The 6 squares of opposition of the  $S5$  modal logical cube.

In addition to these six squares which exhaust the  $S5$  modal logical cube, there are twelve additional squares that are included in the  $S5$  modal logical tetraicosahedron. They are listed in figure 6.

Together these eighteen squares exhaust all possible squares of opposition that can be written within classical  $S5$  modal logic.

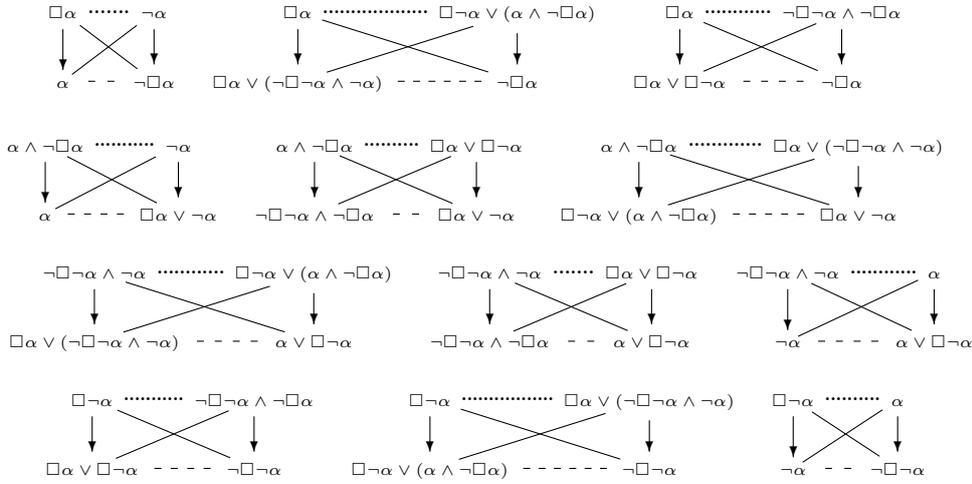


FIGURE 6. The 12 squares of opposition included in the  $S5$  modal logical tetraicosahedron that are not in the  $S5$  modal logical cube.

### 2.3. Towards semantic interpretations of the $S5$ modal logical cube

Following Aristotelian or Medieval tradition, we could use the  $S5$  modal logical cube (2.2) as a model of reasoning within classical  $S5$  modal logic. For instance, in [ADL<sup>+</sup>07], we have used it to supervise the entire process of theory formation while studying a phenomenon, as done by chemists studying properties of a new molecule such as absorption or toxicity.

Let us propose therefore interpretations in natural language of the various vertices. These interpretations are inspired from those usual within temporal, epistemic or doxastic logics:

- $\Box\alpha$ : the fact described by formula  $\alpha$  is proven;
- $\alpha \vee \Box\neg\alpha$ : the fact is either observed or refuted; if the logic were classical, this formula would be rewritten as  $\neg\alpha$  implies  $\Box\neg\alpha$ , which is interpreted as an excessive non-assertion, or hyperbolic doubt as exposed by Descartes;
- $\neg\Box\neg\alpha$ : the fact is not refuted ; it is possible; given as an advice;
- $\alpha \wedge \neg\Box\alpha$ : the fact is observed and not proven; it is conjectured;
- $\Box\alpha \vee \neg\alpha$ : the fact is either proven or its negation has been observed; if the logic were classical, this formula would be rewritten as  $\alpha$  implies  $\Box\alpha$ , which is interpreted as an excessive assertion;
- $\Box\neg\alpha$ : the fact is refuted;
- $\neg\Box\neg\alpha \wedge \neg\alpha$ : the fact is not refuted and not observed; it is taken as a postulate;
- $\neg\Box\alpha$ : the fact is not proven; its negation is possible.

## 3. How modal is the logical tetraicosahedron?

### 3.1. Syntactic translation

The logical tetraicosahedron (1.1) is decorated by modalities from classical  $S5$  modal logic. However we have the following result.

**Proposition 3.1.** *The  $S5$  modal logical tetraicosahedron (1.1) can be translated into the logical tetraicosahedron built on all formulas generated by two propositional variables and admissible within classical propositional logic.*

*Proof.* Let us take  $p = \alpha$  and  $q = \Box\alpha \vee \Box\neg\alpha$ . The tetraicosahedron (2.1) can be rewritten as the tetraicosahedron illustrated in figure 7.

One recognizes here the 14 non-trivial (i.e. all except the always true tautology and the always false contradiction, which correspond to both additional vertices mentioned in section 2.1) truth-value combinations one can build within classical logic with two propositions, which are respectively known as: assertion (respectively of  $p$  and of  $q$ ), negation (respectively of  $p$  and of  $q$ ), disjunction ( $p \vee q$ ), direct implication ( $p \supset q$ , equal to  $\neg p \vee q$ ), converse implication ( $q \supset p$ ), direct non-implication ( $\neg(p \supset q)$ ), converse non-implication ( $\neg(q \supset p)$ ), equivalence ( $p \equiv q$ , equal to  $(p \wedge q) \vee \neg(p \vee q)$ ), alternative ( $p \vee\vee q$ , equal to  $(p \vee q) \wedge \neg(p \wedge q)$ ), incompatibility ( $\neg p \vee \neg q$ ), rejection ( $\neg p \wedge \neg q$ ).  $\square$

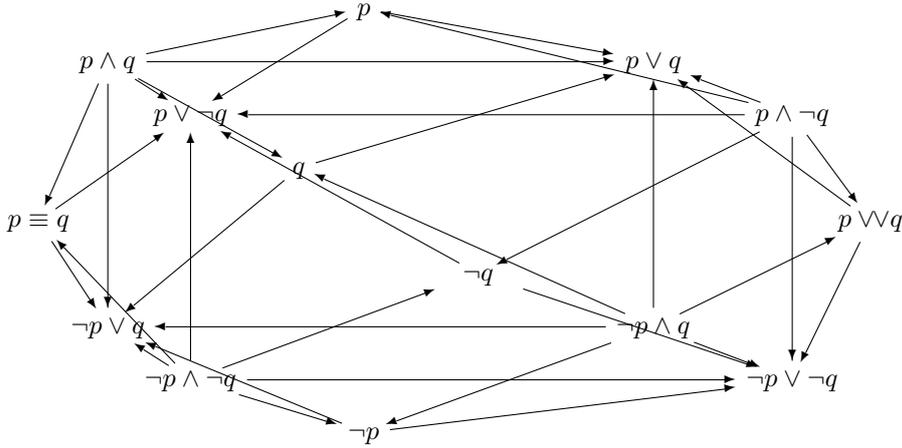


FIGURE 7. The classical logical tetraicosahedron.

Since this translation maps the  $S5$  modal logical tetraicosahedron into the full classical tetraicosahedron that can be obtained from classical logic, it means that the  $S5$  modal logical tetraicosahedron takes its modal flavor only because it is modally decorated, but not from its inherent logic.

This was already noticed in [Pel06], but we have exhibited here another translation which is interesting from a historical point of view, since Blanché has introduced part of the classical logical tetraicosahedron in [Bla57] and parts of the  $S5$  modal logical tetraicosahedron in [Bla66] but did not connect them as we have just done.

### 3.2. Semantic translation

Whereas the previous translation was a mere rewriting of the vertices, we introduce now another translation based on a possible-world semantic interpretation of classical  $S5$  modal logic, which interprets that logic within a multivalued logic [Res69, DDT78, Tha88].

This will yield another translation of the  $S5$  modal logical tetraicosahedron into the classical logical tetraicosahedron built on the Boolean lattice obtained from the set  $\{1, 2, 3, 4\}$ , i.e. the “Boolean” tetraicosahedron (1.1).

**Proposition 3.2.** *The embedding of classical  $S5$  modal logic within a four-valued logic yields a translation of the  $S5$  modal logical tetraicosahedron (2.1) into the “Boolean” tetraicosahedron (1.1).*

*Proof.* The proof is based on the interpretation of classical  $S5$  modal logic within a multivalued logic, obtained by considering a possible-world semantic with 2 worlds: the “current” world  $X$  and a “possible” world  $Y$ .

Let  $V$  be a 4-valued valuation, i.e. a mapping from the set of formulas into  $\{v_1, v_2, v_3, v_4\}$ . To any proposition, a logical value among these four values can be

attributed depending on its truth or falsity in  $X$  and  $Y$ :  $v_1$  if it is true in  $X$  and  $Y$  (meaning necessarily true),  $v_2$  if it is true in  $X$  and false in  $Y$  (meaning currently true but not necessarily true),  $v_3$  if it is false in  $X$  and true in  $Y$  (meaning currently false but not necessarily false),  $v_4$  if it is false in  $X$  and  $Y$  (meaning necessarily false).

Let  $\mathcal{D}$  be the set of designated values  $\{v_1, v_2\}$  and  $\mathcal{U}$  be the set of undesignated values  $\{v_3, v_4\}$ . Remember that in a multivalued logic designated values are the ones which “count as true” when one is not interested in the fine structure of the truth set, and the undesignated values “count as false”.

With this interpretation,  $v_1$  and  $v_2$  count as true,  $v_3$  and  $v_4$  as false,  $v_1$  and  $v_4$  as non contingent,  $v_2$  and  $v_3$  as contingent,  $v_1$  as necessarily true,  $v_4$  as necessarily false,  $v_1, v_2$  and  $v_3$  as possibly true (not necessarily false),  $v_2, v_3$  and  $v_4$  as possibly false (not necessarily true).

This yields the following truth table for a formula  $\alpha$ :

$\alpha$	$\neg\alpha$	$\Box\alpha$	$\Box\neg\alpha$	$\neg\Box\alpha$	$\neg\Box\neg\alpha$
$v_1$	$v_4$	$v_1$	$v_4$	$v_4$	$v_1$
$v_2$	$v_3$	$v_4$	$v_4$	$v_1$	$v_1$
$v_3$	$v_2$	$v_4$	$v_4$	$v_1$	$v_1$
$v_4$	$v_1$	$v_4$	$v_1$	$v_1$	$v_4$

Conjunction and disjunction of formulas are interpreted by introducing a total order on  $\mathcal{D} \cup \mathcal{U}$  defined as  $v_4 < v_3 < v_2 < v_1$ , and interpreting conjunction as the lower bound and disjunction as the upper bound.

A canonical notion of entailment  $\models$  can be then defined by saying that a formula  $\psi$  follows from a formula  $\phi$ , denoted by  $\phi \models \psi$ , whenever all models of  $\phi$  are also models of  $\psi$ , that is  $V(\psi) \in \mathcal{D}$  whenever  $V(\phi) \in \mathcal{D}$ . A look at the previous truth table shows that:  $\Box\alpha \models \alpha \models \neg\Box\neg\alpha$  and  $\Box\neg\alpha \models \neg\alpha \models \neg\Box\alpha$ . One recognizes here the semantic interpretation of the modal 3(3)-graph introduced previously.

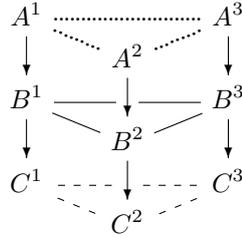
Let us now map any formula  $\phi$  into the set of indices of the designated values taken by  $V(\alpha)$  for which  $V(\phi) \in \mathcal{D}$ . This map is well-defined since all formulas  $\phi$  considered are recursively built from conjunctions, disjunctions, negations, and modal decorations of  $\alpha$ . This yields for instance:  $\Box\alpha \mapsto \{1\}$ ,  $\alpha \mapsto \{1, 2\}$ ,  $\neg\Box\neg\alpha \mapsto \{1, 2, 3\}$ ,  $\Box\neg\alpha \mapsto \{4\}$ ,  $\neg\alpha \mapsto \{3, 4\}$ ,  $\neg\Box\alpha \mapsto \{2, 3, 4\}$ . More generally if a formula  $\alpha_1$  maps into some subset of  $\{1, 2, 3, 4\}$ , and a formula  $\alpha_2$  maps into some subset of  $\{1, 2, 3, 4\}$ , then  $\alpha_1 \vee \alpha_2$  maps into the set of indices of the designated values taken by  $V(\alpha)$  for which  $V(\alpha_1 \vee \alpha_2) \in \mathcal{D}$  or equivalently for which  $V(\alpha_1) \cup V(\alpha_2) \in \mathcal{D}$  (by definition of the interpretation of the disjunction), which is none else that the union of the subsets into which  $\alpha_1$  and  $\alpha_2$  map. Mutatis mutandis, conjunction corresponds to intersection of the subsets.

Thus this mapping is none else than Pellissier’s encoding, which led to the “Boolean” tetracosahedron.  $\square$

### 4. A higher-order generalization

The previous constructions have shown all the extensions of Aristotle's square that can be obtained by adding the modality  $\Box$ . However one could also add further modal operators, as suggested in the construction of higher-order geometrical figures of opposition in [Mor04].

Let us thus define some classical propositional multimodal logic, and following [Mor04], let us start from an arbitrary modal 4(3)-graph:

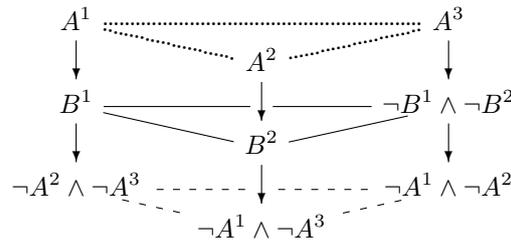


where arrows stand for sub-alternation,  $A$ 's,  $B$ 's and  $C$ 's are formulas, and opposition is defined geometrically as: the disjunction of two vertices of any of the three triangles is contradictory to the vertex that can be obtained by central symmetry relatively to the center of gravity of the middle triangle. In other words, if  $(i, j, k)$  is any permutation of  $\{1, 2, 3\}$ ,  $A^i \vee A^j$  is contradictory to  $C^k$ ,  $C^i \vee C^j$  is contradictory to  $A^k$ ,  $B^i \vee B^j$  is contradictory to  $B^k$ .

As before, we will try now to construct the geometrical figure composed of all the combinations of the modalities obtained by applying  $\wedge$ ,  $\vee$  and  $\neg$ . We have the following proposition.

**Proposition 4.1.** *The 4-oppositional figure generated from an arbitrary modal 4(3)-graph is a tetracosahedron. More precisely, all formulas generated by 4-opposition lie on the vertices or the edges of a "completed" tetracosahedron.*

*Proof.* We will first consider only the  $A$ 's and the  $C$ 's. In the next paragraphs, we adopt the following convention: when we write an expression such as  $A^i \wedge A^j$ , we really mean different exponents which can take any value within  $\{1, 2, 3\}$ ; whereas  $A^i \wedge A^i$  refers to the same  $A^i$ . Using the opposition properties applied to  $C$ 's and  $B$ 's, the previous graph can be rewritten as:



Sub-alternation reads as  $A^i \rightarrow \neg A^j \wedge \neg A^k$ , which implies  $A^i \rightarrow \neg A^j$ . Thus  $A^i \wedge \neg A^j = A^i$  and  $A^i \wedge A^j \rightarrow \neg A^j \wedge A^j = \perp$ , whence  $A^i \wedge A^j = \perp$ , while obviously  $A^i \wedge A^i = A^i$ .

From that follows  $A^1 \wedge A^2 \wedge A^3 = \perp$ . It is also easy to verify that if we express each vertex (indexed by  $i$ ) of the upper triangle as the negation of the disjunction of both opposite vertices  $C^j = \neg A^i \wedge \neg A^k$  and  $C^k = \neg A^i \wedge \neg A^j$  (opposition should define the triangle of  $C$ 's from the triangle of  $A$ 's and conversely), then we obtain  $A^i$  as expected. Hence all opposition requirements do not introduce any new constraints on the  $A$ 's or the  $B$ 's.

Applying negation to the previous results, we have  $\neg A^i \vee \neg A^j = \top$ ,  $\neg A^i \vee A^j = \neg A^i$  and  $\neg A^i \vee \neg A^j \vee \neg A^k = \top$ .

To sum up, the closure of the  $A$ 's by  $\{\wedge, \vee, \neg\}$  yields the terms  $\{\perp, \top, A^i, \neg A^i, A^i \vee A^j, \neg A^i \wedge \neg A^j, A^i \vee A^j \vee A^k, \neg A^i \wedge \neg A^j \wedge \neg A^k\}$  with the sub-alternation relations between terms illustrated in figure 8.

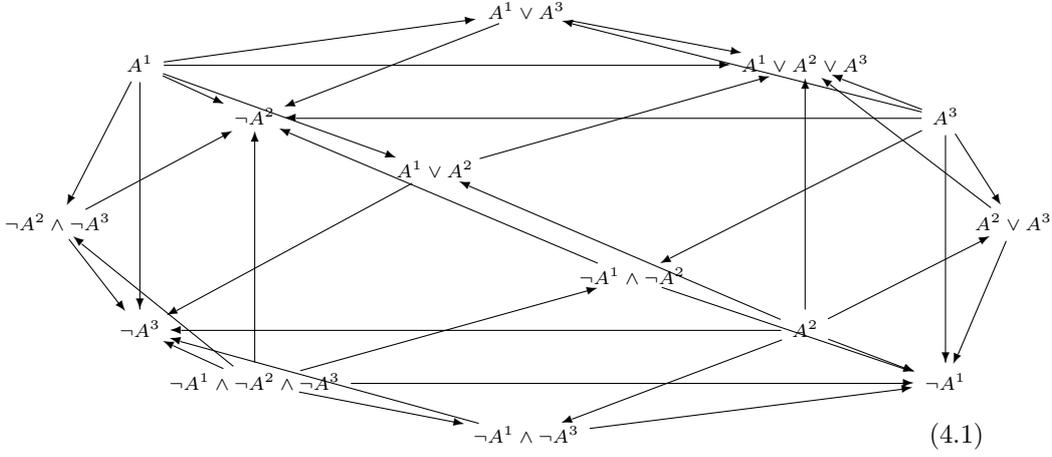


FIGURE 8. The tetraicosahedron corresponding to the  $A$ 's and  $C$ 's of a modal 4(3)-graph.

Actually, this tetraicosahedron has two additional vertices corresponding to  $\top$  and  $\perp$ , as mentioned in section 2.1, with the obvious sub-alternation relations:  $\perp \rightarrow \neg A^1 \wedge \neg A^2 \wedge \neg A^3$  and  $A^1 \vee A^2 \vee A^3 \rightarrow \top$ .

This is no surprise, since up to now, we have restricted our study to  $A$ 's and  $C$ 's, i.e. we have considered actually a modal 4(2)-graph: but if we remember the corollary from [Pel06] recalled in a previous section, a modal 4(2)-graph and a modal 3(3)-graph are translated into the same Boolean lattice corresponding to a set of 4 elements, which is the tetraicosahedron (1.1).

Let us see now how the  $B$ 's fit in this framework (in the sequel we do not use the fact that  $B^3 = \neg B^1 \wedge \neg B^2$  and use only the notation  $B^3$ , in order to have as much symmetry as possible in the notations).

Let us introduce, with the name  $E$ , the collection of the following edges:  $e_1$  stands for any of the three edges  $A^i \rightarrow \neg A^j \wedge \neg A^k$ ,  $e_2$  stands for any of the three edges  $A^i \vee A^j \rightarrow \neg A^k$ ,  $e_3$  is the edge  $\perp \rightarrow \neg A^i \wedge \neg A^j \wedge \neg A^k$ ,  $e_4$  is the edge  $A^i \vee A^j \vee A^k \rightarrow \top$ . Notice that  $e_3$  and  $e_4$  involve the ‘‘completed’’ tetraicosahedron.

Since  $A^i \rightarrow B^i \rightarrow \neg A^j \wedge \neg A^k$ , we have  $A^j \vee A^k \rightarrow \neg B^i \rightarrow \neg A^i$  and:

$$\begin{aligned} A^i \wedge A^j &= \perp \rightarrow B^i \wedge B^j \rightarrow \neg A^i \wedge \neg A^j \wedge \neg A^k, \\ A^i \vee A^j &\rightarrow B^i \vee B^j \rightarrow (\neg A^j \wedge \neg A^k) \vee (\neg A^i \wedge \neg A^k) = \neg A^k, \\ A^i \wedge (A^i \vee A^k) &= A^i \rightarrow B^i \wedge \neg B^j \rightarrow \neg A^j \wedge \neg A^k. \end{aligned}$$

Therefore the terms  $B^i$ ,  $B^i \wedge B^j$ ,  $B^i \vee B^j$ ,  $B^i \wedge \neg B^j$  and their negations lie on the edges of  $E$ . Notice that  $E$  is closed under ‘‘negation’’, i.e. whenever a term lies on an edge defined by two vertices, the negation of the term lies on the edge defined by the negations of both vertices, and this edge is still in  $E$ .

For sake of simplicity, we adopt now the following convention:  $e_1 \wedge A^l$  (resp.  $e_1 \vee A^l$ ) means the set of edges obtained from any edge  $e_1$  by composing its source and target by conjunction (resp. disjunction) with  $A^l$ . Remember that  $i, j, k$  are mute indices within  $\{1, 2, 3\}$  and that when we write different (resp. the same) indices, we really mean they are different (resp. the same).

$e_1 \wedge A^i = A^i \rightarrow A^i$	$e_2 \wedge A^i = A^i \rightarrow A^i$
$e_1 \wedge A^j = \perp \rightarrow \perp$	$e_2 \wedge A^k = \perp \rightarrow \perp$
$e_1 \wedge (A^i \vee A^j) = A^i \rightarrow A^i$	$e_2 \wedge (A^i \vee A^j) = A^i \vee A^j \rightarrow A^i \vee A^j$
$e_1 \wedge (A^j \vee A^k) = \perp \rightarrow \perp$	$e_2 \wedge (A^i \vee A^k) = A^i \rightarrow A^i$
$e_1 \wedge (A^i \vee A^j \vee A^k) = A^i \rightarrow A^i$	$e_2 \wedge (A^i \vee A^j \vee A^k) = A^i \vee A^j \rightarrow A^i \vee A^j$
$e_1 \vee A^i = A^i \rightarrow \neg A^j \wedge \neg A^k$	$e_2 \vee A^i = A^i \vee A^j \rightarrow \neg A^k$
$e_1 \vee A^j = A^i \vee A^j \rightarrow \neg A^k$	$e_2 \vee A^k = A^i \vee A^j \vee A^k \rightarrow \top$
$e_1 \vee (A^i \vee A^j) = A^i \vee A^j \rightarrow \neg A^k$	$e_2 \vee (A^i \vee A^j) = A^i \vee A^j \rightarrow \neg A^k$
$e_1 \vee (A^j \vee A^k) = A^i \vee A^j \vee A^k \rightarrow \top$	$e_2 \vee (A^i \vee A^k) = A^i \vee A^j \vee A^k \rightarrow \top$
$e_1 \vee (A^i \vee A^j \vee A^k) = A^i \vee A^j \vee A^k \rightarrow \top$	$e_2 \vee (A^i \vee A^j \vee A^k) = A^i \vee A^j \vee A^k \rightarrow \top$
$e_3 \wedge A^i = \perp \rightarrow \perp$	$e_4 \wedge A^i = A^i \rightarrow A^i$
$e_3 \wedge (A^i \vee A^j) = \perp \rightarrow \perp$	$e_4 \wedge (A^i \vee A^j) = A^i \vee A^j \rightarrow A^i \vee A^j$
$e_3 \wedge (A^i \vee A^j \vee A^k) = \perp \rightarrow \perp$	$e_4 \wedge (A^i \vee A^j \vee A^k) = A^i \vee A^j \vee A^k \rightarrow A^i \vee A^j \vee A^k$
$e_3 \vee A^i = A^i \rightarrow \neg A^j \wedge \neg A^k$	$e_4 \vee A^i = A^i \vee A^j \vee A^k \rightarrow \top$
$e_3 \vee (A^i \vee A^j) = A^i \vee A^j \rightarrow \neg A^k$	$e_4 \vee (A^i \vee A^j) = A^i \vee A^j \vee A^k \rightarrow \top$
$e_3 \vee (A^i \vee A^j \vee A^k) = A^i \vee A^j \vee A^k \rightarrow \top$	$e_4 \vee (A^i \vee A^j \vee A^k) = A^i \vee A^j \vee A^k \rightarrow \top$

TABLE 1. Proof of the closure of  $E$  under composition with admissible  $A$  terms: note that an edge like  $A^i \rightarrow A^i$  is actually reduced to the vertex  $A^i$ .

Table 1 shows how the edges of  $E$  are transported when composed by conjunction and disjunction with  $A^i$ ,  $A^i \vee A^j$ ,  $A^i \vee A^j \vee A^k$ . For instance,  $e_1 \wedge A^i$

stands for any of the three edges  $A^i \wedge A^i \rightarrow \neg A^j \wedge \neg A^k \wedge A^i$ , i.e.  $A^i \rightarrow A^i$ , since  $\neg A^j \wedge (\neg A^k \wedge A^i) = \neg A^j \wedge A^i = A^i$ .

Looking at all cases, we see that  $E$  is closed under the compositions with the various terms that can be built with the  $A$ 's. Since  $E$  is closed under negation, we conclude that all the terms that can be built from  $A$ 's and  $B$ 's with any of  $\{\wedge, \vee, \neg\}$  lie on some edge of  $E$ .  $\square$

From the previous proof it is straightforward to determine the edges and vertices of the tetraicosahedron (4.1) on which the composite terms with  $A$ 's and  $B$ 's lie. This is helpful when additional relations are given between the various modalities, as in the next section.

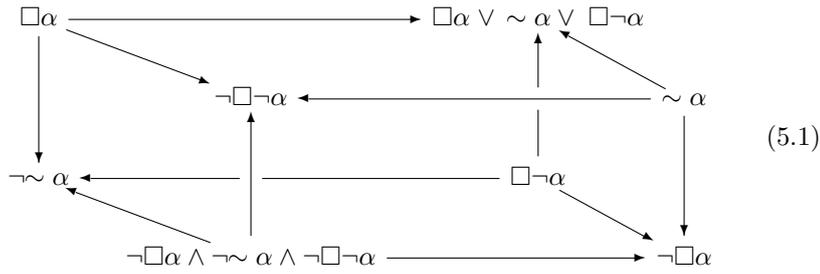
## 5. Special cases of the higher-order generalizations

We show now how the tetraicosahedron (4.1) encompasses all previous logical figures. This will be achieved by taking special values for  $A^1, A^2, A^3, B^1, B^2$ , using the usual modal operator  $\square$  as well as new modal operators.

Let us take  $B^1 = \square\alpha$  and  $B^2 = \square\neg\alpha$ , where  $\square$  satisfies the axioms of classical  $S5$  modal logic. If we assume the triangle of  $A$ 's to be also "strong", then  $A^3 = \neg A^1 \wedge \neg A^2$ , and the sub-alternation relation  $A^1 \rightarrow \square\alpha \rightarrow \neg A^2 \wedge \neg(\neg A^1 \wedge \neg A^2) = \neg A^2 \wedge A^1 = A^1$  implies  $A^1 = \square\alpha$ . In a similar way,  $A^2 = \square\neg\alpha$ . It is easy to see that the tetraicosahedron (4.1) collapses to the classical hexagon from figure 1.

Let us take  $A^1 = \square\alpha, A^2 = \square\neg\alpha, A^3 = \alpha \wedge \neg\square\alpha$ . Take also  $B^1 = \square\alpha$  and  $B^2 = \square\neg\alpha$  which is consistent with all sub-alternation relations. Then the cube underlying the tetraicosahedron (4.1) yields the strong logical cube (2.2).

Let us now take  $A^1 = \square\alpha, A^2 = \square\neg\alpha$ , and  $A^3 = \sim\alpha$  where  $\sim$  is a new modal operator. The choice of  $B^1 = \square\alpha$  and  $B^2 = \square\neg\alpha$  is consistent with all sub-alternation relations. This assignment of modalities corresponds to the modal 4(3)-graph mentioned in [Mor04], where:  $\beta = \square, | = \sim$  and  $\beth = \square\neg$ , and the cube underlying the tetraicosahedron (4.1) yields:



which is constructed from the six squares of opposition illustrated in figure 9.

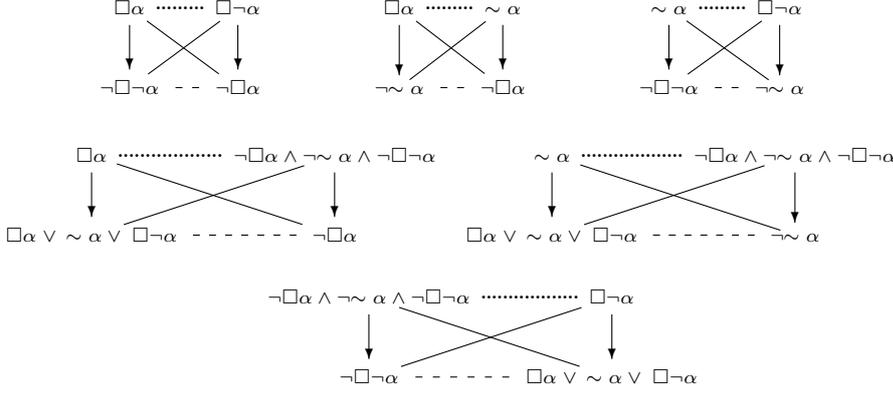


FIGURE 9. The 6 squares included in the multimodal logical cube (5.1).

By analogy with the Aristotelian square, we can have the intuition from the second and third top squares of figure 9 that  $\sim$  should have the flavor of a negation (negative particular assertion) and of a necessity modality (affirmative universal assertion). Obviously classical negation does not work in this case, however an interpretation within society semantics [CLM05] is adequate, as we elaborate below.

## 6. Semantics associated to the modal operator $\sim$

In the next paragraphs, we give a Kripke-based semantics for the modal operator  $\sim$  introduced previously. A Kripke model  $K$  is a triple  $(W, \prec, V)$ , where  $W$  is a non-empty set,  $\prec$  a binary relation on  $W$ , and  $V$  is a mapping (the valuation) that assigns a subset  $V(p)$  of  $W$  to each propositional variable  $p$ .  $W$  is meant to be the set of possible worlds,  $w \prec v$  is meant to say that  $v$  is reachable from  $w$ , and  $V(p)$  is intended to be the set of worlds at which  $p$  is true under the valuation  $V$ .

The following notation is usual:  $w \models_K p$  iff  $w \in V(p)$ . The forcing relation  $\models_K$  is defined recursively on all formulas by:  $w \models_K \neg A$  iff  $w \not\models_K A$ ;  $w \models_K A \vee B$  iff  $w \models_K A$  or  $w \models_K B$ ;  $w \models_K A \wedge B$  iff  $w \models_K A$  and  $w \models_K B$ ;  $w \models_K A \rightarrow B$  iff  $w \not\models_K A$  or  $w \models_K B$ ;  $w \models_K \Box A$  iff for every  $v$  such that  $w \prec v$ ,  $v \models_K A$ .

The semantics presented below is inspired from [Car00] and [CLM05]. It is given by a local forcing relation  $\models_{1,2}$  based on two different Kripke models defined on the same relation  $\prec$ ,  $\models_1$  and  $\models_2$ , which can be interpreted as two different observation or measurement processes that yield potentially different results, due to their intrinsic difference. The formal definition is:

- $w \models_{1,2} p$  iff  $w \models_1 p$  or  $w \models_2 p$
- $w \models_{1,2} \neg A$  iff  $(w \not\models_{1,2} A)$
- $w \models_{1,2} A \wedge B$  iff  $w \models_{1,2} A$  and  $w \models_{1,2} B$
- $w \models_{1,2} A \vee B$  iff  $w \models_{1,2} A$  or  $w \models_{1,2} B$

- $w \models_{1,2} A \rightarrow B$  iff  $w \not\models_{1,2} A$  or  $w \models_{1,2} B$
- $w \models_{1,2} \Box A$  iff  $w \models_1 \Box A$  and  $w \models_2 \Box A$
- $w \models_{1,2} \sim A$  iff  $(\exists v_1, w \prec v_1) (v_1 \models_1 A)$  and  $(\exists v_2, w \prec v_2) (v_2 \models_2 \neg A)$

In plain words,  $\sim A$  means that we might observe  $A$  with measurement process 1 and  $\neg A$  with measurement process 2 (not necessarily at the same time though). Therefore, as stated previously intuitively, the modality  $\sim$  has both a positive and a negative touch, and it behaves in a paraconsistent way, since we can find some  $A$  such that  $x \models_{1,2} \sim A$  and  $x \models_{1,2} \sim \neg A$ , without trivializing the whole logical framework.

We deduce from the former definitions:  $w \models_{1,2} \neg \Box A$  iff  $w \not\models_{1,2} \Box A$  iff  $w \not\models_1 \Box A$  or  $v \not\models_2 \Box A$  iff  $(\exists u_1, w \prec u_1) (u_1 \models_1 \neg A)$  or  $(\exists u_2, w \prec u_2) (u_2 \models_2 \neg A)$ .

We also deduce:  $w \models_{1,2} \neg \Box \neg A$  iff  $(\exists u_1, w \prec u_1) (u_1 \models_1 A)$  or  $(\exists u_2, w \prec u_2) (u_2 \models_2 A)$ .

From this, it is easy to see that  $w \models_{1,2} \sim A$  implies  $w \models_{1,2} \neg \Box A$ , as well as  $w \models_{1,2} \neg \Box \neg A$ . This shows that the consistency relations imposed by the subalternation relations of the modal 4(3)-graph are satisfied.

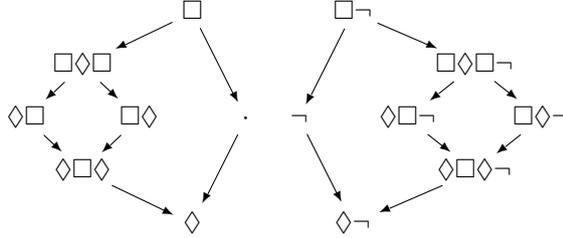
This semantics is actually an extension of what was discussed in section 2.3 with two potentially disagreeing observers instead of a unique observer.

## 7. Changing the logical framework

### 7.1. From classical modal $S5$ logic to classical modal $S4$ modal logic

In the previous paragraphs, the logic considered in order to interpret the various formulas was the classical  $S5$  modal logic, which relies on the chain of affirmative modalities  $\Box \rightarrow \cdot \rightarrow \neg \Box \neg$  and the chain of negative modalities  $\Box \neg \rightarrow \neg \rightarrow \neg \Box$ .

Instead of  $S5$ , one can consider classical  $S4$  modal logic, which relies on the following chain of affirmative and negative modalities:



What has been done previously within classical  $S5$  modal logic can then be done within classical  $S4$  modal logic: the  $n$ -oppositional figures are obtained by considering all the formulas generated from arbitrary (admissible) recursive application of the modal operators, conjunction and negation on some  $\alpha$ , and using the lattice properties (conserved through conjunction and disjunction with any formula) between the modalities in order to deduce the logical implication between all these formulas. It is easy to see that the set of generated formulas

is finite as there is a trivial injective morphism from it into the algebra of all subsets of a set whose cardinality is equal to the number of all possible positive and negative modalities.

Of course, if one switches then to weaker logics, the  $n$ -oppositional figures might not be finite since there might be an infinite number of admissible modalities, as was recalled earlier for some modal logics.

**7.2. From classical modal logic to intuitionistic modal logic**

Instead of considering classical modal logics, i.e. logics built on classical propositional logic, it is possible to change that too, and switch for instance to intuitionistic  $S5$  or  $S4$  modal logics, where the negation operator is no longer involutive. This breaks the duality between affirmative and negative modalities.

[Fon86] shows that intuitionistic  $S5$  modal logic has 5 affirmative modalities ( $\Box \langle \begin{matrix} \nearrow \\ \searrow \\ \Box \neg \neg \end{matrix} \rangle \neg \neg \rightarrow \neg \Box \neg$ ) and 4 negative modalities ( $\Box \neg \rightarrow \neg \rightarrow \neg \Box \neg \neg \rightarrow \neg \Box$ ).

The same reference shows that intuitionistic  $S4$  modal logic has a system of 17 affirmative modalities and 14 negative modalities, with a lack of symmetry (or duality) between these two groups.

Therefore the lattice built from the closure of both these sets of modalities through conjunction and disjunction will yield more complicated geometrical figures than previously, since some of the inferences made in the previous proofs, which reduced the number of non-equivalent terms, are not valid any longer (e.g. double negation does not imply affirmation, and among de Morgan’s laws  $\neg(\phi \wedge \psi) \rightarrow \neg\phi \vee \neg\psi$  does not hold).

However since these logics yield interesting epistemological views on diverse positive and negative introspection capabilities, the obtained  $n$ -oppositional figures could provide a model of formal reasoning on propositions. The challenge is more to find an application, than to exhibit the logical geometrical figures, since the latter are only a matter of combinatorics.

**8. Future work**

In this paper we have seen how Aristotle’s square could be generalized, and we have exhibited various translations between some generalizations. We have only hinted at how the latter could be used as models for formal reasoning by proposing particular semantics for the generalizations proposed. Obviously many other logical frameworks can be defined. Furthermore the methods introduced in this paper can be applied to the higher-order extensions of [Mor04] in a straightforward way: the only difficulty associated with the introduction of additional modal operators (a modal  $n(m)$ -graph with larger  $n$  and  $m$  implies only bigger simplices and more copies of them) is more cumbersome notations and fastidious proofs. More complex geometrical figures arise, and more complex semantics based on society semantics – with the right juggling of the Kripke forcing relations – are needed.

This paves the way for a more general theory of reasoning where Aristotle's square and its generalizations do not model only the reasoning abilities of a unique agent, but of several interacting agents. The modalities between various such agents can be related together, or can arise as the interaction between the agents, which leads then to a model of learner-teacher interactions.

These various ideas are currently formalized using category theory, representing an agent as a topos and the interaction between agents as an adjunction between the corresponding topoi [SLS07].

## References

- [ADL<sup>+</sup>07] Mohammad Afshar, Christopher Dartnell, Dominique Luzeaux, Jean Sallantin, and Yannick Tognetti. Aristotle's square revisited to frame discovery science. *Journal of Computers*, 2(5):54–66, 2007.
- [Béz02] J.-Y. Béziau. S5 is a paraconsistent logic and so is first-order logic. *Logical Investigations*, 9:301–309, 2002.
- [Béz03] J.-Y. Béziau. New light on the square of oppositions and its nameless corner. *Logical Investigations*, 10:218–233, 2003.
- [Béz05] J.-Y. Béziau. Paraconsistent logic from a modal viewpoint. *Journal of Applied Logic*, 3:7–14, 2005.
- [Bla57] R. Blanché. *Introduction à la logique contemporaine*. Armand Colin, 1957.
- [Bla66] R. Blanché. *Structures intellectuelles : essai sur l'organisation systématique des concepts*. Vrin, Paris, 1966.
- [Car00] W.A. Carnielli. Possible-translations semantics for paraconsistent logics. In D. Batens, C. Mortensen, G. Priest, and J.-P. van Bendegem, editors, *Frontiers of Paraconsistent Logic*. Studies in Logic and Computation, no. 8, Research Studies Press Ltd., 2000.
- [CLM05] W.A. Carnielli and M. Lima-Marques. Society semantics and multiple-valued logics. In *XII Encontro Brasileiro de Lógica - Advances in contemporary logic and computer science*, pages 33–52, Series Contemporary Mathematics, American Mathematical Society, 2005.
- [DDT78] M. Davio, J.P. Deschamps, and A. Thayse. *Discrete and switching functions*. Mc Graw-Hill, New York, 1978.
- [Doš84a] K. Došen. Intuitionistic double negation as a necessity operator. *Publications de l'Institut Mathématique*, 35(49):15–20, 1984.
- [Doš84b] K. Došen. Negative modal operators in intuitionistic logic. *Publications de l'Institut Mathématique*, 35(49):3–14, 1984.
- [Fon86] J.M. Font. Modality and possibility in some intuitionistic modal logics. *Notre Dame Journal of Formal Logic*, 27-4:533–546, 1986.
- [Gol05] R. Goldblatt. Mathematical modal logic: a view of its evolution. In D.M. Gabbay and J. Woods, editors, *Handbook of the History of Logic, Volume 6*. Elsevier BV, 2005.

- [Mor04] A. Moretti. Geometry for modalities? Yes: through  $n$ -opposition theory. In J.-Y. Béziau, A. Costa-Leite, and A. Facchini, editors, *Aspects of Universal Logic*, pages 102–145. Travaux de logique 17, Neuchâtel, 2004.
- [Pel06] R. Pellissier. “Setting”  $n$ -opposition. In *UNILOG05*, (to be published), 2006.
- [Res69] N. Rescher. *Many-valued logic*. Mc Graw-Hill, New York, 1969.
- [SLS07] J. Sallantin, D. Luzeaux, and J.-J. Szczéciniaarz. An epistemology of computer science based on a dialectical interaction. In *European Conference on Computing and Philosophy, Philosophy and Computer Science track, Enschede, The Netherlands*, 2007.
- [Tha88] A. Thayse et al. *Approche logique de l'intelligence artificielle, Tome 1 : de la logique classique à la programmation logique*. Dunod, 1988.

Dominique Luzeaux

e-mail: [dominique.luzeaux@polytechnique.org](mailto:dominique.luzeaux@polytechnique.org)

Jean Sallantin

LIRMM, UMR 5506

161 rue Ada

34392 Montpellier Cedex 5

France

e-mail: [jean.sallantin@lirmm.fr](mailto:jean.sallantin@lirmm.fr)

Christopher Dartnell

LIRMM, UMR 5506

161 rue Ada

34392 Montpellier Cedex 5

France

e-mail: [christopher.dartnell@lirmm.fr](mailto:christopher.dartnell@lirmm.fr)

---

**Résumé de la Thèse :** L'utilisation d'outils issus de la découverte scientifique computationnelle a montré un effet inattendu de l'interaction avec une machine apprenante : l'apparition d'un jeu dialectique de localisation et de correction d'erreurs, aussi bien dans les règles apprises par la machine que dans l'ontologie écrite par l'utilisateur. Ces erreurs mettent en évidence des biais de modélisation ou de mesure, ainsi que des biais liés aux exemples fournis à la machine. C'est en réalité ce processus dialectique qui est exploité dans les différentes applications, et l'apprentissage ne sert finalement qu'à générer des contradictions entre la théorie de la machine et celle de l'utilisateur. Cette constatation m'a incité à définir un cadre formel pour cette dialectique. Dans ce contexte, j'ai décidé de positionner ce mode d'interaction avec une machine apprenante par rapports aux principaux paradigmes d'apprentissage, afin de déterminer un protocole d'interaction adapté à la découverte scientifique computationnelle, c'est à dire adapté à l'apprentissage humain comme à l'apprentissage machine, et tenant compte d'une certaine démarche scientifique. Le protocole d'interaction vers lequel je me suis orienté peut être considéré comme une extension de l'apprentissage par *requêtes* [Angluin, 1988] proposant une restriction des requêtes d'appartenance à des expériences finies, et une distribution des requêtes d'équivalences (co-semi-décidables) sur une communauté d'apprenti. Une fois l'intuition de ce protocole dégagée, sa définition logique m'a occupé un certain temps, et je présente dans ce document une définition logique de jugements modaux fondée sur une extension du carré des oppositions.

**Mots clés :** Apprentissage Interactif, Logique, Découverte

---

**Abstract :** The use of tools typical from computational science discovery has shown an unexpected effect of the interaction with a learning machine : the apparition of a dialectic game of localization and error correction of the rules learned by the machine as well as the ontology written by the user. These errors highlight the modeling or measuring biases, as well as biases linked to the set of learning examples provided to the machine. It is actually this dialectic process which is exploited in the various applications, and machine learning is finally only used to generate contradictions between the machine's theory and the user's. This finding led me to defining a formal framework for this dialectic. In this context, I decided to position this mode of interacting with a learning machine in regards with the main learning paradigms, in order to determinate an interaction protocol suited to computational science discovery i.e., suited to both human learning and machine learning, and taking into account a certain scientific method. The interaction protocol toward which I was finally led can be considered as an extension of query learning [Angluin, 1988], with a restriction of membership queries to finite experiments, and a distribution of equivalence queries (co-semi-decidable) on a community of learners. Once the intuition of this protocol was established, I focused on its logical definition and I present in this document a logical definition of modal judgments founded on an extension of the square of oppositions.

**Key Words :** Interactive Learning, Logic, Discovery