Année 2018

MÉMOIRE

présenté en vue de
l'obtention d'une

HABILITATION À DIRIGER DES RECHERCHES

délivrée par

l'Université de Caen Normandie

SPÉCIALITÉ : Informatique

par

Grégory Bonnet

Fiabilité, honnêteté et éthique dans les systèmes d'agents autonomes

Devant le jury composé de :

Mme. Laurence Cholvy	Directrice de recherche, Onera	Rapporteure
M. Nicolas Maudet	Professeur, Sorbonne Université	Rapporteur
M. Nicolas Sabouret	Professeur, Université Paris Sud	Rapporteur
Mme. Amal El Fallah Seghrouchni	Professeure, Sorbonne Université	Examinatrice
Mme. Maroua Bouzid	Professeure, Normandie Université	Garante

Habilitation préparée au sein du Groupe de Recherche en Informatique, Image,
Automatique et Instrumentation de Caen
Équipe Modèles, Agents, Décision

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Les agents autonomes artificiels sont des machines logicielles ou physiques capables de calculer des décisions, de manière individuelle, coordonnées ou non avec d'autres agents ou avec des humains, en vue de la réalisation de buts de haut niveau qui leur ont été spécifiés. Leur introduction dans des domaines tels que le domaine militaire, la justice, le milieu médical ou encore les transports autonomes, soulève de nombreuses questions car les utilisateurs de ces systèmes ont parfois des attentes qui sont distinctes des problématiques d'optimalité ou de conformité légale du comportement des agents. En effet, l'autonomie des agents conduit à nous interroger sur des exigences en termes de *fiabilité* car dans le cas contraire le déploiement d'agents autonomes est problématique, d'*honnêteté* car dans le cas contraire la conception de systèmes coopératifs est rendue difficile et, de manière plus générale, des exigences en matière d'*éthique* car ce qui est technologiquement possible n'est pas toujours humainement ou socialement souhaitable. Ces questions ont d'autant plus d'importance dans le contexte actuel de déploiement d'un nombre croissant d'agents dans notre environnement, collaborant entre eux ou avec des humains.

C'est pourquoi ce mémoire d'Habilitation à Diriger des Recherches présente un projet de recherche qui se fonde sur les travaux que nous avons effectués depuis notre prise de fonction au GREYC (UMR CNRS 6072, Université de Caen Normandie, ENSICAEN). À l'issue de notre doctorat, nous avons fait une année et demie de post-doctorat à l'Université de Technologie de Troyes au sein de l'équipe ERA (Environnement des Réseaux Autonomes). Au cours de cette période, notre travail de recherche portait sur l'*adaptabilité dans les réseaux pair-à-pair*. Au-delà d'une expérience de co-encadrement et de publication, cette thématique a nourri nos réflexions sur la sécurité des systèmes intelligents, qui ont à leur tour fondé deux thèmes qui sont aujourd'hui au cœur de notre projet de recherche.

1. La **gestion de la malveillance dans les systèmes d'agents autonomes** est un questionnement initié lors de notre prise de fonction au GREYC (Université de Caen Normandie). Il s'agit de travaux dans le domaine des systèmes d'agents autonomes ouverts où certains agents peuvent présenter des comportements non fiables, malhonnêtes ou malveillants. Dans ce contexte, nous nous intéressons à des approches formelles comme le raisonnement automatisé, la théorie des jeux ou les systèmes de réputation afin de détecter et prévenir de tels comportements.

2. La question de l'**éthique et agents autonomes** était en gestation depuis 2008 – date à laquelle nous avons intégré le groupe de travail D2A2 (Droits et Devoirs des Agents Autonomes) de l'ancien GRD I3 – mais s'est pleinement développé depuis 2014 avec la coordination du projet ANR CONTINT ETHICAA [1] (Ethics and Autonomous Agents). La question scientifique principale est celle de la conception d'agents artificiels autonomes, interagissant avec des êtres humains, capables de prendre des décisions tenant compte de facteurs éthiques. Dans ce cadre, nous avons mené un travail interdisciplinaire avec des chercheurs en philosophie sociale autour de la modélisation du raisonnement moral et éthique, toujours en utilisant des approches formelles.

Ces thèmes nous ont permis d'identifier trois axes de recherche autour des questions de fiabilité, d'honnêtété et d'éthique des systèmes d'agents autonomes et ce mémoire a pour objectif de présenter un projet de recherche croisant ces trois axes avec trois modèles formels – les systèmes de réputation, les jeux de coalitions hédoniques et les modèles d'agents cognitifs – comme autant de moyens d'assurer des propriétés nécessaires à la mise en œuvre de l'autonomie dans ces systèmes. Ce mémoire est structuré en trois parties.

1. La partie *Positionnement et questions de recherche* forme le cœur de notre mémoire en détaillant les enjeux et les problématiques de notre projet de recherche. Cette partie se divise en deux chapitres. Le chapitre 1 est consacré à la mise en lumière des questions de fiabilité, d'honnêteté et d'éthique dans les systèmes d'agents autonomes tandis que le chapitre 2 s'attache à montrer comment ces questions trouvent écho dans les modèles formels que nous étudions.
2. La partie *Présentation des activités de recherche* a pour objectif de montrer des exemples choisis de nos réalisations présentant de manière concrète le croisement de nos questions de recherche et des modèles formels que nous étudions. Cette partie se divise en trois chapitres. Le chapitre 3 illustre les questions de fiabilité, le chapitre 4 les questions d'honnêteté et le chapitre 5 les questions d'éthique. Chacun de ces chapitres se conclut par un bilan de notre animation scientifique autour de cet axe.
3. La partie *Conclusion* dresse un bilan final de notre projet de recherche. Pour ce faire, cette partie est divisée en deux chapitres. Le chapitre 6 présente des pistes et des perspectives afin d'aller plus loin dans notre questionnement. Enfin, le lecteur pourra trouver au chapitre 7 notre curriculum vitae étendu, détaillant l'ensemble de nos publications, encadrements doctoraux et activités d'animation et de diffusion de la recherche.

1. `https://ethicaa.greyc.fr`

Première partie

Positionnement et questions de recherche

Chapitre 1

Systèmes d'agents autonomes

Sommaire

Ce chapitre a pour objectif d'introduire notre problématique de recherche. Nous clarifions dans un premier temps les concepts d'agent autonome, de système d'agents autonomes et d'autonomie. Nous montrons que l'autonomie des agents implique un certain nombre de contraintes qui, dans l'optique de déployer des agents interagissant avec des humains, nécessitent de s'assurer que trois propriétés fondamentales – la fiabilité des agents, leur honnêteté et leur respect d'une éthique – sont bien présentes dans les systèmes d'agents autonomes. Caractériser, représenter et étudier l'usage de ces trois propriétés forment alors la pierre angulaire de nos travaux de recherche, que nous déclinons en neuf questions.

1 Agents autonomes

Il convient en premier de définir ce que nous appelons *agents autonomes* et *systèmes d'agents autonomes*. Le terme *agent* est originaire du latin *agere* qui signifie diriger, conduire, gérer, agir ou faire. S'il est associé en sciences sociales à la notion d'*acteur*, en informatique il réfère intuitivement à une entité qui peut agir ou réaliser une tâche donnée. Par exemple, nous pouvons penser aux *démons* des systèmes d'exploitation ou aux applications permettant à un périphérique de communiquer avec un gestionnaire dans les réseaux. Toutefois, le terme agent prend dans le domaine de l'intelligence artificielle une dimension particulière.

1.1 Un bestiaire d'agents

Dans le domaine de l'intelligence artificielle, la notion d'agent est une métaphore commune et pratique pour considérer tout à la fois des logiciels, des robots ou même des êtres humains à l'aide d'un même concept qui peut pourtant reposer sur des modèles internes variés. Une manière d'appréhender cela est de considérer un certain nombre de définitions courantes de ce qu'est un agent dans la littérature [Shoham, 1993, Wooldridge et Jennings, 1995, Russell et Norvig, 1995, Franklin et Graesser, 1996, Ferber, 1999, Floridi et Sanders, 2004] que nous résumons dans la table 1.1. Bien que toutes ces définitions diffèrent les unes des autres, un certain nombre d'éléments semblent récurrents. En effet, il est intéressant de remarquer que :

— cinq des six définitions [1] considèrent des entités artificielles (physiques ou virtuelles) ou biologiques faisant partie d'un environnement. En tant que tels, les agents sont alors des *entités finies* disposant de capacités limitées de perception et d'action au sein de cet environnement, qui leur fournit alors les conditions pour exister et qui leur sert de médium d'interaction [Weyns *et al.*, 2007] ;

— cinq des six définitions [2] font explicitement référence à la notion d'*autonomie* sans pour autant la définir clairement. Ceci laisse intuitivement penser qu'il s'agit d'un point central dans la caractérisation de ce qu'est un agent. Nous reviendrons donc sur cette notion dans la section 2.1 et montrerons qu'elle pose tout un ensemble de problématiques spécifiques qui ont guidé notre travail de recherche ;

— quatre des six définitions [3] font référence (parfois indirectement) à la notion de *but* que les agents doivent satisfaire, sans pour autant présumer de la manière dont ces buts ont été adoptés et comment ils peuvent être réalisés. En ce sens, un agent peut être conçu pour satisfaire les buts d'un concepteur ou d'un utilisateur, ou même des buts qui lui seraient propres s'il modélise une entité biologique. De même, différents agents peuvent être conçus selon des modèles de prise de décision distincts.

1. Toutes sauf [Wooldridge et Jennings, 1995].
2. Toutes sauf [Russell et Norvig, 1995].
3. Toutes sauf [Russell et Norvig, 1995, Floridi et Sanders, 2004].

Référence	Définition
[Shoham, 1993]	An agent is an entity whose state is viewed as consisting of *mental components* such as beliefs, capabilities, choices, and commitments, and that functions continuously and *autonomously* in an *environment* in which other processes take place and other agents exist.
[Russell et Norvig, 1995]	An agent is anything that can be viewed as perceiving its *environment* through sensors and acting upon that environment through effectors.
[Wooldridge et Jennings, 1995]	An agent is an *autonomous* rational entity which appears to be the subject of information attitudes and pro-attitudes such as beliefs, *desires*, commitments, etc.
[Franklin et Graesser, 1996]	An *autonomous* agent is a system situated within and a part of an *environment* that senses that environment and acts on it, over time, in pursuit of its *own agenda* and so as to effect what it senses in the future.
[Ferber, 1999]	An agent can be a physical or virtual entity that can act, perceive its *environment* (in a partial way) and communicate with others, is *autonomous* and has skills to achieve its *goals* and tendencies.
[Floridi et Sanders, 2004]	An agent is a system situated within and a part of an *environment*, which initiates a transformation, produces an effect or exerts power on it while having some interactivity, *autonomy* and adaptability properties.

TABLE 1.1 – Quelques définitions d'agents (les emphases sont les nôtres)

Propriété	Description
Situé	Ne s'exécute qu'en fonction des perceptions immédiates de l'agent
Incarné	Agit comme une réaction immédiate à l'environnement
Intelligent	Est une réponse adaptée à l'environnement
Émergent	Doit avoir un rôle au niveau global du point de vue d'un observateur
Minimaliste	Utilise le minimum de ressources ou d'informations
Sans état	Ne doit pas avoir d'état interne (ou mémoire)
Tolérant	Prend en compte l'incertitude ou l'incomplétude des perceptions

TABLE 1.2 – Propriétés des agents réactifs [Brooks, 1991, Werger, 1999]

Ainsi, nous définissons au plus haut niveau d'abstraction un agent comme :

Définition 1.1 (Agent)
Un agent *est une entité autonome finie existant dans un environnement et dotée de buts.*

Bien entendu, dans la littérature, les agents peuvent être caractérisés de manière plus précise par les descriptions internes (ou modèles) qui leur permettent d'agir. Ainsi, à très gros grain, un agent peut être défini comme *réactif* ou *délibératif*.

— Les **agents réactifs** procèdent des premiers travaux sur les architectures de subsomption [Brooks, 1986]. Ils correspondent aux agents à réflexes de [Russell et Norvig, 2003]. Leur caractéristique principale est de ne pas disposer d'un modèle du monde explicite et d'agir uniquement en fonction de leurs perceptions immédiates. Ainsi, un agent réactif peut être décrit comme un ensemble de comportements qui interagissent afin de produire un comportement général plus complexe. Selon [Brooks, 1991, Werger, 1999], ces comportements doivent satisfaire plusieurs propriétés résumées dans la table 1.2. La principale limite de ce type d'agent est la difficulté à formaliser le comportement global attendu et, donc, le risque d'obtenir un système sous-optimal au regard des buts à atteindre [Drogoul, 1995].

— Les **agents délibératifs**, quant à eux, procèdent des travaux de [Dennett, 1971]. Ils regroupent les agents à modèles, à buts, à utilité ou apprenants de [Russell et Norvig, 2003]. Ces agents sont caractérisés par le fait de disposer d'une description interne explicite, que cela représente l'environnement, les actions disponibles, les buts à satisfaire ou d'autres éléments plus complexes comme l'intention, l'engagement, la confiance, voire des modèles d'émotions. Ainsi, ils peuvent avoir des propriétés de *cognition*, de *raisonnement*, de *planification* et de *mémorisation* résumées dans la table 1.3. Ces propriétés regroupant de nombreux concepts, il en découle naturellement un bestaire d'architectures qui répondent à la définition des agents délibératifs comme par exemple les architectures BDI ou les processus décisionnels de Markov [Rao et Georgeff, 1991, Kaelbling *et al.*, 1998]. Les principales

Propriété	Description
Cognition	Représentation du monde en termes d'états mentaux.
Mémorisation	Utilisation des expériences passées lors de la décision.
Planification	Calcul d'actions à réaliser pour atteindre un but désiré.
Raisonnement	Représentation symbolique du monde

TABLE 1.3 – Propriétés des agents délibératifs (inspirées de [Wooldridge et Jennings, 1995])

limites de ce type d'agents sont la difficulté à leur fournir un modèle du monde correct et la complexité algorithmique souvent élevée pour calculer une décision.

Toutefois, la frontière entre ces deux classes d'agents n'est pas aussi distincte qu'elle semble l'être. En premier lieu, il a été montré qu'un ensemble d'agents réactifs pouvaient dans certains cas de figure simuler un agent cognitif [Shiloni *et al.*, 2009]. En second lieu, cela est dû, d'une part, à l'existence d'*agents hybrides* et, d'autre part, à la prise en compte d'*agents humains*.

— Comme leur nom l'indique intuitivement, les **agents hybrides** disposent tout à la fois d'un module réactif et d'un module cognitif qui interagissent au sein de la même architecture, comme par exemple les architectures Réactive Délibérative, InteRRap ou mêmes les Machines de Touring [Lemaître et Verfaillie, 2007, Rodriguez-Moreno *et al.*, 2007, Aschwanden *et al.*, 2006, Muller et Pischel, 1993, Ferguson, 1992]. De manière générale, le module réactif décide de la prochaine action à réaliser en fonction des perceptions immédiates de l'agent tandis que, parallèlement, l'agent calcule l'action à réaliser en fonction d'un modèle du monde. L'action à exécuter finalement est choisie en fonction du temps dont dispose l'agent pour décider.

— Le terme **agent humain** est fréquemment employé dans la littérature pour parler d'*utilisateur humain* ou d'*opérateur humain*. Un utilisateur humain est quelqu'un qui utilise les fonctions d'un agent artificiel sans connaître la manière dont il est conçu (par exemple un agent conversationnel sur un site web). Un opérateur humain est un professionnel qui interagit avec un agent artificiel dans le cadre d'une mission à effectuer conjointement (par exemple piloter un avion de ligne) [Mercier, 2011]. La spécificité d'un agent humain est d'être doté de pulsions, de sentiments, de buts inconscients ou de déficits d'attention qui peuvent parfois conduire à des décisions inattendues ou jugées irrationnelles.

1.2 Des systèmes d'agents autonomes

Dans la littérature, un système d'agents artificiels situés dans un environnement partagé est appelé un système multi-agent. Selon [Ferber, 1999], un *système multi-agent* est composé d'un environnement, d'objets et d'agents, de relations entre ces entités, d'actions qui peuvent être exécutées et de changements qui s'imposent à ce système dans le

temps. Toutefois, nous préférons à ce terme celui de *systèmes d'agents autonomes*. En effet, certains systèmes informatiques peuvent être vus comme plus que de simples systèmes mécaniques car ils impliquent la présence de nombreux agents humains (par exemple, les chat rooms, les sites de vente en ligne, les communautés virtuelles, etc.). Ces systèmes sont appelés *systèmes socio-techniques* et des instances d'entités sociales (agents humains) et techniques (agents artificiels) y interagissent en vue de réaliser un but commun [Hoc, 2000, Whitworth, 2006]. Les systèmes d'agents autonomes, dans leur généralité, nous semblent à la fois capturer les systèmes multi-agents et les systèmes socio-techniques.

Définition 1.2 (Système d'agents autonomes)
Un système d'agents autonomes *est un système multi-agent dans lequel au moins un agent artificiel interagit avec au moins un agent autonome, qu'il soit artificiel ou non.*

En fonction du type de système, les interactions entre agents peuvent être de différentes natures – *indirectes* si elles sont médiées par l'environnement ou *directes* si elles résultent de l'échange explicite de messages entre les agents – et peuvent participer à différents types d'activités. En particulier, les interactions servent à la *coordination* des agents entre eux. Ici, coordination a un sens plus large que la simple *synchronisation* : cela va de la planification d'actions au regard des plans des autres agents jusqu'à la *collaboration* – signifiant calculer des actions jointes pour réaliser des buts communs – ou la *négociation* – signifiant décider comment partager une ressource commune lorsque que les buts des agents sont différents [Durfee, 2001, Ferber, 1999, Nwana *et al.*, 1996].

Dans le contexte de notre travail de recherche, même si nous pouvons considérer n'importe quels types de systèmes d'agents humains qui interagissent les uns avec les autres, nous nous concentrons sur les systèmes d'agents autonomes. Cependant, la notion d'organisation manque à cette définition alors qu'il s'agit d'une dimension fondamentale où les interactions elles-mêmes prennent place [Boissier *et al.*, 2010].

1.3 Des organisations d'agents

Une organisation peut être définie selon deux points de vue :

1. une **entité collective** disposant d'une identité représentant un groupe d'agents doté de structures sociales formalisées [Lemaître et Excelente, 1998, Scott, 1998]. Cette entité n'est pas équivalente au groupe d'agents, peut être considérée comme un agent elle-même et existe en partie indépendamment des agents qui la composent ;
2. une **structure stable** d'activités jointes qui contraignent et affectent les actions et les interactions des agents [Castelfranchi, 1998, Hubner *et al.*, 2002, Sichman *et al.*, 2005]. Cette structure n'est pas un agent en soi et se transforme dynamiquement en fonction des agents qui la composent.

Comme en sociologie [Bernoux, 1985], l'organisation peut concerner la division des tâches, la distribution des rôles, les structures d'autorité, les systèmes de communication

ou même les systèmes de contribution et rétribution. Selon [Gasser, 2001], l'organisation peut aussi s'étendre à la notion de connaissance, culture, mémoire ou histoire.

Définition 1.3 (Organisation)
Une organisation *d'agents est une structure intentionnelle constituée de relations, de protocoles et de normes qui peut être spécifiée par le concepteur du système ou par les agents autonomes eux-mêmes.*

Selon [Horling et Lesser, 2004], trois formes d'organisations peuvent être distinguées, chacune ayant plusieurs variantes.

— Les **groupes** sont des organisations structurellement planes dans lesquelles un ensemble d'agents se synchronisent ou partagent des ressources. Les variantes des groupes comportent les *équipes*, les *coalitions* ou les *congrégations* [Sandholm *et al.*, 1999, Brooks et Durfee, 2003]. Alors que les coalitions sont des groupes temporaires, les équipes et les congrégations ont des cycles de vie long [Horling et Lesser, 2004]. Un groupe peut aussi être une *fédération* qui indique que, fonctionnellement, un agent du groupe joue le rôle de délégué pour permettre l'interaction avec les autres groupes.

— Les **hiérarchies** sont des structures arborescentes fondées sur le principe de « diviser pour régner » dans lesquelles le flot d'information est ascendant tandis que le flot de décision est descendant. De multiples types de hiérarchies existent en fonction de leur profondeur et des liens latéraux qu'elles contiennent. Par exemple, les *systèmes holoniques* sont des hiérarchies imbriquées où des groupes d'agents (appelés holons) sont structurés en de multiples hiérarchies [Fischer *et al.*, 2003].

— Les **sociétés** sont des organisations ouvertes, signifiant que des agents autonomes peuvent rejoindre ou quitter cette organisation au cours du temps [Buzing *et al.*, 2005]. Les sociétés permettent à des agents hétérogènes d'interagir à l'aide de protocoles de communication communs et de cadres déontiques comme des *systèmes normatifs*, des *places de marché* ou toute autre fonction sociale. Ainsi, les sociétés sont des organisations de haut niveau qui peuvent contenir d'autres organisations.

Enfin, nous pourrions nous demander si la présence d'agents humains dans un système d'agents autonomes induit des formes d'organisation spécifiques. L'analyse proposée par [Yanco et Drury, 2004] permet de penser ces organisations selon les types d'agents impliqués et leurs interactions : pour chaque type d'agents, sont-ils isolés, indépendants les uns des autres ou structurés en une sous-organisation ? Il en résulte 8 types de *sociétés* hommes-agents, illustrées sur la figure 1.1 où H représente un acteur humain et R (pour « robot ») un agent artificiel, avec pour chacune autant de variantes qu'il y a de sous-organisations possibles. Par exemple, dans le domaine aéronautique, le pilotage d'un avion est délégué à une société de type D : une *équipe* d'agents humains interagissant avec un unique agent artificiel, le pilote automatique. Remarquons qu'un neuvième type de société – mélangeant type C et E pour représenter plusieurs agents humains indépendants

interagissant avec plusieurs agents artificiels eux-aussi indépendants – n'est pas considéré. Si cela se justifie par le fait que, dans le cadre d'une application réelle, les humains sont toujours soit des opérateurs agissant au sein d'une organisation, soit des utilisateurs indépendants ayant affaire à une organisation d'agents artificiels construite dans un but précis, il est intéressant de constater qu'il s'agit de l'organisation la plus générale, la plus abstraite, celle qui est la plus tributaire de l'autonomie des agents.

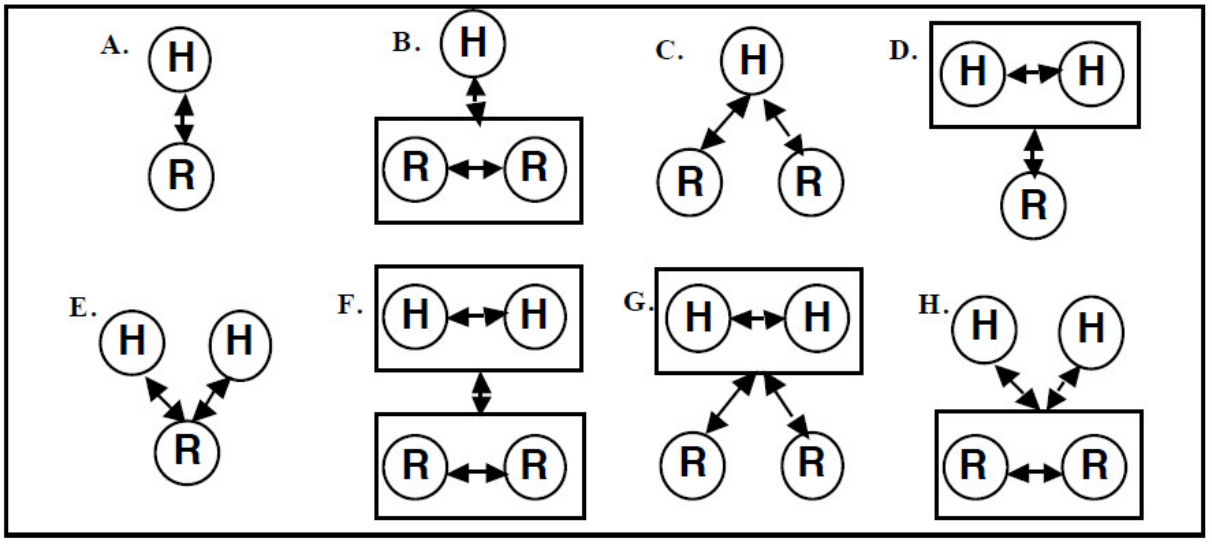

FIGURE 1.1 – Taxonomie des interactions homme-agents selon [Yanco et Drury, 2004]

2 L'autonomie en question

Si la capacité d'organisation est une propriété fondamentale d'un système d'agents autonomes, il n'en reste pas moins qu'elle est influencée par l'autonomie des agents. Intuitivement et sans avoir défini ce qu'était l'autonomie, il semble difficile à un agent n'ayant pas cette propriété d'être autre chose qu'une feuille au sein d'une hiérarchie. L'autonomie semble donc une propriété plus fondamentale encore, qu'il nous faut définir. Il convient alors en premier lieu de distinguer l'autonomie de l'automatisation.

2.1 Autonomie contre automatisation

Autonomie et automatisation diffèrent par la prédictibilité des actions, l'adaptation à l'environnement et la relation entretenue avec les agents humains. En effet, une machine automatisée réalise des séquences d'actions dont l'ordre est prédéterminé.

Selon [Truszkowski *et al.*, 2009], un *processus automatisé* émule un processus manuel en suivant une séquence d'actions étape par étape, pouvant éventuellement inclure la participation d'agents humains. Ainsi, en dehors des situations de panne, les actions d'une machine automatisée sont prévisibles et ne peuvent pas s'adapter à un état non prévu de l'environnement. Une telle machine doit donc opérer dans un environnement connu [Docherty, 2012]. Enfin, même si leurs fonctions peuvent nécessiter la participation d'agents humains, elles sont *conçues* pour donner des résultats prévisibles. Par exemple, une machine à laver exécute toujours les mêmes actions dans le même ordre en fonction de ses données environnementales. Dans le domaine spatial par exemple, les fonctions de calcul d'attitude [4] sont des processus automatisés : elles calculent les attitudes dès que les données stellaires sont disponibles, retournent un simple résultat aux autres fonctions du satellite et ne mettent en place aucune action de recouvrement si une erreur survient.

D'un autre côté, un agent autonome est capable d'opérer et de s'adapter aux environnements ouverts et non structurés. Ainsi, alors que l'objectif est le même – réaliser des tâches ou des actions sans intervention humaine – l'autonomie émule le comportement humain en calculant les actions de l'agent adaptées à l'environnement alors que l'automatisation exécute une série d'actions précalculée pour certains environnements [Jones, 2008, Truszkowski *et al.*, 2009]. Par exemple, un robot d'exploration doit pouvoir adapter son comportement à un terrain inconnu et réagir dynamiquement à ses perceptions comme identifier des zones d'intérêt. Dans le domaine spatial encore, le logiciel de vol d'un satellite surveille les données vitales de l'engin, identifie les détériorations et décide sans intervention de la station sol des actions à réaliser pour maintenir ces mesures dans le domaine requis afin de rester autonome.

2.2 Une notion d'autonomie relative

De nombreuses définitions de l'autonomie ont été proposées dans la littérature [Dorais *et al.*, 1999, Carabelea *et al.*, 2003, Castelfranchi et Falcone, 2003, Bekey, 2005, Jones, 2008, Truszkowski *et al.*, 2009, Defense Science Board, 2012] et nous en présentons quelques-unes dans la table 1.4.

Si les premiers travaux structurent l'autonomie en niveaux explicites [Sheridan et Verplank, 1978], il semble plus pertinent de décrire l'autonomie comme une notion *relative* : d'un point de vue externe – un agent est autonome du point de vue d'un autre pour une certaine fonction dans un certain contexte si son comportement n'est pas imposé par un autre agent [Carabelea *et al.*, 2003, Castelfranchi et Falcone, 2003] – et d'un point de vue interne – l'agent est capable de comportement autonome dans plusieurs situations différentes [Dorais *et al.*, 1999, Bekey, 2005, Jones, 2008]. Pour une même tâche, certaines fonctions nécessitent des interventions humaines tandis que d'autres peuvent être

4. Le contrôle d'attitude consiste à contrôler l'orientation du satellite dans l'espace et ses mouvements d'avant en arrière (tangage), de gauche à droite (roulis) et autour d'un axe vertical (lacet).

Référence	Définition
[Dorais *et al.*, 1999]	A system's level of autonomy can refer to : how complex the commands it executes are, how many of its sub-systems are being autonomously controlled, under what circumstances will the system override manual control, the duration of autonomous operation.
[Carabelea *et al.*, 2003]	An agent is not autonomous in an abstract sense but is autonomous on some level with respect to another entity, be it the environment, other agents, or even the developers of the agent.
[Castelfranchi et Falcone, 2003]	Autonomy is a relationship between the artificial agent and the human agent.
[Bekey, 2005]	Autonomy is the capacity of a robot to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time.
[Jones, 2008]	Autonomy is characterized in terms of [...] metric scores, including the percentage of a mission that is planned and executed by the [agent's] onboard processors, the levels of task decomposition, how easy it is to find a solution in the operating environment, etc.
[Truszkowski *et al.*, 2009]	Autonomy is a system's capacity to act according to its own goals, percepts, internal states, and knowledge, without outside intervention.
[Defense Science Board, 2012]	Autonomy is the capability (or a set of capabilities) enabling a particular action of a system to be automatic, or (within programmed boundaries) *self-governing*. It is not computers making independent decisions and taking uncontrolled action.

TABLE 1.4 – Quelques définitions de l'autonomie

autonomes. Ainsi, un système peut être dans plus d'un niveau d'autonomie à la fois et ces niveaux doivent être considérés en termes d'un continuum de *collaboration homme-machine*.

Définition 1.4 (Autonomie)
*L'*autonomie *est un* continuum dynamique *allant d'un contrôle humain complet sur toutes les décisions d'un agent artificiel aux situations dans lesquelles certaines fonctions – éventuellement de décision – peuvent être déléguées à l'agent artificiel avec seulement un haut niveau de supervision ou de surveillance.*

Ainsi donc, si l'autonomie est un continuum dynamique supervisé par un agent humain, nous pouvons nous demander comment sont décidés ou affectés ces niveaux d'autonomie [Hardin et Goodrich, 2009] et un bestiaire d'approches, que nous pouvons séparer en trois grandes catégories, se penche sur cette question.

— L'**autonomie adaptative** correspond au fait que l'agent artificiel a un contrôle exclusif sur sa propre autonomie, signifiant que l'agent humain ne peut prendre l'autorité – c'est-à-dire décider du contrôle sur l'agent artificiel – que sur demande de ce dernier (et donc sur des critères formalisés).

— L'**autonomie ajustable** correspond au fait que l'agent humain a un contrôle exclusif sur l'autonomie de l'agent artificiel, signifiant qu'il peut prendre l'autorité sur ce dernier en fonction de critères qui lui sont propres (et ne sont pas nécessairement formalisés).

— L'**autonomie en initiative mixte** correspond au fait que l'agent humain et l'agent artificiel peuvent tous les deux décider de l'autonomie de certaines fonctions dans certains contextes. Il s'agit donc d'une situation de partage d'autorité entre les deux agents.

Remarquons que des problèmes peuvent survenir lorsque les agents sont autonomes et que leurs actions sont non interruptibles [Orseau et Armstrong, 2016]. D'autres peuvent aussi survenir lorsque les agents emploient des données en provenance d'un humain qui a une connaissance insuffisante de la manière dont l'agent opère. Dans cette situation, l'autonomie ajustable permet de paramétrer dynamiquement les agents afin qu'ils ne soient pas trop autonomes, ni trop dépendants de l'humain [Beavers et Hexmoor, 2003].

2.3 De l'autonomie à la régulation

Quoi qu'il en soit, le fait de considérer des systèmes d'agents autonomes (à des degrés divers) a un certain nombre d'implications, tant au niveau organisationel qu'individuel, en termes de régulation.

En premier lieu, même si l'organisation, l'interaction et l'environnement ont pour objectif de fournir aux agents un cadre commun, il peut y avoir de l'**hétérogénéité** au

sein d'un système d'agents autonomes. Au regard des définitions données dans les sections précédentes, cette hétérogéité existe à plusieurs niveaux : architectural (agents humains et argents artificiels), décisionnel (agents réactifs et délibératifs) mais aussi opérationnel (les agents n'ont pas nécessairement les mêmes buts ou les mêmes préférences). Cette hétérogénéité est d'autant plus à prendre en considération dans le cadre des organisations de type sociétés ou coalitions qui ont une propriété d'**ouverture**. Il s'agit là-aussi d'une propriété importante qui concerne la *dynamique des entrées et sorties du système*, c'est-à-dire l'ajout ou le retrait au cours du temps de nouveaux agents autonomes (qu'ils soient humains ou artificiels) avec lesquels interagir.

En second lieu, le fait que les agents soient autonomes signifie qu'aucun d'entre eux ne dispose *a priori* de l'autorité sur les autres. Dans sa généralité, du point de vue du système, le processus de décision d'un agent individuel est inaccessible aux autres et peut être vu comme une boîte noire. Une fois les buts ou préférences des agents spécifiés, un concepteur, un opérateur, un utilisateur ou un autre agent artificiel n'a **pas de contrôle direct** sur les actions exécutées. À cela, s'ajoute le fait que, à la manière de la rationalité [Simon, 1990], l'autonomie est une **autonomie limitée** par les informations connues des agents, par le temps à leur disposition pour effectuer les calculs et par les limites inhérentes aux algorithmes. Ceci est d'autant plus important dans lorsqu'il existe une multitude de situations auxquelles les agents autonomes doivent faire face, et qui induisent parfois un manque de prédictabilité sur les actions exécutées.

En troisième lieu, l'introduction d'agents autonomes dans de nombreux domaines comme le milieu médical, militaire, judicaire ou, plus simplement, le transport autonome soulève des **problématiques éthiques**, en particulier dues à l'absence de contrôle direct. En effet, les humains interagissant avec ces agents délèguent, consciemment ou non, une partie de leur pouvoir de décision à ces machines. Or, les humains ont parfois des attentes éthiques distinctes des problématiques d'optimalité ou de conformité légale pour lesquels les agents sont conçus, comme en témoignent les rapports nationaux ou internationaux publiés à ce jour [CERNA, 2014, Demiaux et Si Abdallah, 2017, CERNA, 2017, IEEE, 2017].

Tout ceci conduit au fait que les agents autonomes ou les systèmes d'agents autonomes doivent être soumis à des mécanismes qui identifient et régulent les comportement des agents et des organisations ainsi que leurs interactions. Toutefois, en raison des propriétés présentées précédemment, ces mécanismes de régulation doivent prendre en compte certaines contraintes.

— L'**absence de contrôle direct** implique que la régulation du comportement des agents autonomes passe par le truchement de mécanismes externes comme des mécanismes d'incitation permettant d'adapter les buts des agents ou une délégation de la régulation aux autres agents de l'organisation (via des systèmes normatifs accompagnés de sanctions par exemple).

— L'**autonomie limitée** implique que la régulation du comportement des agents autonomes doit s'appuyer sur des principes généraux, exprimés et représentés le plus indépendamment possible des situations spécifiques que les agents peuvent rencontrer.

— L'**hétérogénéité** implique le fait que la régulation au sein d'une organisation du comportement des agents autonomes doit se faire en minimisant les hypothèses sur ces derniers, c'est-à-dire en les associant à des caractéristiques larges mais pertinentes pour le système, et ce sans préjuger de leurs capacités, ni de leurs buts.

— L'**ouverture** implique que la régulation des organisations elles-mêmes passe par la capacité de ces dernières à se réorganiser en fonction des entrées et des sorties des agents, et à mettre en œuvre des mécanismes d'identification des caractéristiques des agents qui, à un instant donné, en font partie.

— Les **problématiques éthiques** impliquent que les mécanismes de régulation ne doivent pas seulement se servir de critères d'optimalité ou de performance des agents mais aussi doivent être capables de prendre en compte des notions plus abstraites liées à la morale et à l'éthique. De plus, ces notions pouvant être spécifiques à une conception ou un groupe d'utilisateurs, il convient de permettre de réguler des systèmes d'agents autonomes dont les critères éthiques individuels sont hétérogènes.

C'est dans ce contexte de régulation que nos travaux se positionnent. Plus précisement, les propriétés évoquées ci-dessus pointent à la fois des méthodes – utilisation de mécanismes externes et définition de principes généraux – et des besoins d'identification des agents. Quelles sont alors ces caractéristiques à identifier, gérer et implémenter pour assurer le fonctionnement d'un système ou d'une organisation ? En nous plaçant du point de vue d'un agent humain interagissant avec un système d'agents autonomes, trois caractéristiques de haut niveau nous semblent fondamentales : les agents doivent faire preuve de *fiabilité*, d'*honnêteté* et, de manière plus générale, ils doivent faire preuve d'*éthique*.

3 Trois besoins fondamentaux

Ainsi, nous avons identifié comme étant des propriétés désirables de systèmes d'agents autonomes la fiabilité, l'honnêteté et le respect d'une éthique. Avant de formuler notre problématique de recherche, nous devons tout d'abord clarifier ces trois termes et mettre en lumière plusieurs notions – comme celles de confiance, manipulation, morale, valeur, etc. – qui semblent pertinentes pour étudier ces trois propriétés.

3.1 Fiabilité

La nécessité de garantir la correction d'un logiciel est un problème majeur, en particulier pour les systèmes critiques, c'est-à-dire les applications dédiées à des domaines où la sûreté de fonctionnement est nécessaire (comme les transports par exemple). C'est

pourquoi les agents autonomes, qui peuvent aisément être déployés pour ces applications, doivent faire preuve de fiabilité. La fiabilité peut être définie comme à la fois une valeur morale représentant la qualité d'un acteur à être digne de confiance et une valeur technique représentant la qualité d'un appareil à avoir un fonctionnement régulier et sûr, ce que nous retrouvons dans la définition de *reliability* du Oxford Dictionnary : *the quality of being trustworthy or of performing consistently well.* De manière intéressante, cette double définition est à rapprocher de deux grands types de méthodes d'évaluation de la fiabilité.

Une première approche est de **vérifier formellement** les agents. L'une des méthodologies permettant de faire de la vérification formelle consiste à prouver les spécifications des agents et à raffiner leurs comportements jusqu'à l'obtention d'un code exécutable, avec des preuves progressives. Ici, la fiabilité est vue comme le fait qu'un certain comportement, considéré comme désirable par un concepteur, est prouvé. Ces preuves peuvent alors être effectuées soit par des *model-checkers*, soit par des prouveurs de théorèmes qui peuvent ponctuellement faire appel à des *model-checkers*. Les *model-checkers* reposent sur un principe de test exhaustif, tandis que les prouveurs de théorème utilisent le calcul des séquents pour essayer, de manière heuristique, de générer des démonstrations. Notons que la plupart des travaux de vérification menés sur les agents utilisent le *model-checking* [Bordini *et al.*, 2003, Alechina *et al.*, 2004, Raimondi et Lomuscio, 2004, Kacprzak *et al.*, 2004]. Cependant, tous ces travaux partagent la même limite : l'explosion combinatoire des trajectoires possibles du système rend la preuve complexe, difficile, voire impossible. De plus, ces systèmes se réduisent d'ailleurs la plupart du temps à de la preuve sur des formules propositionnelles et non sur des formules de la logique du premier ordre. La principale raison est certainement dûe au fait que, la logique du premier ordre étant semi-décidable, les tentatives de preuves sont faites en utilisant des heuristiques et la preuve d'une propriété peut échouer [Stathis *et al.*, 2004, Bracciali *et al.*, 2006, Mermet et Simon, 2009]. Enfin, quelques travaux reposent sur la programmation logique [Martelli *et al.*, 1997, Giacomo *et al.*, 2000, Shapiro *et al.*, 2002, Baldoni *et al.*, 2005]. Toutefois, toutes ces approches reposent sur une hypothèse, qui est au cœur des méthodes de vérification formelle, à savoir que *la description du comportement interne des agents est accessible* à celui qui fait une vérification. Or, ceci entre en conflit avec les propriétés d'absence de contrôle, d'hétérogénéité des agents et d'ouverture des systèmes qui impliquent généralement de considérer que les descriptions internes des agents ne sont pas accessibles.

Une seconde approche consiste à caractériser une notion de **confiance entre agents**. La notion de confiance a été originellement introduite dans le contexte des systèmes d'agents autonomes par [Marsh, 1994]. Cette notion formalise une estimation du comportement futur d'un agent lorsqu'il existe un risque que celui-ci ait un comportement inattendu et repose sur trois axiomes fondamentaux mis en évidence par [Resnick *et al.*, 2000] :

1. les agents doivent interagir ensemble dans le futur ;
2. les agents doivent partager leur confiance par le biais de témoignages ;

3. ces témoignages doivent être utilisés par les agents pour décider avec qui interagir.

La confiance d'un agent envers un autre est donc une évaluation de ce dernier et plusieurs confiances d'agents envers un même tiers peuvent être agrégées via des témoignages pour construire une notion de réputation concernant ce tiers. Il existe de nombreuses définitions de la confiance dans les systèmes multi-agents [Stephen, 1994, Grandison et Sloman, 2000, Mui *et al.*, 2002, Xiu et Liu, 2005, Golbeck, 2006, Josang *et al.*, 2007] allant d'une mesure de la capacité d'un agent à accepter de dépendre d'un autre, à une estimation de la capacité d'un agent à réaliser de manière fiable une tâche dans un contexte donné, en passant par une mesure du risque pris en déléguant une tâche.

Plus récemment, [Castelfranchi et Falcone, 2010] ont fourni un important travail de synthèse sur la notion de confiance afin d'en caractériser les différents aspects, ce qui inclut les objets sur lesquels la confiance porte, sa dynamique, et la manière dont la confiance intervient dans la construction des décisions et des intentions. Ce qui ressort de ces travaux est que la confiance est une notion aux multiples facettes dont l'élément commun est que l'agent utilise le résultat d'interactions passées pour obtenir une estimation d'un comportement. Ainsi :

Définition 1.5 (Confiance)
La confiance est une estimation subjective d'un comportement futur d'un agent fondée sur l'historique des interactions passées.

La notion de réputation s'applique à des systèmes où les agents interagissent, collectent, partagent et agrègent les résultats de leurs interactions passées afin de décider à quels agents ils peuvent faire confiance pour de futures interactions. La réputation se fonde donc sur la confiance (l'inverse n'est pas vrai) et un agent (ou une autorité centrale) peut fusionner ses observations personnelles et des témoignages reçus afin de calculer des réputations à associer aux autres agents :

Définition 1.6 (Réputation)
La réputation *d'un agent est une agrégation des témoignages des autres agents envers lui, représentant une estimation collective de la confiance qu'un tiers pourrait avoir envers lui.*

Ces deux notions sont alors très pertinentes lorsqu'il s'agit de capturer celle de fiabilité. En effet, elle s'adapte mieux que les approches de vérification formelle au regard des contraintes d'autonomie des agents, dans le sens où elle fait *abstraction des descriptions internes des agents* (elle ne se fonde que sur l'observation de comportements), permettant ainsi de tenir compte de l'hétérogénéité et de l'ouverture des systèmes. Plus encore, la confiance peut se calculer au cours du fonctionnement du système et donc participer directement aux prises de décision des agents. De manière intéressante, ceci pose alors une nouvelle question : quelle influence en fonction de son usage l'évaluation de la fiabilité a-t-elle sur la fiabilité du système lui-même ?

3.2 Honnêteté

Si la caractérisation de la fiabilité est fondamentale pour la régulation des systèmes d'agents autonomes, il en est de même pour l'honnêteté. Dans sa généralité, l'honnêteté est une valeur morale qui représente la qualité d'un agent à agir conformément à une convention pour dire la vérité et faire ce qu'il se doit. Par exemple, le confusianisme décrit l'honnêteté comme à la fois la loyauté envers soi-même et les autres, et la fidélité à la parole donnée. Plus prosaïquement, le Oxford Dictionary définit l'honnêteté comme *free of deceipt, truthful and sincere*, c'est-à-dire exempt de manipulation, exempt de mensonge et sincère. Dans de nombreux domaines, il est désirable que les agents autonomes soient incités à dire la vérité, à révéler leurs informations et à ne pas manipuler le système : en théorie des choix sociaux [Gärdenfors, 1976], dans les systèmes d'enchères [Robinson, 1985], les systèmes de réputation [Schafer *et al.*, 1999] ou bien la sécurité des réseaux [Alpcan et Başar, 2010]. De manière intéressante, aborder l'honnêteté par le prisme de la manipulation nous permet aussi de considérer les questions de mensonge et de sincérité.

Si de nombreuses définitions de la manipulation existent [Gibbard, 1973, Ellison *et al.*, 1997], nous considérons la définition générale suivante.

Définition 1.7 (Manipulation)
Une manipulation est une stratégie permettant à un agent d'influencer et de contrôler les processus de décision d'un ensemble d'agents autonomes à l'aide de fausses informations afin que ces derniers prennent une décision qui lui soit favorable.

Un agent autonome peut manipuler pour deux raisons. La première est d'utiliser stratégiquement le système pour ce pour quoi il est conçu afin d'augmenter son gain indépendamment du gain des autres agents. Nous parlons alors d'*agents malhonnêtes*. Par exemple, faire croire qu'un message est prioritaire pour accéder plus rapidement à la bande passante sur un réseau. La seconde raison est de perturber le fonctionnement du système, c'est-à-dire l'empêcher de réaliser les fonctions pour lesquelles il a été conçu. Nous parlons alors d'*agents malveillants*. Par exemple, supprimer arbitrairement des messages qui transitent dans un réseau pour faire croire à une défaillance de l'agent émetteur.

Définition 1.8 (Agent malhonnête)
Un agent malhonnête *est un agent qui manipule un système afin de maximiser son gain, indépendamment du gain obtenu par les autres agents.*

Définition 1.9 (Agent malveillant)
Un agent malveillant *est un agent qui manipule un système afin de minimiser le gain d'un sous-ensemble d'agents tiers.*

Si cette distinction est à rapprocher de celle introduite par [Conitzer *et al.*, 2003] dans les systèmes de vote, il n'existe toutefois pas de frontière stricte entre agents malhonnêtes et agents malveillants. En effet, la malhonnêteté d'un agent induit généralement des baisses

de gains pour les autres agents et, dans ce cas, être malhonnête suffit à être malveillant. Inversement, si nous représentons le gain d'un agent malveillant par l'opposé de celui des autres agents alors être malveillant implique d'être malhonnête. Remarquons enfin que les agents manipulateurs peuvent aussi faire partie d'une organisation – comme des coalitions – en se regroupant autour du même objectif afin d'avoir une influence plus importante [Robinson, 1985].

Il existe de nombreux types de manipulations [Douceur, 2002, Chang, 2002, Bachrach et Elkind, 2008, Bilge *et al.*, 2009, Hoffman *et al.*, 2009, Waggoner *et al.*, 2012]. La figure 1.2 présente une taxonomie (non exhaustive au niveau des feuilles) des différentes manipulations dans les systèmes d'agents autonomes et nous invitons le lecteur à se référer à l'état de l'art de [Vallée, 2015] pour plus de détails.

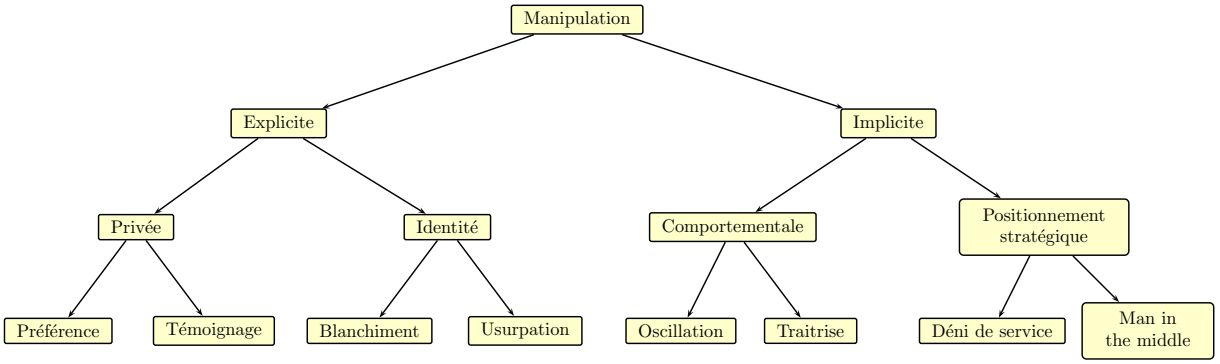

FIGURE 1.2 – Taxonomie des manipulations selon [Vallée, 2015]

Comme les processus de décision des agents sont fondés en partie sur leurs connaissances, manipuler un agent consiste à biaiser ses connaissances et ainsi contrôler indirectement ce processus. Pour ce faire, l'agent manipulateur peut soit partager *explicitement* avec sa cible des informations qu'il sait être fausses, soit *implicitement* l'amener à déduire de fausses connaissances.

Une **manipulation explicite** consiste à fournir volontairement à un agent tiers de fausses informations. En effet, les agents n'ayant pas une perception complète et parfaite de l'environnement, l'échange d'une partie de leurs informations vient renforcer mutuellement leurs connaissances [Stone et Veloso, 2000]. Or, l'apport de ce partage d'informations repose toujours sur la véracité de ces dernières. Nous pouvons alors distinguer deux catégories d'informations : les *informations privées* d'un agent, c'est-à-dire sa représentation interne de l'environnement et les *informations publiques*, c'est-à-dire l'ensemble des connaissances observables par tous les agents du système. Par exemple dans le domaine du choix social, fournir de *faux profils de préférence* est une manipulation sur les informations privées tout comme l'est le *faux témoignage* dans les systèmes de réputation. Les manipulations fondées sur l'identité des agents – usurpation d'identité [Koops et Leenes,

2006, Bilge *et al.*, 2009, Angin *et al.*, 2010], blanchiment [Feldman *et al.*, 2006] ou attaques Sybil [Douceur, 2002, Cheng et Friedman, 2005a, Bachrach et Elkind, 2008, Aziz et Paterson, 2009, Waggoner *et al.*, 2012] – sont des manipulations sur les informations publiques. Par exemple, les *attaques Sybil* consistent à utiliser de fausses ou multiples identités alors que les systèmes d'agents autonomes font généralement l'hypothèse que les agents ne disposent que d'une et une seule identité, pour obtenir des données privées, réinitialiser des mécanismes d'apprentissage ou construire des coalitions virtuelles.

À l'inverse des manipulations explicites, une **manipulation implicite** consiste à interagir avec le système afin que les autres agents déduisent de leurs observations de fausses connaissances. Nous distinguons alors les *manipulations comportementales* lorsque l'agent manipulateur fournit de fausses informations par l'intermédiaire d'un comportement observable particulier et les manipulations par *positionnement stratégique* lorsque l'agent manipulateur va agir afin de réduire la capacité d'observation des autres agents. Par exemple, dans le cadre des jeux répétés ou des systèmes de réputation, les agents estiment le comportement futur des autres à l'aide de leurs observations lors d'interactions passées. Une *traîtrise* est une manipulation comportementale qui consiste à adopter un comportement fiable pendant une période de temps afin d'être identifié en tant que tel puis subitement adopter un comportement non fiable, mais associé à un gain important [Marti et Garcia-Molina, 2006]. Dans les réseaux, une *attaque éclipse* est une manipulation par positionnement stratégique qui consiste à intercepter et supprimer des messages afin d'isoler du réseau un sous-ensemble des agents de manière que ces derniers ne puissent pas interagir avec les autres agents du système [Specht et Lee, 2004, Singh *et al.*, 2006].

3.3 Éthique

En préambule des deux sections précédentes, nous avons introduit les notions de fiabilité et d'honnêteté comme des valeurs morales ou techniques. Toutefois, nous avons indiqué en section 2.3 que l'autonomie des agents peut soulever des questions d'ordre éthique. Ainsi, selon les applications, les utilisateurs ont parfois des attentes éthiques distinctes des problématiques d'optimalité ou de conformité légale du comportement des agents. Une manière de gérer ces problématiques est de concevoir des agents autonomes capables d'exhiber des comportements qui pourraient être qualifiés d'éthiques, ce qui implique de caractériser des valeurs éthiques et morales et de permettre à des agents autonomes de raisonner et décider en fonction de ces dernières.

Si les deux mots *morale* et *éthique* désignent initialement la même chose, l'usage a introduit une différence entre les deux notions selon leur champ d'application prétendu (universel pour la morale, particulier pour l'éthique), leur statut (absolu ou relatif) et leur visée (le bon ou le juste) [Ricoeur, 1995, Comte-Sponville, 2012, Timmons, 2012]. Nous renvoyons le lecteur aux états de l'art présentés dans [Cointe, 2017, Boissier *et al.*, 2017] pour de plus amples précisions. Toutefois, clarifions quelques points permettant de distinguer les théories morales des théories éthiques.

1. L'**objectivisme** moral affirme, d'une part, qu'il existe des valeurs indépendantes de nos désirs ou préférences et, d'autre part, que dans l'ensemble des énoncés évaluatifs moraux possibles, certains sont vrais et d'autres faux. À l'inverse, le **particularisme** éthique pose la question des implications pratiques des jugements moraux mais pas de la vérité ou de la fausseté de ces derniers.
2. L'**absolutisme** moral s'oppose au **relativisme** éthique dans le sens où le relativisme affirme la relativité de toute valeur et donc de toute évaluation. Ainsi, une éthique est relative à un certain sujet, une certaine histoire, une certaine culture, à un certain désir, voire à tout cela à la fois.
3. Le **bon** au sens moral affirme qu'une chose est positive en raison d'une description en termes de valeurs tandis que le **juste** au sens éthique affirme qu'une chose doit être choisie au regard d'une procédure décrivant comment les valeurs doivent être employées.

Ainsi, le juste s'appuie sur le bon et donc l'éthique s'appuie sur la morale. Dans tous les cas, la notion de *valeur* reste au cœur de cet édifice. En nous appuyant sur les travaux en axiologie, psychologie et sociologie [Rokeach, 1973, B. Williams, 1990, Swchartz, 1992, Hitlin et Piliavin, 2004, Schwartz, 2012, Schroeder, 2016], nous définissons une valeur au regard de plusieurs critères.

— **Valeurs intrinsèques et extrinsèques**. Une distinction courante s'appuie sur la différence entre valeur finale, attributive ou *intrinsèque* et valeur instrumentale, prédicative ou *extrinsèque* d'une chose. Une chose a une valeur intrinsèque si elle possède cette valeur en elle-même, indépendamment des autres choses. Si elle était seule à exister, elle possèderait encore cette valeur. Par exemple, la dignité est une valeur intrinsèque et le beau est extrinsèque.

— **Concepts épais et fins**. Une deuxième distinction oppose les valeurs spécifiques ou *concepts épais* aux valeurs générales ou *concepts fins* [B. Williams, 1990]. Les premières sont des valeurs reposant principalement sur une description alors que les secondes reposent principalement sur des jugements. Par exemple, la sincérité au sens de dire ce que nous croyons être vrai est un concept épais tandis que le juste est un concept fin.

— **Systèmes de valeurs**. Les valeurs semblent exister en nombre fini, et être présentes dans toute culture, en variant seulement en importance [Swchartz, 1992]. Elles sont organisées au sein d'un *système de valeurs*, un ensemble de valeurs structuré par des relations hiérarchiques et des relations d'opposition auxquelles des agents accordent des importances plus ou moins grandes.

Quoi qu'il en soit, nous parlons d'agents autonomes *moraux* (ou éthiquement neutres) si leur comportement satisfait simplement des valeurs, et d'agents autonomes *éthiques*[5]

5. Cette expression est un raccourci langagier. Un agent autonome ne peut pas être éthique mais simplement exhiber des comportements qui peuvent être qualifiés d'éthiques par un observateur humain.

si leur processus de décision réalise un arbitrage entre des valeurs et leurs buts. C'est ce second type d'agent qui nous intéresse car rares sont les applications qui n'ont pour objectif que de simplement satisfaire des valeurs morales : il est bien souvent nécessaire de composer avec les intérêts du domaine d'application.

Définition 1.10 (Agent autonome éthique)
Un agent autononome éthique *est un agent dont le processus de décision intègre de manière explicite des valeurs ainsi qu'un arbitrage entre ces valeurs et les buts de l'agent dans son domaine d'application.*

La littérature traite de la question de ces agents éthiques selon six angles principaux.

— La **conception sensible aux valeurs** [Friedman, 1996, Friedman *et al.*, 2013, Aldewereld *et al.*, 2015] est une méthodologie de génie logiciel permettant de prendre en compte des valeurs morales. Pour cela, des experts ont pour objectif de guider l'implémentation afin d'obtenir un logiciel dont le comportement est conforme aux attentes éthique du domaine. La littérature fournit un grand nombre d'heuristiques, de précisions méthodologiques, de valeurs à examiner et de cas d'exemples, mais cela ne participe pas explicitement aux processus de décision des agents autonomes.

— L'**aide à la décision éthique** propose des implémentations permettant à des utilisateurs humains d'analyser des problématiques éthiques. Certaines approches s'appuient sur des langages de modélisation pour représenter un processus de décision [Frize *et al.*, 2005, Okada *et al.*, 2007, Chatterjee *et al.*, 2009], d'autres proposent des mécanismes pour éliciter des critères éthiques [Chae *et al.*, 2005, Anderson *et al.*, 2006, Mathieson, 2007, Robbins et Wallace, 2007] mais aucune ne propose des mécanismes de résolution.

— Le **raisonnement moral et éthique** est souvent modélisé par des logiques déontiques [Chisholm, 1963, Gensler, 1996, Powers, 2005, Bringsjord et Taylors, 2012, Lorini, 2012] ou des logiques non-monotones [Ganascia, 2007, Ganascia, 2012, Berreby *et al.*, 2015]. Ces travaux s'attachent en particulier à exprimer des théories spécifiques (raisonnement aristotélicien, raisonnement kantien, modélisation de la responsabilité morale, etc.). Une des approches les plus abstraites s'appuie sur les systèmes d'argumentation valués décrivant des valeurs associées à des arguments, et des préférences sur ces valeurs [Atkinson et Bench-Capon, 2008, Bench-Capon et Atkinson, 2009, Atkinson et Bench-Capon, 2016].

— L'**éthique empirique**, de son côté, s'appuie sur des mécanismes d'apprentissage pour fonder une théorie en accord avec des observations données par des experts [Braithwaite, 1955, Honarvar et Ghasem-Aghaee, 2009, Anderson et Anderson, 2014, Abel *et al.*, 2016, Anderson *et al.*, 2017]. Dans ces approches, le caractère éthique

Comme le remarquaient, non sans humour, [Turkle et Shapiro, 2011] : il ne faut pas confondre simuler l'amour et l'amour lui-même.

d'une action est généralement représenté par des poids qui sont aggrégés puis comparés à un critère de décision (généralement une maximisation de l'utilité). Ainsi, ces travaux combinent des fonctions d'utilité données *a priori* et des requêtes auprès d'un utilisateur humain.

— Quelques travaux s'intéressent à la **vérification formelle de l'éthique** [Abramson et Pike, 2011, Winfield *et al.*, 2014, Dennis *et al.*, 2015]. Toutefois ces approches présentent encore des limites car il s'agit pour l'essentiel de vérifier qu'un agent satisfait des règles portant sur des choix ponctuels d'action, et non des règles correspondant à des aspects plus généraux de leur comportement.

— Enfin, certains travaux proposent des **implémentations d'agents éthiques** [McLaren, 2003, Arkin, 2009, Honarvar et Ghasem-Aghaee, 2009, Saptawijaya et Pereira, 2014]. Par exemple, [Saptawijaya et Pereira, 2014] s'appuie sur de la programmation logique, [Arkin, 2009] sur des contraintes architecturales et [Honarvar et Ghasem-Aghaee, 2009] sur un réseau de neurones. Beaucoup d'implémentations s'appuient aussi sur des agents normatifs [Guerini et Stock, 2005, Caire, 2009, Chopra et White, 2011, Tufis et Ganascia, 2012].

Dans la littérature en éthique computationnelle, la plupart des travaux s'intéressent uniquement à l'éthique à l'échelle du comportement individuel de l'agent. Or, dans un système d'agents autonomes, une simple contrainte de son comportement peut permettre à un agent d'agir individuellement de manière éthique dans un collectif, mais le laisse démuni lorsqu'il doit tenir compte de l'éthique des autres agents. De plus, outre des variations dans l'éthique individuelle, différentes éthiques coexistent au sein d'une même société, parfois même au sein d'un même individu. Dans ces circonstances, toute approche de mise en œuvre du raisonnement éthique dans des agents autonomes doit prendre en considération cette dimension plurielle des éthiques.

4 Questionnement central

Au vu des éléments présentés précédemment, notre projet de recherche consiste à *étudier la notion de fiabilité, d'honnêteté et de respect de l'éthique dans les systèmes d'agents autonomes*. Ce projet se structure autour de neuf questions, notées de Q1 à Q9 dans la suite :

1. Comment caractériser qualitativement (Q1) et quantitativement (Q2) la fiabilité d'un agent et quelle influence cette caractérisation a-t-elle sur un système d'agents autonomes en fonction de la manière dont les agents en question s'en servent (Q3) ?
2. Comment caractériser qualitativement (Q4) et quantitativement (Q5) l'honnêteté d'un agent et quel mode d'organisation des agents permet de garantir le respect de cette valeur (Q6) ?
3. Comment représenter et raisonner sur des valeurs morales et éthiques ou modéliser des principes ou des théories éthiques issus de la philosophie (Q7) ainsi que vérifier

que des agents respectent ces valeurs et principes (Q8) et puissent interagir avec des agents aux éthiques et morales différentes (Q9) ?

Afin de répondre à ces questions, notre projet de recherche s'appuie sur des approches formelles – les systèmes de confiance et réputation, les jeux de coalitions et les modèles d'agents cognitifs – qui les traitent de manière croisée, ce que nous présentons au chapitre suivant.

Chapitre 2

Modéliser la fiabilité, l'honnêteté et l'éthique

Sommaire

Répondre à nos questions de recherche ne peut se faire sans considérer des modèles formels permettant de représenter les notions de fiabilité, d'honnêteté et d'éthique évoquées au chapitre précédent. Ce chapitre a donc pour objectif de dégager des questions concrètes en croisantr ces notions avec différentes approches formelles : les systèmes de réputation, les modèles de formation de coalition et les modèles d'agents cognitifs. Avoir choisi ces approches nous permet de considérer une large game de modèles, entre modèles individuels et modèles de collectifs, et entre modèles quantitatifs et modèles qualitatifs. Nous présentons dans un premier temps pour chaque approche un état de l'art orienté (et donc volontairement partiel car ce n'est pas l'objet du présent mémoire) autour des questions de fiabilité, d'honnêteté et d'éthique. Dans un second temps, nous positionnons nos neuf questions de recherche par rapport à ces modèles.

1 Systèmes de réputation

Cette section a pour objectif de présenter une première approche formelle pour étudier les systèmes d'agents autonomes : l'usage de systèmes de réputation. Nous montrons ici que les notions de confiance et de réputation classiquement utilisées dans ces systèmes permettent de traiter la question de fiabilité mais qu'elles peuvent être adaptées pour traiter celle de l'honnêteté via la notion de crédibilité.

1.1 Approches quantitatives contre qualitatives

Comme indiqué au chapitre précédent, la notion de confiance formalise une estimation du comportement futur d'un agent lorsqu'il existe un risque que celui-ci ait un comportement inattendu. Cette confiance peut être représentée soit de manière quantitative, soit qualitative. Dans tous les cas, elle se fonde sur l'observation d'interactions interpersonnelles et elle peut être agrégée au niveau collectif en une notion de réputation, c'est-à-dire la représentation d'un consensus au niveau de l'ensemble des agents. C'est pourquoi les travaux sur la confiance peuvent se diviser en deux classes :

— des approches qualitatives qui, en s'appuyant sur des modèles logiques, sont intéressantes pour représenter l'intentionalité d'un agent ;

— des approches quantitatives qui, en comptant et agrégeant les interactions, sont intéressantes pour représenter la dynamique de la confiance.

Ces deux approches ne s'excluent pas mutuellement et sont pertinentes pour penser les notions de fiabilité, d'honnêteté et d'éthique. Toutefois, l'approche quantitative – en s'intéressant surtout aux protocoles d'interaction entre agents – permet de faire l'économie de leurs modèles internes et donc d'être plus à même de tenir compte de la propriété d'autonomie des agents. C'est pourquoi nous traitons des approches qualitatives en section 3.2 tandis que nous nous intéressons ici aux approches quantitatives.

Ces approches quantitatives s'appuient sur des **systèmes de réputation** où les agents interagissent, collectent, partagent et agrègent les résultats de leurs interactions passées afin de décider à quels agents ils peuvent faire confiance pour de futures interactions. La confiance et la réputation peuvent alors prendre la forme d'une *valeur booléenne* [Stephen, 1994], *discrète* comme très mauvais, mauvais, neutre, bon, très bon [Abdul-Rahman et Hailes, 2000] ou *continue*[1] [Stephen, 1994, Kamvar *et al.*, 2003, Cheng et Friedman, 2005a, Jøsang *et al.*, 2007, Zhou et Hwang, 2007]. Ces systèmes de réputation se déclinent ensuite en trois grandes familles selon leur type de fonction d'agrégation, c'est-à-dire l'algorithme utilisé par les agents (ou une autorité centrale) pour calculer la réputation.

1. Dans le cas continu, la réputation peut avoir une sémantique de rang ou de valeur. Un rang de réputation permet d'ordonner qualitativement les agents entre eux : un agent ayant une plus grande réputation qu'un autre est considéré comme plus fiable. Une valeur de réputation permet non seulement d'ordonner les agents mais aussi d'associer un sens quantitatif à ces derniers comme la probabilité que leur prochaine interaction soit de bonne qualité.

— Les **systèmes symétriques** sont des systèmes dans lesquels l'ordre d'agrégation des témoignages n'influe pas sur la valeur de réputation des agents. C'est le cas de fonctions d'agrégation naïves qui peuvent consister à faire la moyenne des témoignages comme sur eBay par exemple.

— Les **systèmes asymétriques globaux** sont des systèmes dans lesquels l'ordre d'agrégation est structuré. Les confiances des agents sont représentées sous forme d'un graphe orienté valué G – appelé *graphe de confiance* – où les nœuds désignent les agents, les arcs des interactions passés et leurs poids sont la valeur de confiance. Un système est asymétrique si au moins un nœud du graphe a une importance privilégiée dans le calcul de la réputation, généralement représentée par une confiance *a priori* dans l'agent associé à ce nœud. Le système est *global* si, malgré cette propriété d'asymétrie, la valeur de réputation ne dépend pas de l'agent qui la calcule – comme dans le cas du système EigenTrust [Kamvar *et al.*, 2003].

— Les **systèmes asymétriques personnalisés** sont des systèmes dans lesquels non seulement l'ordre d'agrégation des témoignages est structuré comme précédemment mais aussi dans lesquels la valeur de réputation d'un agent dépend de celui qui la calcule – comme dans le cas des systèmes BetaReputation [Jøsang et Ismail, 2002] et FlowTrust [Cheng et Friedman, 2005a].

À titre d'exemple, nous détaillons ci-dessous les trois derniers systèmes de réputation cités qui sont couramment étudiés dans la littérature. Nous ne présentons ici aucun système symétrique car leur fonctionnement est généralement trivial (comme lorsqu'il s'agit de faire la moyenne des confiances accordées par les agents afin de calculer une valeur de réputation).

— EigenTrust est un système de réputation global asymétrique inspiré du Google PageRank [Page *et al.*, 1999] qui utilise la matrice d'adjacente a_3 du graphe de confiance et produit un rang de réputation. La confiance y est modélisée par la somme des interactions satisfaisantes, notée $sat(i,j)$, diminuée de la somme des interactions non satisfaisantes, notée $unsat(i,j)$. Cette valeur de confiance, notée $s_{ij} \in \mathbb{Z}$, est ensuite normalisée en une valeur $c_{i,j} \in [0,1]$. Il s'agit donc d'une répartition d'un poids entre tous les agents. De plus, EigenTrust attribue sous forme d'un vecteur $\vec{p}$ une valeur minimale par défaut à des agents de confiance. Pour un paramètre d'exploration $a \in [0,1]$, la réputation sous forme d'un vecteur $\vec{t}$ des agents est la probabilité qu'un marcheur aléatoire sur le graphe de confiance partant de l'agent a_i s'arrête sur l'agent a_j. C'est le point fixe de la fonction ci-dessous lorsque k est incrémenté :

$$\vec{t}^{k+1} = (1-a)C^T\vec{t}^{k} + a\vec{p}$$

— BetaReputation est un système de réputation fondé sur une approche bayésienne et produisant une valeur de réputation continue. La confiance est modélisée par un couple $\langle r_{ij}, s_{ij} \rangle$ correspondant respectivement à la partie positive et négative de

l'évaluation d'un agent a_i des interactions qu'il a eues avec un agent a_j. Ces deux valeurs doivent appartenir à un même domaine de définition fini et doivent correspondre au gain réel positif et négatif obtenu lors d'une interaction. La réputation d'un agent est alors modélisée par une fonction de densité Beta. Sémantiquement, la réputation correspond à la valeur espérée de la qualité d'une future interaction avec cet agent. Lorsque l'agent a_i reçoit un témoignage $\langle r_{jk}, s_{jk} \rangle$ de l'agent a_j vis-à-vis de l'agent a_k, il l'agrège avec ses propres observations comme suit :

$$r_k^{i:j} = \frac{2 r_{ij} r_{jk}}{(s_{ij}+2)(r_{jk}+s_{jk}+2)+2r_{ij}}$$
$$s_k^{i:j} = \frac{2 r_{ij} s_{jk}}{(s_{ij}+2)(r_{jk}+s_{jk}+2)+2r_{ij}}$$

Intuitivement, lorsque l'agent a_i reçoit un témoignage provenant de l'agent a_j, le témoignage est pondéré par la confiance que a_i a envers l'agent a_j. De ce fait, BetaReputation utilise une fonction de réputation personnalisée. En effet, la réputation de a_k (notée $Rep(r_k, s_k)$) est calculée à partir de l'agrégation de l'ensemble des témoignages reçus par la fonction :

$$Rep(r_k, s_k) = \frac{r_k - s_k}{r_k + s_k + 2}$$

Des extensions de BetaReputation proposent d'intégrer un facteur d'oubli $\lambda \in [0, 1]$ pondérant l'importance des interactions les plus anciennes ou une troisième composante de la confiance u_{ij} représentant un degré d'incertitude lors de l'évaluation des interactions.

— FlowTrust est un système de réputation asymétrique personnalisé se fondant sur le graphe de confiance. La confiance c_{ij} est une valeur réelle unique représentant la proportion d'interactions satisfaisantes de j pour i. La réputation de a_k est le flot maximal pour l'ensemble $\mathbb{P}_{i,k}$ de chemins disjoints allant de a_i vers a_k sur le graphe de confiance G :

$$f(G,k)_i = \max_{\mathcal{P}_{i,k} \in \mathbb{P}_{i,k}} \sum_{P \in \mathcal{P}_{i,k}} \min\{c_{xy} | (x,y) \in P\}$$

où les P est un chemin de l'ensemble $\mathcal{P}_{i,k}$ parmi tous les ensembles de chemins disjoints possibles $\mathbb{P}_{i,k}$. Il est à noter que [Cheng et Friedman, 2005a] ont montré que remplacer l'opérateur $\sum$ par max garantit une robustesse aux diffamations et à certaines formes d'attaques Sybil.

1.2 Honnêteté et crédibilité

S'arrêter à ce que nous avons décrit ci-dessus revient toutefois à resteindre drastiquement la notion de confiance à un aspect particulier. En effet, [Castelfranchi et Falcone,

2010] ont étudié la hiérarchie des différents composants fondamentaux de la confiance mais aussi ses aspects dynamiques, en particulier dans le cadre de la décision, de la construction d'intentions, de l'acte de faire confiance ou de s'autoriser à déléguer des actions, et il en ressort que la confiance est une notion plurielle aux multiples facettes. Modéliser la confiance dans sa globalité semble alors peu pertinent car produit une trop grande abstraction. C'est pourquoi, à y regarder de plus près, les approches présentées précédemment s'intéressent principalement à la confiance et la réputation dans les actions des agents.

Plus précisement, tous ces systèmes modélisent une notion de **confiance en la fiabilité** d'un agent. En effet, les systèmes de réputation s'intéressent principalement à l'estimation de la fiabilité des agents même si certains systèmes de réputation modélisent la confiance et la réputation comme des **tuples de valeurs** représentant différents critères d'évaluation [Sabater et Sierra, 2001, Jøsang et Ismail, 2002, Theodorakopoulos et Baras, 2006, Jin *et al.*, 2007] ou différents contextes d'application [Page *et al.*, 1999, Sabater *et al.*, 2006, Danezis et Mittal, 2009].

Toutefois, certains travaux se sont intéressés à l'évaluation de l'honnêteté au travers de la notion de **crédibilité**. Cette notion de crédibilité part du constat que, dans de nombreux systèmes de réputation [Mui *et al.*, 2002, Kamvar *et al.*, 2003, Cheng et Friedman, 2005a, Yu *et al.*, 2006], les témoignages d'un agent a_j sont pondérés par la confiance que l'agent a_i a envers a_j. Ainsi, la confiance joue un double rôle : mesurer la fiabilité de l'agent a_j lors de ses interactions et mesurer sa fiabilité lorsqu'il communique un témoignage. Or, il s'agit de deux notions différentes et faire l'hypothèse d'un transfert de l'une à l'autre est problématique car, par exemple dans le cadre d'un système de service, un agent dont les services sont fiables pourra formuler des témoignages malhonnêtes pour évincer des concurrents.

La notion de crédibilité consiste alors à définir une mesure de confiance spécifique à la production de témoignages. Comme la confiance, la crédibilité vient ensuite affecter l'agrégation des témoignages, soit en pondérant les témoignages [Sabater et Sierra, 2001, Srivatsa *et al.*, 2005, Koutrouli et Tsalgatidou, 2011], soit en les filtrant et les retirant [2] directement du processus d'agrégation [Muller et Vercouter, 2004, Whitby *et al.*, 2004, Zhao et Li, 2009]. La crédibilité peut prendre plusieurs formes :

— un degré de similarité entre le témoignage reçu et soit les observations directes de l'agent [Sabater et Sierra, 2001, Srivatsa *et al.*, 2005, Koutrouli et Tsalgatidou, 2011], soit le résultat de l'interaction suivante [Zhao et Li, 2009] ;

— une quantité d'inconsistance entre les témoignages reçus par plusieurs agents [Muller et Vercouter, 2004, Muller et Vercouter, 2005] ;

— le gain d'information produit par le témoignage reçu [Whitby *et al.*, 2004].

2. Remarquons que certaines techniques de filtrage sont drastiques en mettant sur liste noire tous les témoignages d'un agent non crédible.

Ainsi, nous pouvons tracer une analogie entre la confiance des systèmes de réputation qui est une *confiance en la fiabilité des actions* au sens d'une confiance dispositionnelle dans le contexte où elle s'exprime, et la crédibilité qui est une *confiance en l'honnêteté des témoignages*. Remarquons que, dans ce dernier cas, nous aurions pu penser qu'il s'agissait d'une confiance en la fiabilité des informations. Cependant, il est important de pouvoir faire confiance à des sources incohérentes entre elles car cela permet d'obtenir de l'information sur la variance du comportement des agents. Ainsi, au-delà de l'usage classique de la confiance pour évaluer la fiabilité des agents, il nous semble pertinent d'aborder la question de l'honnêteté par le prisme de la crédibilité.

1.3 De l'influence du processus de décision sur la confiance

Nous pouvons remarquer que les travaux présentés précédemment portent essentiellement sur le calcul des valeurs de confiance, de réputation et de crédibilité mais non pas sur la manière dont les agents les utilisent. Or, la politique d'utilisation de ces valeurs a nécessairement une influence sur le fonctionnement du système tant sur ce pour quoi il a été conçu que sur la manière dont les valeurs sont calculées. En effet, dans le premier cas par exemple, si l'ensemble des agents décide de n'interagir qu'auprès des agents ayant la plus haute valeur de réputation, il sera difficile pour un agent malveillant seul d'interagir. Cependant, une telle politique risque, d'une part, de surcharger de requêtes ces agents réputés et, d'autre part, d'aller à l'encontre de l'ouverture du système en ne permettant pas à des agents nouveaux entrants d'interagir. Dans ce cas (et toujours pour cette politique), n'interagir qu'avec les agents les plus réputés ne permet – s'ils sont effectivement fiables – que de confirmer leur réputation mais, en empêchant des interactions avec d'autres agents, réduit la capacité du système à évaluer tous les agents.

Ces deux exemples mettent en lumière le fait qu'il est pertinent d'étudier la confiance et la réputation en la fiabilité ou l'honnêteté d'un point de vue dynamique, dans sa construction et dans son usage. C'est pour ces raisons qu'il serait intéressant de disposer d'un modèle générique pour l'analyse des systèmes de réputation qui permettrait cela. Or dans la littérature, ce type de modèle n'existe pas. Par exemple, considérons le modèle générique proposé par [Cheng et Friedman, 2005b] pour étudier les manipulations et qui permet de représenter la classe des systèmes de réputation personnalisés.

Définition 2.1
Soit $\mathcal{G} = (V, E)$ un graphe orienté où V est l'ensemble des agents $\{a_1 \dots a_n\}$ et $E \subseteq V \times V$ une relation d'interaction étiquetée par une valeur de confiance $c : E \mapsto [0, 1]$. La réputation de a_j selon a_i est donnée par une fonction $f_{\mathcal{G}} : V \times V \mapsto [0, 1]$ où :

$$f_{\mathcal{G}}(a_i, a_j) = \oplus_{P \in \mathcal{P}_{ij}} \odot (P)$$

$\mathcal{P}_{ij}$ est un ensemble de chemins entre a_i et a_j dans $\mathcal{G}$; $\odot$ est un opérateur d'agrégation de c le long d'un unique chemin P ; $\oplus$ est un opérateur d'agrégation de $\odot$ sur tous les chemins $\mathcal{P}_{ij}$.

Après avoir calculé la réputation de a_j, l'agent a_i doit décider s'il a confiance ou non. Or, dans ce modèle, il n'y a pas de méciseme permettant de décider si un agent a confiance dans un autre, ni aucune définition de l'évolution des confiances interpersonnelles, ce qui est représentatif de l'absence d'étude sur l'influence de cette fonction de décision sur le système de réputation.

De plus, définir cette influence, en étudier les effets et permettre à un agent de raisonner à son sujet (c'est-à-dire décider de la manière de construire la confiance qu'il accorde) prend du sens dans un contexte de questionnement éthique. En effet, cela permettrait de consider des **éthiques de la confiance**. Par exemple, une *éthique de la responsabilité* pourrait consister à n'interagir qu'avec des agents de confiance, une *éthique de l'indulgence* à ne pas tenir compte des premières interactions avec un agent, ou une *éthique de la réciprocité* à n'accorder la confiance qu'aux agents qui nous l'accordent ; et, à notre connaissance, de tels mécanismes n'ont pas été étudiés dans la littérature.

2 Formation de coalitions

Des agents autonomes sont parfois amenés à coopérer temporairement dans le but de réaliser collectivement une tâche qu'ils ne peuvent pas faire seuls. Dans ce cas, les agents doivent se demander avec quels agents coopérer. Ce problème est appelé un problème de *formation de coalitions*, couramment étudié au travers des jeux de coalitions.

2.1 Un bestiaire de modèles

Les *jeux de coalitions* ont été le sujet de très nombreuses publications dans la littérature [Nash, 1950, Shapley, 1952, Morgenstern et Von Neumann, 1953, Gamson, 1961, Kelso Jr et Crawford, 1982, Okada, 1996, Sandhlom et Lesser, 1997, Bloch, 1997, Ray et Vohra, 1999, Rahwan et Jennings, 2007, Elkind et Wooldridge, 2009, Génin, 2010, Hoefer *et al.*, 2014]. De tels jeux consistent, pour un ensemble $N = \{a_1, \ldots, a_n\}$ d'agents partageant le même environnement, à décider des sous-ensembles d'agents qui vont temporairement coopérer afin de réaliser collectivement une même tâche.

Définition 2.2 (Coalition)
Soit N, l'ensemble des agents. Une coalition $C \subseteq N$ *est un sous-ensemble non vide d'agents. La* coalition singleton *d'un agent $a_i \in N$ désigne la coalition $\{a_i\}$. La* grande coalition *est la coalition contenant l'ensemble des agents : $C = N$. L'ensemble des coalitions possibles pour N est noté par $\mathcal{C}^N$, et l'ensemble des coalitions possibles contenant l'agent $a_i \in N$ est noté $\mathcal{C}^N_{a_i}$.*

Exemple 2.3
Soit un système d'agents autonomes où $N = \{a_1, a_2, a_3\}$. La figure 2.1 représente $\mathcal{C}^N$.

Ici, l'ensemble des coalitions contenant l'agent a_1 est :

$$\mathcal{C}^N_{a_1} = \{\ \{a_1\}, \{a_1, a_2\}, \{a_1, a_3\}, \{a_1, a_2, a_3\}\ \}$$

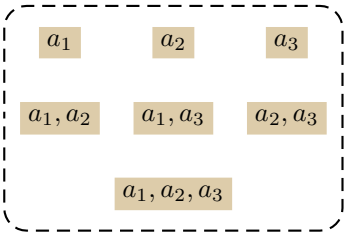

FIGURE 2.1 – Ensemble des coalitions possibles

Calculer à un instant donné quelles coalitions les agents vont former revient à trouver un partitionnement de l'ensemble des agents tel que chaque agent appartient à une seule et unique coalition. Une telle partition est appelée une *structure de coalitions*.

Définition 2.4 (Structure de coalitions)
Soit N, l'ensemble des agents. Une structure de coalitions *est un partitionnement de N, c'est-à-dire un ensemble de coalitions $\Pi = \{C_1, \ldots, C_k\}$ tel que les coalitions de Π sont :*

1. *non vides : $\forall i \in [1,k], C_i \neq \emptyset$;*
2. *deux à deux disjointes : $\forall i, j \in [1,k], i \neq j \implies C_i \cap C_j = \emptyset$;*
3. *couvrantes : $\forall a_i \in N, \exists C \in \Pi : a_i \in C$.*

Nous dénotons par $C^{\Pi}_{a_i}$ la coalition de l'agent a_i dans la structure de coalitions Π, et par $\mathcal{P}_N$ l'ensemble des structures de coalitions possibles à partir de N.

Notons que s'il est généralement considéré qu'un agent ne peut appartenir à un instant donné qu'à une seule et unique coalition, [Shehory et Kraus, 1998] étendent le problème aux cas des coalitions chevauchantes, c'est-à-dire celles où un agent peut appartenir simultanément à plusieurs coalitions. Cette généralisation leur permet de modéliser le problème d'affectation de tâches comme un problème de formation de coalitions afin d'obtenir une affectation qui maximise une fonction d'utilité.

La figure 2.2 montre l'ensemble des structures de coalitions possibles pour un ensemble d'agents $N = \{a_1, a_2, a_3\}$. Les arcs entre les différentes structures de coalitions représentent le passage d'une structure de coalitions à une autre lorsqu'un agent quitte sa coalition pour en rejoindre une autre.

Pour n agents, il existe $2^n - 1$ coalitions possibles, chaque agent étant présent dans 2^{n-1} de ces coalitions. Comme le montre [Wieder, 2008], le nombre de structures de coalitions possibles correspond au nombre de Bell :

$$B_{n+1} = \sum_{k=1}^{n} \binom{n}{k} B_n \text{ où } B_0 = 1$$

Afin de donner un ordre de grandeur, nous présentons dans la table 2.1 les 10 premiers nombres de Bell. Cet ordre de grandeur nous donne un aperçu de la complexité d'énumérer l'ensemble des structures de coalitions possibles afin de décider laquelle former.

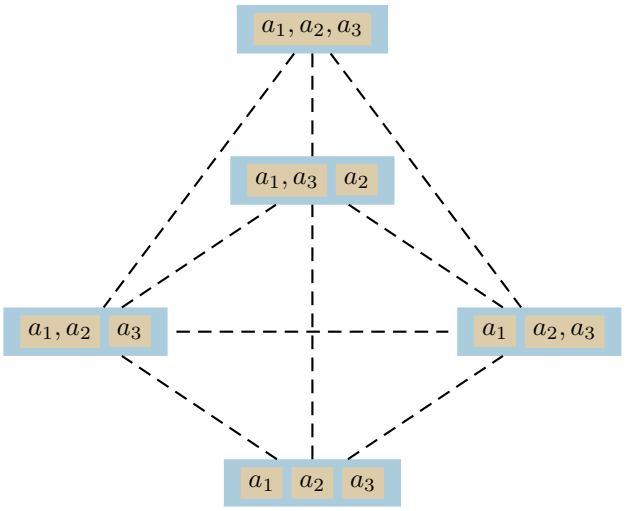

FIGURE 2.2 – Ensemble des structures de coalitions possibles

$\|N\|$	1	2	3	4	5	6	7	8	9	10
$B_{\|N\|}$	1	2	5	15	52	203	877	4140	21147	115975

TABLE 2.1 – Nombre de structures de coalitions possibles

Pour décider avec quels autres agents du système coopérer, les agents doivent pouvoir comparer les différentes coalitions. L'une des approches classiques de la théorie des jeux coopératifs [Shapley, 1952] est de considérer le gain que chaque agent va recevoir en formant cette coalition. Pour ce faire, les agents disposent d'une *fonction d'utilité* – aussi appelée *fonction caractéristique* lorsqu'elle décrit l'utilité de toutes les coalitions d'un point de vue global – qui définit pour chaque coalition C une valeur réelle correspondant aux gains que reçoit un agent si la coalition C se forme.

Définition 2.5 (Fonction d'utilité)
*Soit $N = \{a_1, \ldots, a_n\}$ un ensemble d'agents. L'*utilité d'une coalition $C \subseteq N$ *pour l'agent $a_i \in C$ est définie par $u_{a_i} : 2^N \to \mathbb{R}$.*

De nombreux travaux portent sur les formes de la fonction caractéristique. Le cas classique est celui des *jeux de coalitions à utilité transférable* où l'utilité d'une coalition est répartie entre les agents qui composent cette coalition [Shapley, 1952, Gamson, 1961, Hart et Kurz, 1983, Winter, 1989] et où les agents cherchent à former une structure de coalitions qui produit le maximum d'utilité [Aumann et Dreze, 1974, Rahwan, 2007]. Toutefois, d'autres travaux se sont intéressés à d'autres formes de jeux.

— Les **jeux de coalitions bayésiens** prennent en considération une notion d'incertitude. Pour cela, la fonction caractéristique est définie par une distribution de

probabilité sur l'utilité que tirent les agents d'une coalition après distribution des gains [Chalkiadakis *et al.*, 2007, Ieong et Shoham, 2008, Yang et Gao, 2014].

— Les **jeux de coalitions recouvrantes** modélisent des jeux où les agents peuvent distribuer leur participation entre plusieurs coalitions et où chaque coalition produit une utilité dépendante de la participation de ses membres [Chalkiadakis *et al.*, 2010]. Une coalition C est désormais un vecteur $\vec{r}$ où chaque composante $r_i \in [0; 1]$ représente la participation de l'agent a_i à C. La fonction caractéristique devient donc de la forme $v : \mathbb{R}^n \to \mathbb{R}$.

— Les **jeux de compétences** introduisent explicitement la notion de tâches à réaliser. Pour cela, chaque agent dispose de compétences et les tâches en nécessitent pour être accomplies. Les coalitions sont alors définies par leur pouvoir, c'est-à-dire l'ensemble des compétences de ses membres. Ces modèles sont assez proches des *jeux de coalitions quantitatifs* et *jeux de ressources* où les compétences sont remplacées par des ressources [Bachrach et Rosenschein, 2008].

— Les **jeux de coalitions à externalités** expriment le fait que la valeur d'une coalition dépend (en partie) des autres coalitions qui co-existent dans la même structure stable [Ray et Vohra, 1999, De Clippel et Serrano, 2005, Michalak *et al.*, 2009, Grabisch et Funaki, 2012]. Pour cela, la fonction caractéristique est remplacée par une fonction de partition de la forme $\mathcal{P} : 2^N \times 2^{2^N} \to \mathbb{R}$. Ce modèle permet de représenter des jeux de votes où chaque coalition vote pour une option qui peut donc avoir une utilité différente selon les agents.

— Enfin, les **jeux de coalitions à utilité non-transférable** représentent des jeux où les agents ne peuvent pas se distribuer l'utilité des coalitions une fois qu'elles sont formées [McKelvey *et al.*, 1978, Harsanyi, 1963, Aumann, 1985, Winter, 1991, Suzuki *et al.*, 2015]. Ici, il n'y a donc plus de fonction caractéristique mais les agents expriment une relation de préférence entre les coalitions. Si ces jeux sont généralement des *jeux hédoniques* décrits en détail dans la section suivante [Dreze et Greenberg, 1980], certains modèles plus spécifiques comme les *jeux de coalitions qualitatifs* ou les *jeux fractionnels* ont été proposés [Aziz *et al.*, 2013a].

2.2 Le cas des jeux hédoniques

Si l'utilisation des fonctions caractéristiques permet une évaluation quantitative des coalitions auxquelles chaque agent peut appartenir, une autre approche consiste à définir un opérateur de comparaison ordinal entre les structures de coalitions. Cette approche est celle des *jeux hédoniques* [Drèze et Greenberg, 1980, Bogomolnaia et Jackson, 2002].

Définition 2.6 (Jeu hédonique)
Un jeu hédonique *est défini par un couple* $HG = \langle N, \succeq \rangle$ *où* N *désigne l'ensemble des agents et* $\succeq$ *l'ensemble des profils de préférence des agents.*

Le profil de préférence d'un agent désigne un ordre total sur l'ensemble des $2^{|N|-1}$ coalitions auquel il peut appartenir. Pour deux coalitions C_1 et C_2, $C_1 \succ_{a_i} C_2$ signifie que l'agent a_i préfère strictement la coalition C_1 à la coalition C_2. Remarquons que de nombreux travaux s'intéressent aussi à leurs représentations compactes [Hajduková *et al.*, 2003, Ballester, 2004, Aziz *et al.*, 2014] sous diverses formes [3].

— Les *listes de coalitions individuellement rationnelles* (IRCL) modélisent des agents rationnels qui ne vont pas accepter de former une coalition moins préférée à leur coalition singleton [Ballester, 2004]. Ainsi, toute coalition $C \in \mathcal{C}^N_{a_i}$ telle que $\{a_i\} \succ_{a_i} C$ n'a pas besoin d'être modélisée dans le profil de préférence de l'agent $a_i \in N$.

— Les *jeux à additivité séparable* modélisent les préférences des agents vis-à-vis des autres agents et non plus vis-à-vis de l'ensemble des coalitions. Chaque agent dispose d'une fonction $v_{a_i} : N \longrightarrow \mathbb{R}$ et la valeur d'une coalition est une agrégation (les opérateurs diffèrent selon les auteurs) des valeurs des agents qui la composent [Hajduková *et al.*, 2003, Hajduková *et al.*, 2004, Aziz *et al.*, 2011]. Des règles de départage comme un ordre lexicographique permettent d'obtenir un ordre strict sur les structures de coalitions lorsque la représentation ne permet pas de comparer deux structures.

Au-delà de ces considérations, indépendamment du fait que les agents comparent les coalitions par une fonction d'utilité ou par des profils de préférence, les travaux portant sur les jeux de coalitions, s'intéressent principalement à deux questions :

1. quelles sont les structures de coalitions acceptables pour les agents ?
2. comment former une telle structure de coalitions ?

Ces propriétés sont caractérisées par un *concept de solution* qui définit les propriétés que doit satisfaire une partition pour être considérée comme *stable*, c'est-à-dire une partition où aucun agent ne désire changer de coalition. La table 2.2 présente les concepts de solution classiquement considérés dans la littérature [Greenberg, 1994, Bogomolnaia et Jackson, 2002, Ballester, 2004, Elkind et Wooldridge, 2009, Aziz *et al.*, 2011, Aziz *et al.*, 2013c, Aziz *et al.*, 2013b, Brandl *et al.*, 2015, Peters et Elkind, 2015].

Notons que tous ces concepts peuvent être généralisées aux jeux à utilité transférable en considérant que pour deux coalitions C_1 et C_2 : $C_1 \succ_{a_i} C_2 \iff u_{a_i}(C_1) > u_{a_i}(C_2)$ et $C_1 \sim_{a_i} C_2 \iff u_{a_i}(C_1) = u_{a_i}(C_2)$. De plus, ils peuvent être déclinés en **formes affaiblies** en considérant des préférences non-strictes ($\succeq$). Enfin, chacun de ces concepts de solution représente un comportement spécifique que doivent suivre les agents dans le processus de formation de coalitions. À titre d'exemple, la meilleure structure de coalitions – appelée *structure optimale* – est celle qui satisfait parfaitement l'ensemble des participants. Un autre exemple est celui des *structures Pareto-optimales.*

3. Il existe une généralisation appelée *réseaux de jeux hédoniques* [Elkind et Wooldridge, 2009].

Concept de solution	**Propriété**
Stabilité de Nash (NS)	$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi)$
Stabilité Individuelle (IS)	$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi)$ $\wedge \forall a_j \in C, C \cup \{a_i\} \succeq_j C$
Stabilité Individuelle Contractuelle (ICS)	$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi)$ $\wedge \forall a_j \in C, C \cup \{a_i\} \succeq_j C$ $\wedge \forall a_k \in C_i(\Pi), a_k \neq a_i, C_i(\Pi) \setminus \{a_i\} \succeq_k C_i(\Pi)$
Stabilité Contractuelle de Nash (CNS)	$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi)$ $\wedge \forall a_k \in C_i(\Pi), a_k \neq a_i, C_i(\Pi) \setminus \{a_i\} \succeq_k C_i(\Pi)$
Stabilité du Cœur (CS)	$\forall a_i \in N, \nexists C \in G_i : C \succ_i C_i(\Pi)$ $\wedge \forall a_j \in C, C \succeq_j C_j(\Pi)$
Optimalité (O)	$\forall a_i \in N, \nexists C \in G_i : C \succ_i C_i(\Pi)$
Pareto-Optimalité (PO)	$\nexists \Pi_2 : \forall a_i \in N, C_i(\Pi_2) \succeq_i C_i(\Pi)$ $\wedge \exists a_j \in N, C_j(\Pi_2) \succ_j C_j(\Pi)$

TABLE 2.2 – Principaux concepts de solution

Exemple 2.7
Considérons un jeu $HG = \langle N, \succeq \rangle$ *tel que :*

$$
\begin{aligned}
N &= \{a_1, a_2, a_3\} \\
\succeq_{a_1} &= \{a_1, a_2\} \succ_{a_1} \{a_1, a_3\} \succ_{a_1} \{a_1, a_2, a_3\} \succ_{a_1} \{a_1\} \\
\succeq_{a_2} &= \{a_1, a_2\} \succ_{a_2} \{a_2, a_3\} \succ_{a_2} \{a_1, a_2, a_3\} \succ_{a_2} \{a_2\} \\
\succeq_{a_3} &= \{a_1, a_3\} \succ_{a_3} \{a_2, a_3\} \succ_{a_3} \{a_1, a_2, a_3\} \succ_{a_3} \{a_3\}
\end{aligned}
$$

La structure de coalitions $\{ \{a_1\}, \{a_2\}, \{a_3\} \}$ *est dominée au sens de Pareto par* $\{ \{a_1, a_2\}, \{a_3\} \}$ *car* a_1 *et* a_2 *préfèrent être ensemble tandis que* a_3 *est indifférent (car dans les deux cas, il est dans sa coalition singleton). La figure 2.3 montre les dominances au sens de Pareto pour les différentes structures de coalitions. Ici, les trois structures* $\{ \{a_1, a_2\}, \{a_3\} \}$, $\{ \{a_1, a_3\}, \{a_2\} \}$, $\{ \{a_1\}, \{a_2, a_3\} \}$ *sont optimales au sens de Pareto.*

Par définition, certains de ces concepts sont des généralisations des autres : il existe une relation d'inclusion entre les ensembles stables au sens de Nash, individuellement stables et individuellement contractuellement stables. La figure 2.4 résume les relations entre les différents concepts que nous avons présentés. Un arc allant du concept A au concept B ($A \longrightarrow B$) signifie que A est inclus dans B. L'hyperarête en pointillés indique les concepts de solution *irrationnels*, c'est-à-dire les concepts dont la satisfaction ne garantissent pas

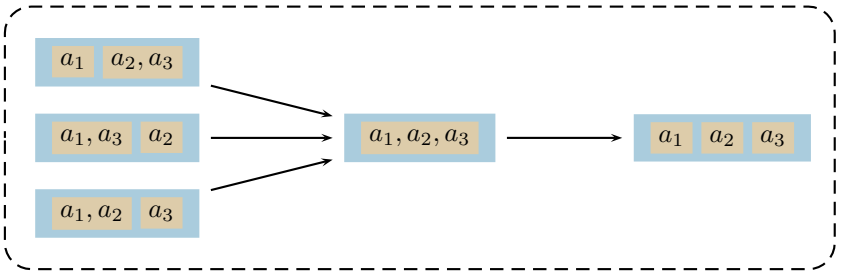

FIGURE 2.3 – Dominance au sens de Pareto des structures de coalitions

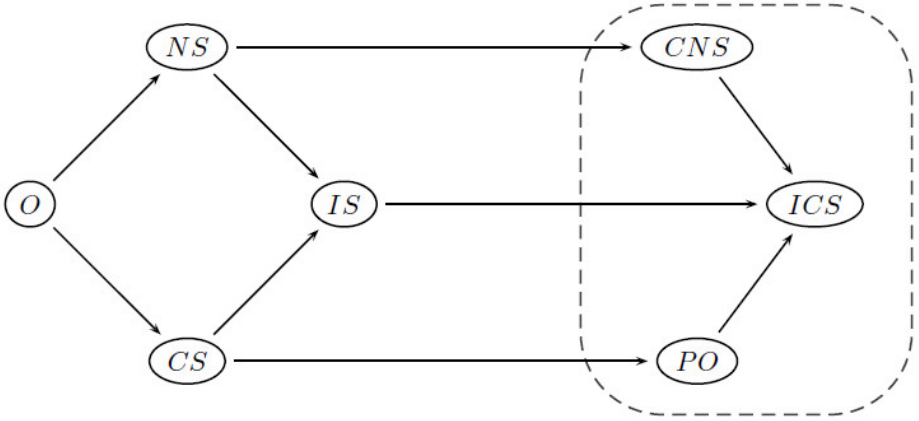

FIGURE 2.4 – Relations d'inclusions entre les concepts de solution

	CS	NS	IS
IRCL de taille $\leq n^3$	NP-c	NP-c	NP-c
Réseaux de coalitions hédoniques	NP-h	NP-c	NP-c
$\mathcal{W}$-préférences (avec règle de départage)	P	NP-c	?
$\mathcal{W}$-préférences	NP-c	NP-c	NP-c
$\mathcal{W}\beta$-préférences (avec règle de départage)	P	NP-c	?
$\mathcal{W}\beta$-préférences	NP-c	NP-c	NP-c
Jeux à additivité séparable	NP-h	NP-c	NP-c

TABLE 2.3 – Problèmes d'existence des concepts selon les modèles de préférences

aux agents d'être dans une coalition *a minima* équivalente en termes de préférences à leur coalition singleton.

Au-delà de la définition de concepts de solution, décider de l'existence d'une structure de coalitions appartenant à l'un de ces concepts est un problème important. La table 2.3

présente quelques résultats de complexité pour différentes représentations des préférences et concepts de solution, extraits du travail de [Peters et Elkind, 2015]. Remarquons que les problèmes qui en résultent sont presque tous difficiles. Malgré cela, de nombreux algorithmes de formation de coalitions ont été proposés [Sandholm, 1999, Larson et Sandholm, 2000, Génin, 2010] et nous invitons le lecteur à se référer au travail de synthèse très complet réalisé par [Rahwan *et al.*, 2015]. Synthétiquement, en dehors des approches naïves qui consistent à énumérer toutes les structures de coalitions possibles, nous distinguons deux grandes familles d'algorithmes de formation de coalitions :

— les *algorithmes de génération des structures de coalitions* sont des algorithmes *anytime* centralisés et, bien souvent, distribués qui reposent sur une borne permettant de trouver une structure de coalitions proche de l'optimum en termes de bien-être social sans avoir à parcourir nécessairement l'ensemble de structures de coalitions existantes [Sandholm, 1999, Larson et Sandholm, 2000, Dang et Jennings, 2004, Rahwan, 2007, Rahwan et Jennings, 2007, Rahwan *et al.*, 2007, Rahwan et Jennings, 2008, Keinänen, 2010, Michalak *et al.*, 2010] ;

— les *protocoles de négociation* sont des algorithmes décentralisés dans lesquels les agents échangent des propositions de coalitions à former avec, selon les protocoles, la possibilité de quitter une coalition en cours de formation ou non [Kelso Jr et Crawford, 1982, Vauvert et El Fallah-Seghrouchni, 2001, Aknine *et al.*, 2004, Génin, 2010, Génin et Aknine, 2011].

2.3 Hétérogénéité des concepts de solution et valeurs éthiques

Les jeux de coalitions et les jeux hédoniques sont particulièrement intéressants pour étudier les notions de fiabilité et d'honnêteté. En effet, la fonction caractéristique et les préférences des agents peuvent être **fondées sur une mesure de fiabilité**, comme une notion de confiance entre les agents. La question de l'honnêteté peut, quant à elle, être abordée par **le prisme des manipulations** car un agent peut avoir intérêt à mentir sur ses préférences ou la valeur qu'il accorde à une coalition pour en tirer un avantage. Mais qu'en est-il des questions d'éthique ?

Classiquement en théorie des jeux, les notions de stabilité et d'équité sont souvent vues comme répondant à des critères éthiques dont le sens est intimement lié à leur définition.

— La **stabilité** exprime sous quelles conditions il est acceptable de changer de coalition sachant la distribution de l'utilité d'une partition [Driessen, 1991]. Par exemple, le *cœur* exprime le fait qu'il est acceptable de changer de coalition si un agent ne reçoit pas une utilité supérieure ou égale à celle qu'il pourrait recevoir s'il était seul, exprimant une éthique individualiste. Le *dernier cœur* rend acceptable le fait qu'au moins un agent sacrifie une partie de son utilité pour assurer la stabilité et que le sacrifice maximal parmi les agents est minimisé. Cependant, le dernier cœur peut être dictatorial s'il existe un sous-ensemble d'agents pouvant forcer les autres à accepter

un sacrifice afin de trouver une solution stable. Le *nucléole* est plus équitable car il s'agit d'un dernier cœur qui minimise le sacrifice parmi tous les agents [Schmeidler, 1969].

— L'**équité** exprime une notion de justice sur la distribution de l'utilité entre les agents. Les premières approches ont considéré une distribution autour des valeurs de Shapley [Shapley, 1953] ou de l'indice de Banzhaf [Banzhaff III, 1964], exprimant un critère de mérite en fonction de la contribution des agents aux coalitions. Ces approches sont axiomatisées par des propriétés de symétrie, d'efficacité, de monotonie, d'additivité et de joueur nul. Relâcher ces axiomes permet de définir des familles de valeurs [Yang, 1997] et de capturer des critères éthiques autres que le mérite. Par exemple, la *rationing value* ne considère plus l'axiome d'efficacité et autorise les agents à ne pas distribuer toute l'utilité [Yang, 1997] et la valeur de solidarité relâche l'axiome de joueur nul afin de permettre aux agents de donner plus d'utilité aux agents qui contribuent le moins [Nowak et Radzik, 1994].

Cependant, au vu de l'état de l'art précédent, nous pouvons remarquer que le concept de solution est toujours une propriété globale du jeu, c'est-à-dire qu'il est le même pour tous les agents. Or, cela revient à considérer les critères éthiques sous-jacents (plus ou moins de mérite, plus ou moins de solidarité, plus ou moins d'individualisme) comme étant les mêmes pour tous, faisant fi d'une certaine hétérogénéité. Afin d'illustrer notre propos, considérons l'exemple suivant dans un contexte de jeu hédonique.

Quatre agents (a_1, a_2, a_3 et a_4) – ont pour objectif de faire un trajet commun et cherchent une solution de co-voiturage. Chaque agent dispose de préférences, représentant ses intérêts, vis-à-vis des passagers avec qui il peut partager une voiture. Supposons que a_1 et a_4 ne s'apprécient pas et refusent de partager la même voiture, c'est-à-dire préfèrent être seuls qu'avec l'autre. En revanche, tous deux préfèrent être avec a_2 et a_3 en même temps plutôt qu'être seulement avec l'un d'entre eux. a_2 préfère partager sa voiture avec a_3 qui, lui, préfère être avec a_1.

Si la répartition consiste à faire deux voitures, a_1 seul et a_2, a_3 et a_4 ensemble, alors selon la stabilité au sens de Nash a_3 peut décider de changer de voiture pour rejoindre a_1. Ce choix s'opère en fonction des intérêts de a_3 et de sa préférence pour a_1. Toutefois, ce choix peut aussi résulter de la manière dont a_3 définit une solution acceptable au vu de critères éthiques qui lui sont propres (par exemple il pourrait considérer que le bien-être des agents qu'il rejoint est plus important que son propre bien-être). Intuitivement, il semble alors intéressant d'exprimer des concepts de solution hétérogènes, propres à chaque agent. Par exemple, nous pourrions supposer que a_1, a_2, a_3 et a_4 se comportent différemment (selon des concepts de solution canonique ou des comportements singuliers) :

— a_1 ne rejoint une voiture que si les passagers de cette dernière l'acceptent,

— a_2 ne rejoint une voiture que si les passagers de cette voiture ainsi que ceux de la voiture qu'il quitte l'acceptent,

— a_3 rejoint une voiture s'il préfère voyager avec les passagers de cette dernière,
— a_4 ne rejoint une voiture que si cela est préféré par tous les autres agents, sans considération pour ses propres préférences.

Supposons alors la répartition des passagers en deux voitures avec a_1 et a_3 dans l'une et a_2 et a_4 dans l'autre. Cette répartition satisfait les agents car aucun ne désire alors changer de voiture. Elle fait alors consensus au sens qu'elle respecte au mieux les intérêts de chacun et **les manières hétérogènes que chacun considère comme acceptables pour former des coalitions**. Ainsi, dans le contexte de la formation de coalitions, il nous semble pertinent d'aborder les questions d'éthique par la prise en compte de cette hétérogénéité.

3 Modèles d'agents cognitifs

Les deux modèles formels présentés précédemment, qu'il s'agisse des systèmes de réputation ou des jeux de coalitions, s'intéressent aux agents autonomes du point de vue du système dans lequel ils sont plongés. Ce sont les interactions entre les agents qui sont au cœur de ces modèles. C'est pourquoi il est aussi important de se tourner vers les modèles d'agents cognitifs qui, eux, s'attachent aux mécanismes internes des agents leur permettant de prendre des décisions. Cette section présente alors une vue de haut niveau de ces modèles en mettant l'accent sur les architectures BDI qui nous semblent les plus à même de traiter les questionnements liés à l'honnêteté et l'éthique dans les systèmes d'agents autonomes.

3.1 Architectures BDI

Les architectures BDI [4] reposent sur un des modèles d'agents cognitifs le plus étudié dans la littérature. Originaire des travaux philosophiques de [Bratman, 1990] et [Dennett, 1987] sur la définition du *raisonnement pratique* et de la *posture intentionnelle*, ce modèle guide la sélection des actions réalisées par l'agent en distinguant différents types d'états mentaux : (1) les *croyances* constituent la représentation mentale des informations dont l'agent dispose sur l'état du monde courant ; (2) les *désirs* ou *buts* de l'agent déterminent les objectifs à atteindre ; (3) les *intentions* de l'agent sont des engagements à réaliser un sous-ensemble de désirs au regard de ses croyances.

À partir des premiers travaux de [Bratman, 1987] et [Dennett, 1987], deux modèles logiques principaux ont été définis [Herzig *et al.*, 2016, Meyer *et al.*, 2015] :

1. [Cohen et Levesque, 1990] construit une logique BDI fondée sur une **logique temporelle linéaire** quantifiée avec des modalités d'action et de croyance, distinguant les intentions potentielles des intentions concrètes. Pour cela, il introduit quatre

4. Beliefs, Desires and Intentions.

étapes passant de la génération des buts choisis (les états que l'agent désire atteindre), les buts réalisables (buts choisis que l'agent croit ne pas avoir atteints), les buts persistants (buts réalisables qui ne sont abandonnés que si l'agent les pense réalisés ou irréalisables) et, enfin, les intentions (buts persistants que l'agent est prêt à réaliser).

2. [Rao et Georgeff, 1991] considère une **logique temporelle arborescente** où chaque état mental BDI dispose d'opérateurs qui lui sont propres ainsi que d'opérateurs de compatibilité avec les autres états mentaux. En particulier, l'intention est une modalité et non plus un prédicat comme dans l'approche de [Cohen et Levesque, 1990].

Ces logiques se sont vues accompagnées de langages de programmation dédiés comme AgentSpeak(L) [Rao, 1996], JACK [Howden *et al.*, 2001], 2APL [Dastani, 2008] ou Jason [Bordini *et al.*, 2007]. À titre d'exemple, nous présentons ci-dessous deux de ces langages :

— JACK étend la syntaxe de Java pour représenter des objets orientés-agents avec des classes spécifiques aux éléments du modèle BDI. Par exemple, la classe `BeliefSet` maintient le modèle du monde de l'agent – encapsulée dans la classe `View` – en respectant des contraintes de cohérence et propose des méthodes pour y faire des requêtes. JACK a été étendu en JACK Teams pour représenter les activités coordonnées au sein d'équipes d'agents, chaque équipe disposant de croyances, de désirs et d'intentions séparés.

— Jason [Howden *et al.*, 2001] est un langage de programmation et un interpréteur d'AgentSpeak(L). Au-delà des classes traditionnelles encapsulant les éléments d'un modèle BDI, Jason propose des actes de langage pour la communication, l'étiquetage des plans pour être réutilisés dans les modules de décision, des fonctions de confiance et la possibilité d'importer du code. L'interpréteur Jason a été intégré dans une plateforme multi-agent – appelée JaCaMo [Boissier *et al.*, 2016] – qui intègre aussi une dimension normative et des artefacts extérieurs aux agents.

Une des propriétés les plus intéressantes des architectures BDI est leur modularité, ce qui fait que, selon la manière dont la gestion des engagements sur les désirs et les intentions est réalisée, plusieurs types d'agents peuvent être définis : par exemple des agents fanatiques ou agents à obligation aveugle qui maintiennent leurs intentions jusqu'à ce que le but associé soit réalisé, ou des agents ouverts ou agents à obligation ouverte qui maintiennent leurs intentions jusqu'à ce qu'ils croient que le but associé n'est plus réalisable. Plus généralement, il est assez naturel de vouloir étendre le modèle BDI en y introduisant de nouvelles notions qui peuvent être exprimées par un raisonnement logique, comme des normes, des notions de confiance, des émotions, etc. Qu'en est-il de la fiabilité, de l'honnêteté et de l'éthique ?

3.2 Logiques de la confiance

Nous avons déjà vu en section 1.1 que les modèles de confiances se déclinaient en approches quantitatives et qualitatives. Ces dernières s'appuient essentiellement sur le fait que la confiance se construit à partir d'états mentaux [Castelfranchi et Falcone, 2010]. Dans ce contexte, les logiques modales, en permettant l'expression de modalités d'intention, croyance, action ou but, sont bien adaptées pour modéliser la confiance et peuvent naturellement s'intégrer dans une architecture BDI. Par exemple, [Herzig *et al.*, 2010] considèrent la confiance comme un prédicat signifiant que l'agent a_i a confiance en un agent a_j à propos d'une action α qui a pour conséquence la proposition ϕ si, et seulement si, chacune des expressions suivantes est vraie :

1. a_i a pour but que ϕ,
2. a_i croit que :
   (a) a_j est capable de faire l'action α,
   (b) a_j en faisant l'action α va permettre ϕ,
   (c) a_j a l'intention de faire α.

Un autre exemple de confiance est celle définie par [Smith *et al.*, 2011] :

1. a_i a pour but que ϕ,
2. a_i croit que a_j réalise ϕ,
3. a_i a l'intention que :
   (a) a_j réalise ϕ,
   (b) a_i ne réalise pas ϕ.
4. a_i a pour but que a_j a l'intention de ϕ,
5. a_i croit que a_j a l'intention de ϕ.

Ces deux définitions diffèrent sur le contexte d'application de la confiance. Contrairement à la première formulation, la seconde exprime de manière sous-jacente une notion de délégation d'action : l'agent qui fait confiance a l'intention de ne pas réaliser l'action et il a l'intention que l'autre agent la réalise. Toutefois, dans les deux cas, il s'agit de prédicats de *confiance occurrente* exprimant la confiance d'un agent à l'instant présent : l'agent a_j se prépare à accomplir l'action pour laquelle l'agent a_i lui fait confiance. D'autres travaux s'intéressent à la *confiance dispositionnelle* exprimant que l'agent a_i a confiance en l'agent j à propos d'une proposition ϕ dans un contexte spécifique ψ qui n'est pas nécessairement le contexte présent. Ici, la confiance est généralement représentée par une modalité [Liau, 2003, Dastani *et al.*, 2004, Singh, 2011]. Par exemple, [Singh, 2011] propose une modalité $T^d_{i,j}(\psi, \phi)$ signifiant que l'agent a_i a confiance dans a_j pour réaliser ϕ dans un contexte ψ et la confiance occurente devient un cas spécifique exprimé par $T^d_{i,j}(\top, \phi)$.

Cependant, tous ces modèles traitent d'une notion de **confiance en la fiabilité** d'un agent. D'autres aspects de la confiance sont plus rarement traités, comme par exemple la

confiance en la sincérité ou la confiance en l'honnêteté. Par exemple, [Demolombe, 2004] propose une logique multimodale – avec les modalités K_i, B_i, $Com_{i,j}$, O, P et E_i qui sont respectivement la connaissance, la croyance, le fait de communiquer, l'obligation, la permission et une modalité d'intention – et définit la **confiance en l'honnêteté** comme suit :

$$Thon_{i,j}(\phi) \triangleq K_i(E_j\phi \Rightarrow PE_j\phi)$$

Ici, un agent a_i a confiance dans l'honnêteté d'un agent a_i si, et seulement si, a_i sait que si a_j a l'intention de ϕ alors il est permis à a_j de faire ϕ. Avec la même logique, [Demolombe, 2004] propose aussi une notion de **confiance en la sincérité** signifiant qu'un agent a_i a confiance dans la sincérité de a_j si, et seulement si, a_i sait que, si a_j lui communique ϕ, alors a_j croit que ϕ. Formellement :

$$Tsinc_{i,j}(\phi) \triangleq K_i(Com_{j,i}\phi \Rightarrow B_j\phi)$$

Bien entendu, d'autres travaux se sont intéressés à ces questions. Par exemple, [Christianson et Harbison, 1997] ont, eux aussi, caractérisé la confiance en l'honnêteté mais cette dernière est en fait équivalente à la confiance en la sincérité de [Demolombe, 2004] ; [Liau, 2003] a proposé une modalité de confiance dans le jugement d'un autre agent sur une proposition mais, d'un point de vue formel, il s'agit d'une confiance en la fiabilité d'un discours plutôt qu'une confiance en la sincérité ou l'honnêteté d'un agent.

Ainsi, de nombreux travaux existent pour traiter de la caractérisation de la confiance en la fiabilité. Toutefois, les autres aspects de la confiance sont beaucoup moins traités dans la littérature, ou de manière limitée. Par exemple, dans les travaux de [Demolombe, 2004], l'honnêteté est réduite à l'absence de violation de normes et, bien que la notion de sincérité soit capturée, elle ne se comporte pas comme un système KD, ce qui peut poser problème puisqu'il sera alors possible de faire confiance à un agent sur une proposition alors que cet agent annonce tout à la fois cette proposition et son contraire.

3.3 Éthique et modèles BDI

Si certains travaux dans le domaine – comme ceux de [Broersen *et al.*, 2001, Wiegel, 2006] – ont prosaïquement ajouté un composant déontique aux composants doxastiques et intentionnels du modèle BDI, l'éthique ne peut être réduite aux obligations et permissions. Deux [5] grands types d'approches ont été proposés :

Certains travaux comme ceux de [Lorini, 2012, Vanhée, 2015] s'attachent à représenter une **morale utilitariste rationnelle**. Pour cela, ils proposent d'ajouter explicitement à l'architecture BDI un ensemble de valeurs morales (ou plus généralement culturelles pour

5. Nous avons volontairement omis les travaux déjà cités au chapitre précédent sur le raisonnement moral et éthique – comme ceux de [Chisholm, 1963, Gensler, 1996, Powers, 2005, Ganascia, 2007, Bringsjord et Taylors, 2012, Berreby *et al.*, 2015] – qui ne s'appuient pas sur des modèles BDI.

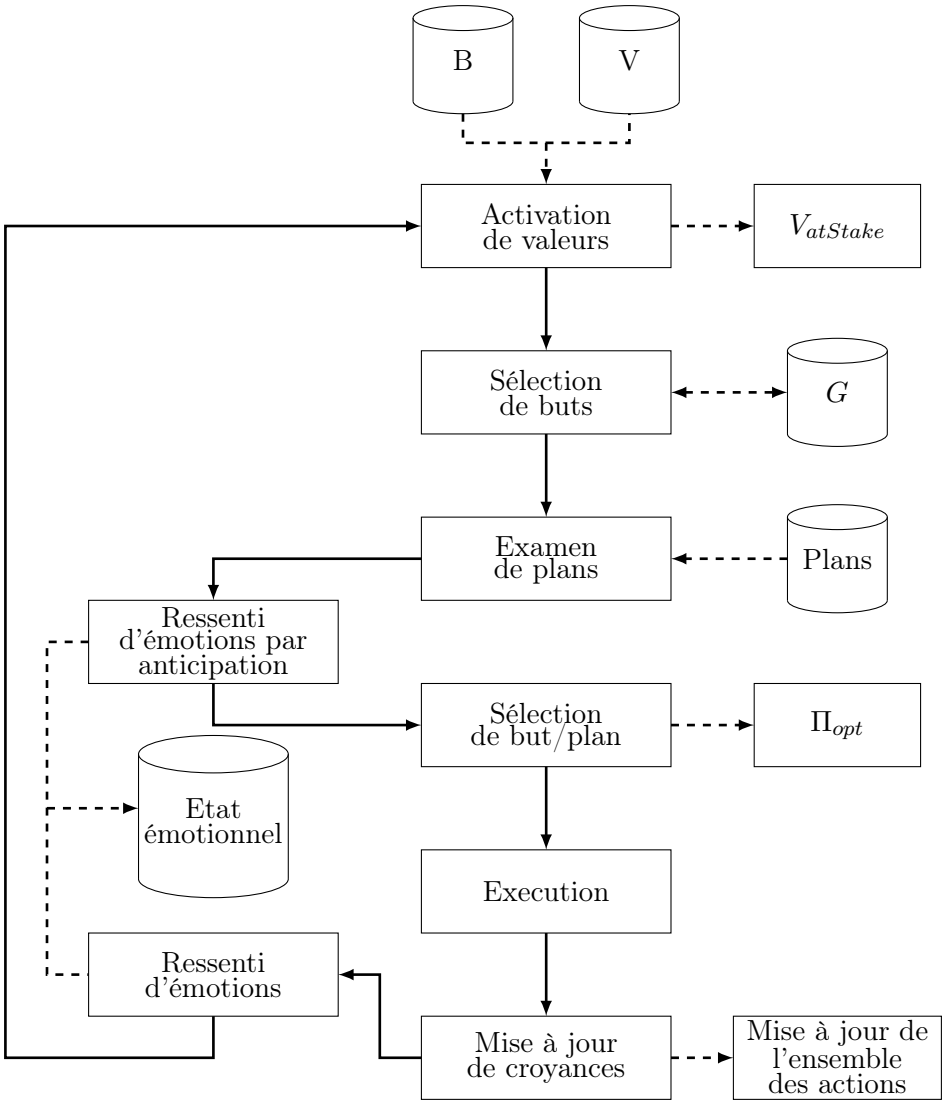

FIGURE 2.5 – Architecture des agents BDI émotionnels de [Battaglino *et al.*, 2013]

[Vanhée, 2015]). Deux types de problèmes éthiques en résultent : soit des conflits entre les désirs et la morale, soit des conflits entre valeurs morales elles-mêmes. L'éthique proposée est alors un principe reposant sur un calcul d'utilité et une relation de préférence entre désirs et valeurs. Selon le paramétrage des agents, certains peuvent avoir une définition de l'utilité sous un angle purement hédoniste ou au contraire définir l'utilité exclusivement en fonction de la satisfaction morale apportée par l'action. Remarquons que ces propositions ne représentent pas de jugement des comportements des autres et se focalisent sur un point précis du raisonnement de l'agent qui est celui du raisonnement sur la conciliation des désirs et des valeurs dans un cadre BDI opérationnel.

Partant des travaux de [Antunes et Coelho, 1999] sur les architectures BVG [6], certains travaux viennent ajouter au modèle BDI des valeurs et des états émotionnels, et décrivent une **morale intuitionniste** [Coelho et da Rocha Costa, 2009, Coelho *et al.*, 2010, Battaglino *et al.*, 2013]. Une des architectures représentatives est illustrée en figure 2.5. De manière générale, ces architectures s'appuient sur un cycle de raisonnement de l'agent dans lequel viennent se mêler des étapes de raisonnement classiques des modèles BDI et des modifications d'un état émotionnel de l'agent. Les plans sont ainsi non seulement évalués pour les possibilités qu'ils offrent d'assouvir les désirs de l'agent, mais également pour les émotions qu'ils évoquent. Les contraintes morales sont des règles associées à des états de connaissance et des états émotionnels qui contraignent alors les plans. De même, le changement d'état du monde perçu après exécution de l'action peut entraîner des modifications de l'état émotionnel. Cette approche déclarative cherche à représenter de manière explicite les aspects émotionnels du jugement éthique. L'apport d'éléments intuitionnistes au modèle BDI a pour but, selon les auteurs, d'introduire une certaine empathie dans l'interaction avec d'autres agents, artificiels ou humains. La généricité de cette approche vis-à-vis du domaine applicatif et de la morale confiée à l'agent est assurée par la séparation entre l'architecture du processus décisionnel d'une part et les connaissances (valeurs morales, plans, buts, croyances) données en paramètres.

Les approches présentées ci-dessus proposent des méthodes intéressantes pour construire des agents exhibant des comportements éthiques. Toutefois, dans un système d'agents autonomes, ces derniers doivent interagir et travailler ensemble pour échanger des données, des ressources ou réaliser des actions jointes. Les approches précédentes considèrent généralement les autres agents comme des éléments de l'environnement. Cependant, dans une perspective de systèmes d'agents autonomes, ces derniers doivent être capable de représenter, juger et tenir compte de l'éthique des autres. Ceci nous permet d'identifier deux besoins fondamentaux :

1. Les agents ont besoin d'une **représentation explicite de l'éthique** comme le suggère la théorie de l'esprit. En effet, l'éthique des autres ne peut être comprise sans une représentation individuelle de l'éthique [Kim et Lipson, 2009]. Pour exprimer et concilier les théories morales et les théories éthiques, nous proposons de représenter distinctement les deux théories : une théorie du bien, divisée en valeurs morales et règles morales, et une théorie du juste, divisée entre principes éthiques et préférences éthiques. De plus, une telle représentation facilite la configuration des agents par des non-spécialistes de l'intelligence artificielle.
2. Les agents ont besoin d'un **processus explicite de jugement éthique** leur permettant de raisonner sur les théories du bien et les théories du juste aussi bien d'un point de vue individuel que d'un point de vue collectif. Nous considérons un jugement éthique comme une évaluation de la conformité des actions des agents au regard des théories du bien et du juste. Nous proposons alors plusieurs types de

6. *Beliefs, Values and Goals.*

	Fiabilité	Honnêteté	Éthique
Confiance	Q2	Q5	Q8
Coalitions	Q3	Q6	Q9
Cognition	Q1	Q4	Q7

TABLE 2.4 – Croisement des axes et des approches formelles

jugements fondés sur une capacité à substituer la morale et l'éthique d'un agent par celles d'un autre.

4 Croisement des questionnements

En conclusion de ce chapitre, nous positionnons les questions évoquées au chapitre précédent au regard des trois approches formelles que nous avons détaillées et certains travaux choisis [7] que nous avons réalisés. Pour rappel, nos questions de recherche sont :

1. Comment caractériser qualitativement (Q1) et quantitativement (Q2) la fiabilité d'un agent et quelle influence cette caractérisation a-t-elle sur un système d'agents autonomes en fonction de la manière dont les agents en question s'en servent (Q3) ?
2. Comment caractériser qualitativement (Q4) et quantitativement (Q5) l'honnêteté d'un agent et quel mode d'organisation des agents permet de garantir le respect de cette valeur (Q6) ?
3. Comment représenter et raisonner sur des valeurs morales et éthiques ou modéliser des principes ou des théories éthiques issus de la philosophie (Q7) ainsi que vérifier que des agents respectent ces valeurs et principes (Q8) et puissent interagir avec des agents aux éthiques et morales différentes (Q9) ?

La table 2.4 indique comment les notions de fiabilité, d'honnêteté et d'éthique se croisent avec nos trois approches formelles afin de répondre, à chaque fois, à une de ces questions. Les lignes *Confiance*, *Coalitions* et *Cognition* réfèrent respectivement aux modèles de confiance et systèmes de réputation, aux jeux de coalitions et jeux hédoniques, modèles logiques, et aux architectures BDI.

Si les systèmes de confiance permettent classiquement d'évaluer la fiabilité ou une forme approchée d'honnêteté via la crédibilité, il semble pertinent de généraliser ces approches à des valeurs plus larges, comme le respect de valeurs éthiques. De plus, nous avons vu l'importance d'étudier la dynamique de la construction de la confiance. Nous présentons d'une part le travail réalisé au cours de la thèse de Thibaut Vallée sur un

7. Dans ce mémoire, nous avons fait le choix de ne présenter que des travaux réalisés en collaboration avec des doctorants ou des post-doctorants que nous avons encadrés.

modèle générique de système de réputation pour en étudier la dynamique de la confiance au chapitre 3 (Q2). D'autre part, nous présentons un modèle de crédibilité augmentant la robustesse des systèmes de réputation aux manipulations au chapitre 4 (Q5). Enfin, un travail sur le lien entre confiance et éthique a vu le jour au cours de la thèse de Nicolas Cointe et est présenté au chapitre 5 (Q8).

Concernant les modèles de formation de coalitions, nous présentons au chapitre 4 un travail toujours réalisé durant la thèse de Thibaut Vallée sur l'étude de la robustesse des jeux hédoniques aux manipulations, permettant de caractériser des conditions sous lesquelles ces jeux sont insensibles aux agents malhonnêtes (Q6). De plus, un autre travail réalisé en collaboration avec Thibaut Vallée est présenté au chapitre 5, proposant un modèle de jeux hédoniques pour des agents hétérogènes dans leurs mécanismes de formation de collectifs, représentant des valeurs éthiques (Q9). La question du lien entre coalitions et confiance (Q3) n'est pas abordée dans ce mémoire mais est présenté sous forme de sujet de thèse au chapitre 6.

Enfin, concernant les modèles d'agents cognitifs, nos travaux réalisés au cours de la thèse de Christopher Leturc et présentés au chapitre 4 proposent une logique modale pour représenter la confiance en la sincérité d'un agent (Q4) tandis que ceux réalisés au cours de la thèse de Nicolas Cointe (et présentés au chapitre 5) proposent une architecture BDI pour intégrer un modèle de raisonnement éthique dans un agent autonome (Q7). Là encore, le lien entre modèles cognitifs et fiabilité (Q1) n'est pas abordé dans ce mémoire mais est détaillé en tant que proposition de sujet de thèse au chapitre 6.

Deuxième partie

Présentation des activités de recherche

Chapitre 3

Premier axe : étude de la fiabilité

Sommaire

Ce chapitre est consacré à l'étude de la fiabilité dans les systèmes d'agents autonomes. Nous nous concentrons ici sur les systèmes de réputation. Dans la littérature, les études sur ces systèmes ne s'intéressent généralement pas à la manière dont les valeurs de réputation sont utilisées. Il convient donc de se poser une question : comment utiliser les valeurs de réputation pour décider avec qui interagir ? Pour traiter de cette question, nous proposons en section 1 un modèle générique de système de réputation fondé sur un modèle de bandits manchots. Cette modélisation nous permet de considérer en section 2 un ensemble de politiques d'utilisation des valeurs de réputation et, en section 3, un ensemble de manipulations inspirées de ce qui se fait dans la littérature de ce domaine. Nous étudions en section 4 l'influence de ses politiques sur la fiabilité de ces systèmes de réputation. Nous concluons ce chapitre par un bilan de l'animation et l'encadrement scientifique réalisés autour de ce travail.

1 Un modèle de bandit manchot

Usuellement, dans les systèmes de réputation, un agent demande à interagir avec l'agent ayant la meilleure réputation, même si certains auteurs proposent des heuristiques probabilistes [Kamvar *et al.*, 2003]. Ce problème de décision a été étudié dans un autre contexte, celui des bandits manchots (MAB) [Robbins, 1952]. La définition canonique d'un problème MAB est la suivante : considérons une machine à sous avec plusieurs bras, chacun ayant une fonction de gain suivant une loi de distribution *a priori* inconnue, quelle séquence de bras un agent doit-il tirer afin de maximiser son gain ?

1.1 Systèmes d'agents autonomes et bandits manchots

Si de nombreux modèles de MAB existent – à plusieurs joueurs [Liu et Zhao, 2010], à fonction de gain stationnaire ou non [Koulouriotis et Xanthopoulos, 2008], à possibilité de tirer plusieurs bras simultanément [Anantharam *et al.*, 1987], ou même avec adversaire [Auer *et al.*, 1995] – l'agent dispose dans tous les cas d'une politique de sélection lui permettant de minimiser son regret, c'est-à-dire la différence entre le gain obtenu et le gain qu'aurait eu l'agent si, à chaque pas de temps, il avait choisi le meilleur bras ; et toutes ces politiques – telles que UCB, Poker ou ε-glouton [Vermorel et Mohri, 2005, Auer et Ortner, 2010] – proposent des compromis entre l'exploitation et l'exploration.

Il nous semble alors pertinent de faire l'analogie entre les deux problèmes de décisions – la sélection d'agents évalués par un système de réputation et la sélection de bras évalués par une estimation de leur fonction de gain – pour proposer une méthodologie d'étude des systèmes de réputation. Ce lien a déjà été mis en évidence par [Awerbuch et Kleinberg, 2008] mais uniquement en se concentrant sur la question de la mise à jour des valeurs de confiance, tandis que nous proposons de ce servir de cette analogie pour étudier l'influence d'une politique de sélection afin de décider avec qui interagir.

Considérons un système d'agents autonomes, décrit par la définition 3.1, où chaque agent peut fournir des services et demander à d'autres de lui en fournir. Afin de ne pas perdre en généralité, nous considérons ces services comme abstraits et, lorsqu'un agent a besoin d'un service qu'il ne peut pas réaliser lui-même, il doit décider à quel autre agent demander de le lui fournir avec pour objectif de recevoir le service désiré avec la meilleure qualité possible. Ce problème est similaire à celui des bandits manchots et le tableau 3.2 résume cette analogie entre système d'agents autonomes et MAB.

Définition 3.1
Un système d'agents autonomes *est un tuple $\langle N, S\rangle$ où N est l'ensemble des agents et S l'ensemble des services qui peuvent être fournis. Notons par $N_x \subseteq N$ l'ensemble des agents capables de réaliser le service $s_x \in S$.*

Pour cela, considérons un joueur et une machine à sous ayant plusieurs bras. Chacun de ces bras est associé à une fonction de gain dont la loi de probabilité est *a priori* inconnue.

Système d'agents autonomes	
N :	Ensemble des agents
S :	Ensemble des services
N_x :	Ensemble des agents pouvant fournir le service s_x
$\varepsilon_{i,x}$:	Expertise de l'agent a_i pour le service s_x
v_i :	Fonction d'évaluation de l'agent a_i
π_i :	Politique de sélection de l'agent a_i
Système de réputation	
f_i :	Fonction de réputation de l'agent a_i,
$O_{i,k,x}$:	Ensemble des observations de l'agent a_i vis-à-vis de $\varepsilon_{k,x}$
$F_{i,j,k,x}$:	Témoignages de l'agent a_j fournis à a_i vis-à-vis de $\varepsilon_{k,x}$
$\mathcal{F}_i$:	Ensemble des témoignages et des observations de l'agent a_i

TABLE 3.1 – Récapitulatif des notations pour le chapitre 3

Le problème est alors de décider quelle séquence de bras tirer afin de maximiser le gain cumulé. Considérons un agent $a_i \in N$ et un service $s_x \in S$. L'agent a_i peut modéliser son problème de sélection de fournisseur par un bandit manchot m_x où il associe un bras $m_{x,k}$ à chaque agent $a_k \in N_x$. L'espérance de gain du bras $m_{x,k}$ (inconnue par a_i) représente la capacité de l'agent a_k à fournir le service s_x. Demander le service s_x à l'agent a_k correspond alors à tirer le bras $m_{x,k}$.

1.2 Intégration d'un mécanisme de réputation

Dans un bandit manchot tout comme dans le système d'agents autonomes, les expériences passées sont utilisées pour estimer la qualité d'un service futur (le gain) d'un agent (d'un bras) s'il est sélectionné. Dans le cadre des systèmes d'échange de services, les agents peuvent aussi échanger des informations et approximer l'estimation du gain espéré par une fonction de réputation. Sous hypothèse de corrélation entre réputation d'un agent et estimation de l'espérance de gain d'un bras, les témoignages du système de réputation sont les observations du bandit manchot.

Hypothèse 3.2
$\forall\ a_i, a_j, a_k \in N$ et $\forall\ s_x \in S$ si la réputation de a_j selon a_i pour le service s_x (notée $f_i(a_j, s_x) \in \mathbb{R}$) est supérieure à la réputation de a_k selon a_i pour le service s_x (notée $f_i(a_k, s_x) \in \mathbb{R}$) alors l'estimation du gain espéré du bras $m_{x,j}$ est supérieure que celle du bras $m_{x,k}$.

Cependant, certains agents, appelés agents malveillants, peuvent volontairement fournir des services de mauvaise qualité (par exemple fournir un virus). C'est pourquoi il nous faut considérer dans le MAB la présence d'adversaires qui choisissent le gain fourni par leurs bras [Auer *et al.*, 1995].

Nous définissons donc un agent comme :

	Système d'agents autonomes	**MAB**
Objectif	Maximiser la qualité des services reçus	Maximiser le gain
Acteurs	Consommateurs et fournisseurs	Joueurs et bandits
Interactions	Demander un service	Tirer un bras
Capacité	Expertise	Fonction de distribution des gains
Gain	Qualité d'un service	Gain d'un bras
Observations	Évaluation de la qualité	Observations passées
Communication	Témoignage sur un autre agent	Témoignage sur un bras
Réputation	Comportement futur espéré	Gain espéré
Politique de sélection	Déterminer le futur fournisseur de service	Déterminer le futur bras à utiliser
Manipulations	Agents malveillants	Adversaire

TABLE 3.2 – Analogie entre système d'agents autonomes et MAB

Définition 3.3
Un agent $a_i = \langle \overrightarrow{\varepsilon_i}, v_i, \mathcal{F}_i, f_i, \pi_i \rangle$ *est une entité autonome qui peut fournir et recevoir des services où* $\overrightarrow{\varepsilon_i} \in \mathbb{R}^{|S|}$ *désigne un vecteur d'expertise ;* v_i *une fonction d'évaluation ;* $\mathcal{F}_i$ *un ensemble de témoignages ;* f_i *une fonction de réputation ;* π_i *une politique de sélection.*

Pour un service $s_x \in S$, l'*expertise* de l'agent $a_k \in N_x$, notée $\varepsilon_{k,x} \in \mathbb{R}$, est sa capacité à fournir le service s_x avec une bonne qualité lorsqu'un autre agent le lui demande. Même si la qualité d'un service dépend de l'expertise de son fournisseur, elle est sujette à l'évalutation du demandeur. Cette évalutation peut être fondée sur de nombreux facteurs et est donc subjective. Par exemple, supposons que l'agent a_i demande à l'agent a_k de convertir un fichier `.doc` en fichier `.pdf` car il ne dispose pas d'une fonction `doc2pdf`. La qualité de ce service peut être fondée sur le fait de recevoir le fichier sans erreur d'encodage, mais aussi sur le temps mis par a_k pour le lui fournir.

Afin de rester général, nous considérons qu'un agent $a_i \in N$ ayant demandé le service $s_x \in S$ à l'agent $a_k \in N_x$ à l'instant t reçoit une *observation* $v_i(a_k, s_x, t) \in V_x$ où v_i désigne la fonction d'évaluation de l'agent a_i et V_x une échelle d'évaluation commune à l'ensemble des agents pour le service s_x. Pour simplifier la lecture, nous notons $v^t_{i,k,x}$ pour désigner $v_i(a_k, s_x, t)$. L'objectif d'un agent $a_i \in N$ qui désire le service $s_x \in S$ est de le recevoir avec la meilleure qualité possible. Pour cela, a_i peut utiliser ses observations sur le système afin de demander le service à un agent $a_k \in N_x$ en lequel il aurait le plus confiance. Dans toute la suite, nous notons $O_{i,k,x}$ l'ensemble des observations de l'agent a_i pour le service s_x fourni par a_k. Nous supposons *a priori* que les observations des agents sont sans erreur [1].

Comme il est possible que les agents aient individuellement peu d'observations sur les autres agents, ils peuvent partager ces dernières par le biais de témoignages. Un agent a_j peut fournir à un agent a_i à propos du service s_x rendu par a_k un *témoignage* noté

1. Le cas où les observations sont incertaines est en partie pris en compte lors de l'utilisation de la mesure de crédibilité présentée au chapitre 4 section 2.

$F_{i,j,k,x}$. Ce témoignage est équivalant aux observations de a_j ($F_{i,j,k,x} = O_{j,k,x}$) sauf en cas de manipulation comme présenté en section 3. Nous notons $\mathcal{F}_i$ l'union des observations de a_i et de l'ensemble des témoignages qu'il a reçu pour l'ensemble des services.

L'agent a_i peut alors utiliser $\mathcal{F}_i$ pour estimer l'expertise d'un agent pour un service donné. Cette estimation est la *réputation* de l'agent pour ce service. Nous supposons que chaque agent dispose d'une *fonction de réputation* $f_i : N \times S \times 2^{\mathcal{F}} \to \mathbb{R}$ (par abus de notation $\mathcal{F}$ est l'ensemble des témoignages possibles) qui calcule la réputation des agents. Cette fonction abstraite doit être instanciée et nous présentons en section 2.1 les différentes fonctions que nous considérons ici.

Un agent $a_i \in N$ qui désire le service $s_x \in S$ doit ensuite décider à quel autre agent le demander. La *politique de sélection* de l'agent a_i $\pi_i : S \to N$ permet à a_i de choisir un fournisseur pour le service s_x à partir des valeurs de réputation de tous les agents. Cette politique doit être instanciée et nous présentons en section 2.2 les différentes politiques que nous considérons ici.

Ainsi, l'architecture générale du système est résumée sur la figure 3.1. Débutant par un besoin de service (centre de la figure), chaque agent utilise sa politique de sélection π_i pour déterminer à quel agent le demander (flèche 1). Une fois ce service reçu (flèche 2), l'agent l'évalue et met à jour ses observations. Il fournit ensuite ces observations comme témoignages aux autres agents (flèche 5) et obtient à son tour des témoignages (flèche 6). La fonction de réputation f_i agrège les observations de l'agent avec les témoignages reçus afin d'obtenir une estimation de l'expertise des agents. Parallèlement, lorsqu'un agent reçoit une demande de service (flèche 3), il le fournit s'il en est capable (flèche 4).

2 Politiques d'utilisation de la confiance

Afin d'étudier comment des politiques de sélection et une notion de crédibilité influent sur la robustesse du système, nous devons dans un premier temps considérer différentes instanciations de la fonction de réputation, puis définir quelles sont ces politiques de sélection.

2.1 Modélisation des fonctions de réputation

Comme vu au chapitre 2 section 1.1, il existe de nombreuses fonctions de réputation. Nous nous intéressons à cinq fonctions différentes. La première n'est pas véritablement une fonction de réputation mais une fonction de décision classique dans les problèmes de bandits manchots que nous appellons *estimation personnelle*. Cette dernière consiste à ne pas prendre en compte les témoignages et à considérer la réputation d'un agent comme le gain moyen des services qu'il a fournis. Cela revient à condiérer une fonction de sélection qui ne s'appuie pas sur la réputation et qui nous sert de référence.

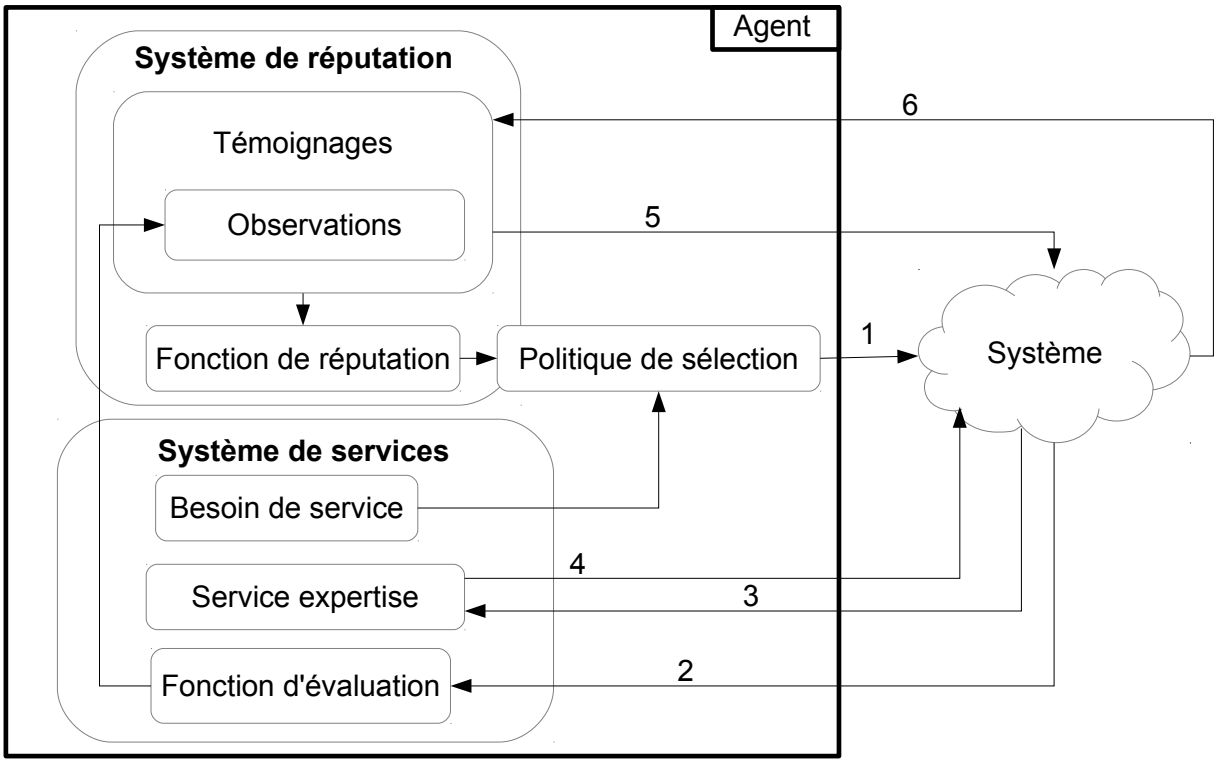

FIGURE 3.1 – Architecture schématique d'un système d'agents autonomes

Définition 3.4
*L'*estimation personnelle *de a_i est la fonction de réputation $f_i(a_k, s_x, \mathcal{F}_i) = \mu_{i,k,x}$ où $\mu_{i,k,x}$ est la moyenne de $O_{i,k,x}$, les observations de l'agent a_i vis-à-vis de $\varepsilon_{k,x}$.*

Nous considérons également une fonction de réputation symétrique, que nous appellons *estimation collective*, et qui consiste à prendre en compte tous les témoignages et définir la réputation comme le gain moyen calculé à partir des observations et des témoignages.

Définition 3.5
*L'*estimation collective *de a_i est la fonction de réputation $f_i(a_k, s_x, \mathcal{F}_i) = \mu_{N,k,x}$ où $\mu_{N,k,x}$ désigne la moyenne des observations présentes dans l'ensemble $\bigcup_{a_j \in N} F_{i,j,k,x}$.*

Nous considérons enfin les trois fonctions de réputation présentées en détail au chapitre 2 : EigenTrust, BetaReputation et FlowTrust. Ces trois fonctions sont fondées sur un *graphe de confiance* calculé à partir des témoignages : un graphe orienté dont les arcs (a_i, a_j) représentent le fait qu'il y ait eu au moins un service fourni à a_i par a_j et sont étiquettés par les observations de a_i envers a_j. Les définitions formelles de ces systèmes, instanciés dans le cadre de notre modèle, sont les suivantes.

— Dans EigenTrust, la réputation d'un agent est la probabilité qu'un marcheur aléatoire sur le graphe de confiance atteigne le nœud qui lui est associé où la probabilité de transition d'un nœud à l'autre est proportionnelle à la différence entre les bons services et les mauvais services fournis.

Définition 3.6
Soit $\mathbb{C}$ la matrice d'adjacence d'un graphe de confiance. La fonction EigenTrust *de a_i est : $f_i(a_k, s_x, \mathcal{F}_i) = (1-a)(\mathbb{C}^T)^n \vec{c_i} + a p_k$, où $a \in [0,1]$ est un facteur d'exploration, $\vec{c_i}$ le vecteur normalisé de confiance de a_i dans lequel $c_{i,k,x} = \max(0, r_{i,k,x} - s_{i,k,x})$ où r_{ij} (respectivement s_{ij}) est le nombre d'évaluation positives (respectivement négatives) d'un service s_x de l'agent a_k par un agent a_i et p_k une valeur de confiance a priori envers a_k.*

— BetaReputation considère la réputation d'un agent comme une espérance de gain modélisée par une loi de bêta densité. C'est une approche assymétrique où chaque agent communique à ses voisins dans le graphe de confiance la réputation qu'il a calculée et pondère les témoignages reçus par la réputation des agents qui les fournissent.

Définition 3.7
Soit $\mathbb{P}_i$ l'ensemble des chemins entre a_i et a_k sur un graphe de confiance, $r_{i,k,x}$ (resp. $s_{i,k,x}$) le nombre d'évaluation positives (respectivement négatives) d'un service s_x de l'agent a_k par un agent a_i. La fonction BetaReputation *de a_i est $f_i(a_k, s_x, \mathcal{F}_i) = \frac{r_{i,k,x} - s_{i,k,x}}{r_{i,k,x} + s_{i,k,x} + 2}$ où :*

$$r_{i,k,x} = \sum_{P \in \mathcal{P}_{i,k}} \prod_{(a_j, a_{j'}) \in P} \frac{2 r_{j,j',x} r_{j',k,x}}{(s_{j,j',x} + 2)(r_{j',k,x} + s_{j',k,x} + 2) + 2 r_{j,j',x}}$$

$$s_{i,k,x} = \sum_{P \in \mathcal{P}_{i,k}} \prod_{(a_j, a_{j'}) \in P} \frac{2 r_{j,j',x} s_{j',k,x}}{(s_{j,j',x} + 2)(r_{j',k,x} + s_{j',k,x} + 2) + 2 r_{j,j',x}}$$

— Dans FlowTrust, la réputation d'un agent a_k selon un agent a_i est le flot maximum de a_i vers a_k sur le graphe de confiance où la capacité d'un arc (a_i, a_j) est la moyenne des gains $\mu_{i,k,x}$ de a_i.

Définition 3.8
Soit $\mathcal{P}_{i,k}$ l'ensemble des chemins disjoints *entre a_i et a_k sur un graphe de confiance et $\mu_{i,k,x}$ la moyenne des $O_{i,k,x}$. La fonction* FlowTrust *de a_i est : $f_i(a_k, s_x, \mathcal{F}_i) = \sum_{P \in \mathcal{P}_{i,k}} \min\{\mu_{j,j',x} | (a_j, a_{j'}) \in P\}$.*

2.2 Adaptation des politiques classiques

Si la réputation d'un agent est une estimation du gain espéré lors des futures interactions, il convient de définir comment un agent l'utilise pour maximiser son gain.

Définition 3.9
Soit $a_i \in N$ un agent désirant recevoir le service $s_x \in S$. La politique de sélection Π_i *définit à quel agent $a_k \in N_x$ demander ce service.*

Nous proposons ici d'utiliser les politiques canoniques des bandits manchots comme politiques de sélection. Ces politiques permettent aux agents d'interagir majoritairement avec les fournisseurs de services ayant une bonne réputation (puisqu'ils sont supposés être ceux maximisant l'espérance de gain) et d'interagir occasionnellement avec les autres agents afin de vérifier si leur mauvaise réputation n'est pas due à un manque d'observations. Ce sont donc des compromis entre l'exploitation des connaissances des agents et l'exploration du système permettant d'affiner ces connaissances. Nous adaptons ici deux d'entre elles, UCB [2] et la politique ε-gloutonne, et en proposons une troisième : l'ε-élitisme. Il existe bien entendu d'autres politiques mais nous considérons celles-ci car UCB est une des politiques les plus performantes, ε-glouton est une politique naïve qui sert classiquement de point de comparaison et l'ε-élitisme est la politique implicitement utilisée dans les systèmes de réputation.

La première est l'une des plus utilisées dans le cadre des bandits manchots. En effet, UCB permet de borner le regret des agents (différence entre le gain total obtenu et le gain maximum si le meilleur bras avait toujours été sélectionné) [Robbins, 1952, Auer *et al.*, 1995]. Dans le cadre des systèmes de réputation, UCB consiste à demander un service à l'agent maximisant la réputation à laquelle s'ajoute un facteur d'exploration. Ce facteur a pour objectif d'inciter les agents à interagir avec ceux sur lesquels il dispose de peu d'information. UCB garantit ainsi un minimum d'interactions avec chaque fournisseur afin que la fonction de réputation retourne une estimation suffisament précise de l'expertise. Il

2. Upper Confidence Bound.

est important de noter que ceci permet de maintenir la propriété d'ouverture du système en incitant à interagir avec tout nouvel arrivant.

Définition 3.10
L'agent $a_i \in N$ désirant le service $s_x \in S$ suit la politique UCB *s'il sélectionne l'agent $a_k \in N_x$ qui maximise :*

$$f_i(a_k, s_x, \mathcal{F}_i) + \sqrt{\frac{2\ln(1+n_x)}{1+n_{k,x}}}$$

où $n_{k,x}$ est le nombre de fois que l'agent a_k a fourni le service s_x à a_i et n_x le nombre de fois que a_i a reçu le service s_x.

La politique ε-gloutonne consiste à demander le service désiré à l'agent capable de le fournir avec la meilleure valeur de réputation, tout ayant une certaine probabilité d'explorer uniformément le système [Auer *et al.*, 2002]. Ce facteur d'exploration est la probabilité de sélectionner aléatoirement uniformément un autre agent que celui maximisant la valeur de réputation. Comme UCB, la politique ε-gloutonne garantit une propriété d'ouverture du système mais, contrairement à UCB, sans pour autant avantager les nouveaux entrants.

Définition 3.11
L'agent $a_i \in N$ suit une politique ε-gloutonne *s'il demande le service $s_x \in S$ à l'agent $a_k \in N_x$ qui maximise $f_i(a_k, s_x, \mathcal{F}_i)$ avec une probabilité $1-\varepsilon$ ou, avec une probabilité ε, sélectionne aléatoirement uniformément a_k dans N_x.*

Nous proposons une troisième politique appelée ε-élitisme. Un agent suivant cette politique sélectionne le futur fournisseur de services uniformément parmi les $\lceil \varepsilon \times |N_x| \rceil$ agents de N_x ayant les meilleures valeurs de réputation. Contrairement aux politiques précédentes, la politique ε-élitiste ne respecte pas la propriété d'ouverture du système car il n'y a pas de facteur d'exploration permettant d'éventuellement sélectionner des nouveaux entrants mais elle permet de ne pas surcharger de demandes l'agent ayant la meilleure réputation. Remarquons que dans la littérature, les systèmes de réputation appliquent classiquement une politique purement élitiste (soit $\frac{1}{|N_x|}$-élitiste).

Définition 3.12
Soit $a_i \in N$ un agent désirant recevoir le service $s_x \in S$. Soit $N_{x,\varepsilon} \subseteq N_x$ tel que $|N_{x,\varepsilon}| = \lceil \varepsilon \times |N_x| \rceil$ et que $\forall a_j \in N_{x,\varepsilon}, \nexists a_k \in N_x \setminus N_{x,\varepsilon} : f_i(a_j, s_x, \mathcal{F}_i) < f_i(a_k, s_x, \mathcal{F}_i)$. L'agent a_i suit une politique ε-élitiste *s'il sélectionne aléatoirement uniformément a_k dans $N_{x,\varepsilon}$.*

3 Modélisation des manipulations

L'utilisation de systèmes de réputation a pour objectif de garantir aux agents de recevoir les services avec la meilleure qualité. Cependant, dans un système ouvert, certains agents malveillants peuvent avoir comme objectif de fournir des services de mauvaise qualité. Par exemple, dans un système pair-à-pair d'échange de fichiers tel que Gnutella [Ripeanu, 2001], un agent malveillant peut chercher à propager des virus.

Définition 3.13
Soit un agent malveillant $a_j \in N$ ayant une expertise $\varepsilon_{j,x}$. a_j fournit un bon service s_x s'il le fournit avec une qualité correspondante à son expertise. a_j fournit un mauvais service s_x s'il le fournit avec une qualité $\min(V_x)$.

Si un tel agent malveillant peut être détecté par des systèmes de réputation, plusieurs agents malveillants peuvent former une coalition afin de manipuler le système. Par ailleurs, un agent malveillant seul peut s'introduire dans le système sous de multiples fausses identités (appelées agents Sybil [Douceur, 2002]) et ainsi former une coalition malveillante. Nous considérerons ici les trois types de manipulations classiquement utilisés dans la littérature : les faux témoignages, le blanchiment et l'attaque oscillante [Hoffman *et al.*, 2009].

3.1 Manipulations individuelles

Comme la réputation des agents est fondée sur l'utilisation de témoignages, une manipulation consiste à fournir de faux témoignages. Nous considérons deux types de faux témoignages : la promotion et la diffamation. La première consiste à fournir des témoignages afin d'augmenter artificiellement la valeur de réputation d'un agent. À l'inverse, la diffamation a pour objectif de diminuer la réputation de l'agent. Dans les deux cas, si la manipulation est efficace, les agents malveillants apparaîtront comme les meilleurs fournisseurs, leur permettant ainsi de fournir de mauvais services.

Définition 3.14
Soit un agent malveillant $a_j \in N$. Soit un service $s_x \in S$ et deux agents $a_i \in N$ et $a_k \in N_x$. L'agent a_j fournit un faux témoignage *à l'agent a_i vis-à-vis de l'expertise de a_k pour le service s_x s'il lui communique des témoignages $F_{i,j,k,x}$ tels que $F_{i,j,k,x} \neq O_{j,k,x}$. Soit $\mu_{j,k,x}$ la moyenne des véritables observations de a_j (fondée sur $O_{j,k,x}$) et $\mu_{i,j,k,x}$ la moyenne des observations fournies en témoignage (fondée sur $F_{i,j,k,x}$).*

- *si $\mu_{j,k,x} < \mu_{i,j,k,x}$ alors l'agent a_j* promeut *a_k,*
- *si $\mu_{j,k,x} > \mu_{i,j,k,x}$ alors a_j* diffame *a_k.*

Comme nous considérons un système multi-agent ouvert, si un agent malveillant a une valeur de réputation trop faible, il lui est possible de quitter le système pour revenir sous une autre identité. Cette manipulation, appelée blanchiment, a pour objectif de réinitialiser la réputation de l'agent en obtenant la même réputation qu'un nouvel agent qui vient de joindre le système pour la première fois.

Définition 3.15
Soit un agent malveillant $a_j \in N$. a_j effectue un blanchiment *s'il quitte le système pour le rejoindre sous une autre identité a'_j.*

Nous considérons ici que les agents malveillants peuvent changer d'identité à volonté. Notons qu'il est difficile d'empêcher une telle manipulation tout en satisfaisant la propriété d'ouverture du système. Cependant, certaines approches telles que l'utilisation de *captchas*, de frais d'inscription ou de puzzles cryptographiques permettent de rendre le blanchiment coûteux [Borisov, 2006].

3.2 Manipulations collectives

Si la promotion, la diffamation et le blanchiment peuvent être effectués par un agent seul en un court instant, une coalition d'agents malveillants peut également manipuler le système d'agents autonomes sur le long terme. L'une de ces manipulations est l'attaque oscillante. Dans une telle manipulation, la coalition d'agents malveillants est partitionée en deux sous-ensembles M_1 et M_2. Ces sous-groupes ont alors un comportement coordonné. Les agents du premier groupe fournissent des services de bonne qualité afin de bénéficier d'une bonne valeur de réputation. Dans le même temps, ils promeuvent les agents du second groupe afin d'accroître la réputation de ces derniers. Les agents du second groupe fournissent quant à eux volontairement de mauvais services et diffament les agents honnêtes.

Lorsque la réputation d'un agent du second groupe tombe en dessous de celle d'un des agents du premier groupe, ils échangent leurs rôles : l'agent de M_1 fournit désormais les mauvais services et diffame tandis que celui du groupe M_2 fournit de bons services et promeut. Notons que l'attaque oscillante peut être combinée avec du blanchiment au moment où les agents de M_1 et de M_2 changent de rôles.

Définition 3.16
Soit $M \subset N$ une coalition d'agents malveillants. Soit M_1 et M_2 un partitionnement de M. La coalition M effectue une attaque oscillante *en appliquant la stratégie suivante :*

- *les agents de M_1 promeuvent ceux de M_2,*
- *les agents de M_2 diffament les agents de $N \setminus M$,*
- *les agents de M_1 fournissent les services en fonction de leur expertise,*
- *les agents de M_2 fournissent volontairement de mauvais services,*
- *si la réputation d'un agent de M_2 est inférieure à celle d'un agent de M_1, l'agent de M_2 se blanchit et intègre M_1 tandis que l'agent de M_1 intègre M_2.*

4 Résultats expérimentaux

Afin d'évaluer les politiques de sélection face aux manipulations, nous considérons deux mesures : le regret des agents honnêtes et le coût de la manipulation. Le regret d'un agent est une mesure classique dans les problèmes de bandits manchots. Intuitivement, le regret d'un agent désigne la différence entre le gain qu'il aurait pu gagner s'il avait toujours

demandé les services aux meilleurs fournisseurs et le gain qu'il a réellement obtenu en suivant sa politique de sélection. Le regret des agents peut donc être vu comme une mesure d'efficacité du système puisque minimiser le regret et maximiser le gain des agents sont équivalents.

Définition 3.17
Soit un agent $a_i \in N$ ayant un ensemble d'observations $O_i = \{v^1_{i,k_1,x_1}, \ldots, v^n_{i,k_n,x_n}\}$. Soit $\varepsilon^t_{\star,x}$ l'expertise du meilleur fournisseur du service s_x à l'instant t. Le regret *de a_i est donné par :*

$$r_i = \sum_{t=1}^{n} \varepsilon^t_{\star,x_t} - v^t_{i,k_t,x_t}$$

Comme certaines manipulations telles que l'attaque oscillante impliquent que les agents malveillants fournissent parfois de bons services (ce qui est contraire à leur objectif) afin de maintenir une haute valeur de réputation, nous considérons le coût de la manipulation comme le ratio de bons services fournis par les agents malveillants sur l'ensemble des interactions passées.

Définition 3.18
Soit $M \subset N$ une coalition d'agents malveillants. Soit $n_{i,k,x}$ le nombre de fois que l'agent a_k a fourni le service s_x à l'agent a_i et $n^+_{i,k,x}$ le nombre de fois où l'agent a_k a fourni le service s_x avec une bonne qualité. Le coût *de la manipulation est donné par :*

$$\mathfrak{C} = \frac{\sum_{a_k \in M} \sum_{s_x \in S} n^+_{i,k,x}}{\sum_{a_k \in N} \sum_{s_x \in S} n_{i,k,x}}$$

Nous considérons dans nos expérimentation un système d'agents autonomes $\langle N, S \rangle$ où initialement $|N| = 100$ et $|S| = 10$. Parmis les 100 agents, 10 sont considérés comme appartenant à une même coalition malveillante et ils appliquent une attaque oscillante. L'expertise des agents pour un service est tirée aléatoirement uniformément entre 0 et 1, chaque agent pouvant fournir entre 0 et 5 services. La qualité des services est évalué sur l'intervalle $[-1, 1]$. Nous considérons le temps comme discret. À chaque pas de temps, les agents demandent un service qu'ils ne peuvent pas fournir eux-même. Afin de simplifier notre étude, nous supposons que chaque service est fourni en un pas de temps et qu'un agent peut fournir simultanément autant de services que demandé. À chaque pas de temps, un nouvel agent peut rejoindre le système ou le quitter avec une probabilité de $0,01$.

Dans nos simulations, nous considérons que nos agents n'ont *a priori* aucune connaissance initiale et vont interagir durant 200 pas de temps. Nous réitérons ces simulations 50 fois et calculons la moyenne des métriques présentées précédemment. Dans ces simulations, nous considérerons trois politiques de sélection : UCB, $0,1$-gloutonne et $0,1$-élitiste

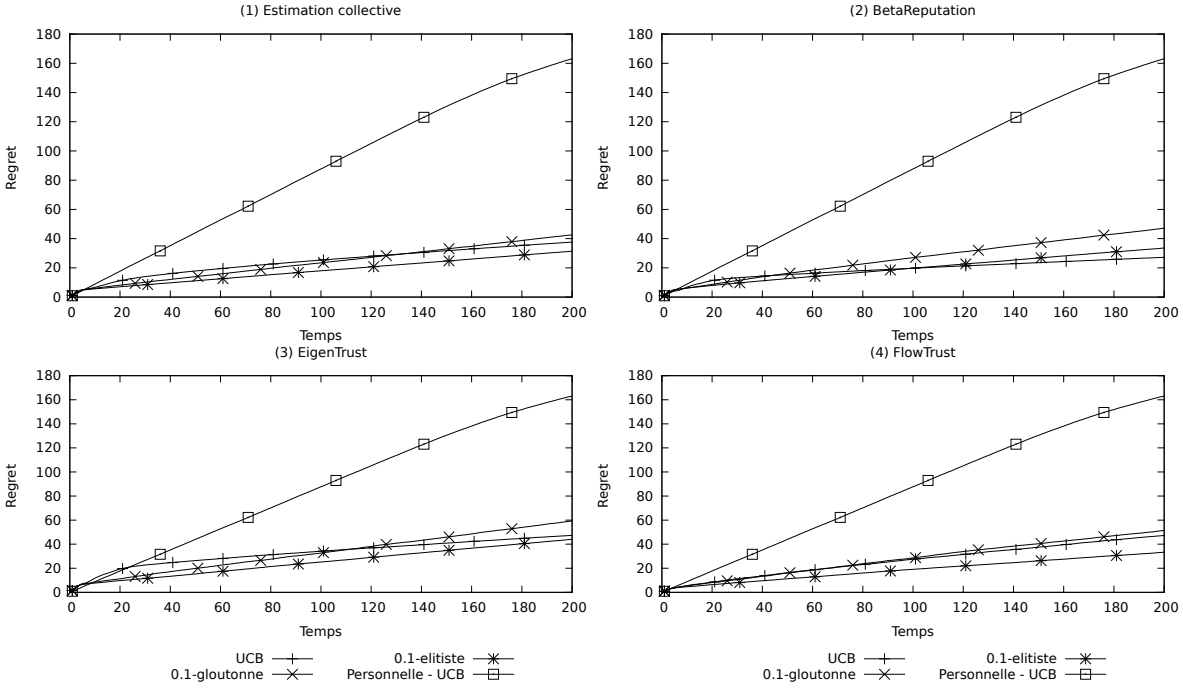

FIGURE 3.2 – Regret en l'absence de manipulation

que nous appliquons sur l'estimation collective, BetaReputation, EigenTrust et FlowTrust. Nous comparons ces résultats avec l'estimation personnelle utilisant UCB, ce qui correspond à un problème classique de bandits manchots. Intuitivement, l'estimation collective doit être très sensible aux manipulations alors que l'estimation personnelle (n'utilisant aucun témoignage) n'est sensible qu'aux changements de comportements.

4.1 Regret des systèmes de réputation

La figure 3.2 nous montre l'intérêt des agents à coopérer en l'absence de manipulation dans le système. Nous considérons ici que les agents malveillants fournissent uniquement des mauvais services mais n'appliquent pas d'attaque oscillante. Indépendamment de la politique de sélection utilisée, l'échange d'information permet aux agents d'avoir un regret très bas contrairement à l'estimation personnelle qui nécessite que les agents explorent chaque fournisseur, ce qui leur confère un regret important. La politique $0,1$-élitiste est celle qui minimise le regret sur l'estimation collective (figure 3.2.1) et FlowTrust (figure 3.2.4). UCB devient rapidement la politique qui minimise le regret sur BetaReputation (figure 3.2.2). Enfin, si UCB est initialement la politique de sélection la moins performante sur EigenTrust (figure 3.2.3), la vitesse de croissance de son regret devient rapidement quasi-nulle, tendant ainsi à minimiser le regret. Dans les quatre cas, la politique $0,1$-gloutonne est

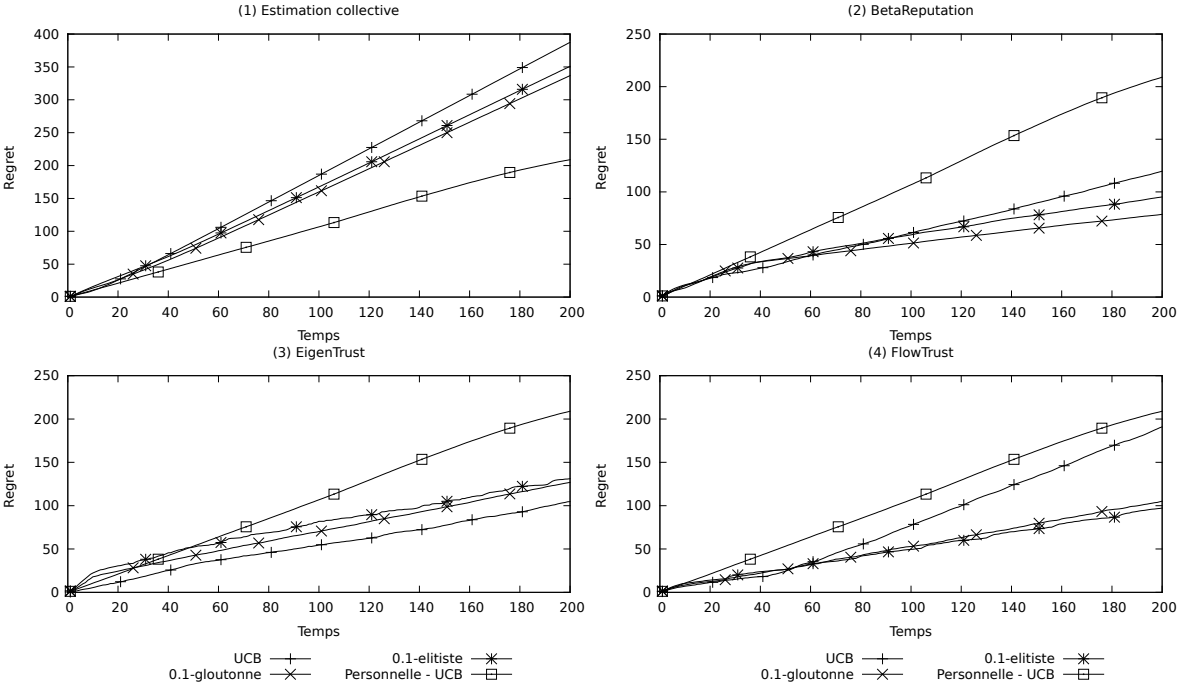

FIGURE 3.3 – Regret en présence de manipulations

celle qui fournit le plus haut regret. En effet, le facteur d'exploration de cette politique amène souvent les agents à interagir avec des agents pourtant identifiés comme ayant une faible expertise.

La figure 3.3 présente le même système en présence de manipulations. Ainsi, le regret des agents est globalement plus important (y compris avec l'estimation personnelle puisque les agents malveillants affectuent une attaque oscillante). Ici, l'estimation collective (figure 3.3.1) est si peu robuste que toutes les politiques de sélection tendent à interagir avec les agents malveillants. Alors que UCB est la politique la plus efficace dans les systèmes de réputation triviaux, elle dégrade largement les performances de BetaReputation et FlowTrust (figures 3.3.2 et 3.3.4) en raison de son facteur d'exploration qui incite à interagir avec les agents qui viennent d'effectuer un blanchiment. À l'inverse, EigenTrust (figure 3.3.3) qui est le système de réputation le plus sensible aux manipulations est plus performant avec UCB : le facteur d'exploration va permettre aux agents honnêtes diffamés d'être tout de même sélectionnés. Remarquons que la politique $0,1$-gloutonne qui est la moins performante en l'absence de manipulation est ici celle qui minimise le regret sur BetaReputation.

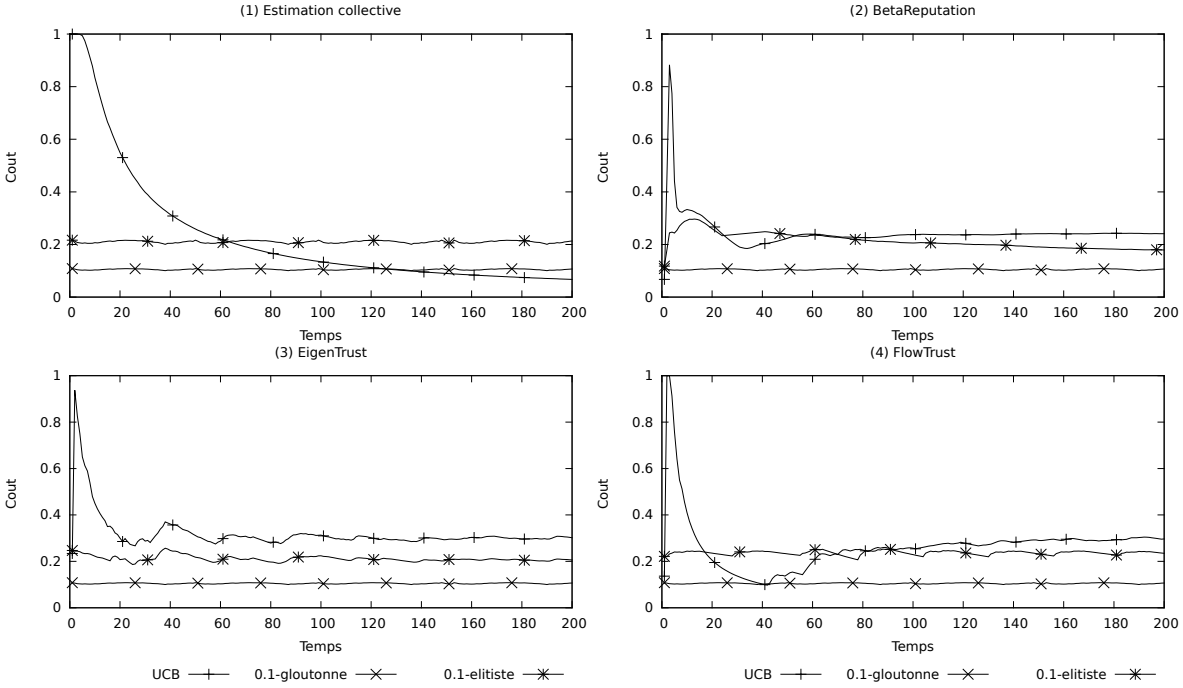

FIGURE 3.4 – Coût de la manipulation

4.2 Coût des manipulations

Si manipuler les systèmes de réputation permet aux agents malveillants de fournir de mauvais services, ceux-ci fournissent également des services de bonne qualité afin de maintenir leurs valeurs de réputation. La figure 3.4 nous montre ce coût (définition 3.18). Nous pouvons constater qu'avec UCB, le coût initial est très élevé puis chute rapidement. En effet, les agents honnêtes commencent par explorer et interagissent avec les agents malveillants qui ne sont pas promus et qui fournissent de bons services puis, très rapidement, le facteur d'exploration s'annule et les agents vont interagir avec les agents malveillants qui fournissent de mauvais services. BetaReputation, EigentTrust et FlowTrust (figures 3.4.2, 3.4.3 et 3.4.4) présentent des hausses occasionnelles du coût de la manipulation dues aux blanchiments. Avec les politiques de sélection $0,1$-gloutonne et $0,1$-élitiste, le coût de la manipulation est globalement constant pour toutes les politiques : comme les agents malveillants fournissant de bons services ne sont pas diffamés, ils appartiennent aux 10 % des meilleurs fournisseurs des services et peuvent donc être sélectionnés par le facteur d'exploration.

Nous pouvons donc conclure que les politiques de sélections ont une forte influence sur l'efficacité des fonctions de réputation. En l'absence de manipulation, UCB est la politique qui minimise le regret mais elle rend les fonctions de réputation beaucoup plus sensibles

au blanchiment et au changement de comportement. Cependant, de telles manipulations ont un fort coût puisque les agents malveillants doivent fournir des bons services après blanchiment. À l'inverse, la politique $0,1$-gloutonne qui est la moins efficace en l'absence de manipulation devient la plus performante en présence de manipulations. En effet, son facteur d'exploration lui permet d'interagir avec des agents honnêtes même s'ils ont une faible réputation due aux diffamations. Malgré les manipulations, les trois politiques appliquées sur BetaReputation, EigenTrust et FlowTrust donnent un regret inférieur à celui de l'estimation personnelle. L'estimation collective, elle, reste toujours sensible aux faux témoignages. Cependant, il convient de remarquer que l'influence des politiques diffèrent selon la fonction de réputation utilisée. Ainsi, l'étude de l'influence des paramètres de ces fonctions (comme l'ajout d'un facteur d'oubli pour BetaReputation ou une variation des confiances *a priori* d'EigenTrust) serait pertinente pour généraliser complètement nos résultats.

Bilan et animation scientifique

Nous avons présenté dans ce chapitre un exemple de travaux réalisés dans le cadre de l'axe de recherche sur la fiabilité des agents autonomes. Cet axe s'est construit à partir de premiers travaux (non présentés dans ce mémoire) sur la robustesse des systèmes de réputation où nous nous sommes intéressés à des modèles fondés sur la théorie des jeux. Ces travaux initiaux ont conduit à une publication internationale [Bonnet, 2012] et deux publications nationales [Bonnet, 2013, Bonnet, 2014], et nous avons été invité en 2013 par l'équipe DESIR à donner un séminaire au LIP6 afin de les présenter.

Les limites de ces travaux initiaux nous ont conduit à proposer un modèle générique de système de réputation qui, en s'inspirant des modèles de bandits manchots, permet de prendre en considération les différents composants d'un tel système : les observations des agents, leur agrégation mais surtout le processus de prise de décision. Ce modèle a résulté du co-encadrement de la thèse de Thibaut Vallée (2012 – 2015) sous la direction de François Bourdon (Université de Caen Normadie) et a fait l'objet de plusieurs publications internationales [Vallée et Bonnet, 2015, Vallée *et al.*, 2014b] et nationales [Vallée *et al.*, 2015, Vallée *et al.*, 2014a]. De plus, nous nous sommes appuyés sur ce modèle dans des travaux ultérieurs. Nous nous en sommes servi afin de caractériser et étudier une notion de crédibilité des agents (présentée au chapitre suivant). Ce modèle a aussi permis une collaboration au niveau régional et a été utilisé pour le stage de master de Damien Lelerre en 217, que nous avons co-encadré avec Laurent Vercouter (INSA Rouen). Ce stage a eu pour objectif de proposer un système de réputation à témoignages confidentiels et d'étudier l'influence de mécanismes d'anonymisation sur la robustesse du système en fonction de la politique de sélection.

Ce travail sur la fiabilité des systèmes de réputation nous a naturellement amené à nous interroger sur les raisons qui font que ces systèmes peuvent être mis en défaut, et en particulier lorsque cela est dû à la présence d'agents menteurs ou manipulateurs. Il s'avère que traiter de la fiabilité des agents autonomes ne peut se faire sans penser la question des agents malhonnêtes, ce qui a conduit à développer un second axe de recherche autour de l'étude de l'honnêteté que nous présentons dans le chapitre suivant.

Chapitre 4

Second axe : étude de l'honnêteté

Sommaire

Les interactions entre agents autonomes sont généralement régies par des règles formant un protocole. Forcés de respecter ce protocole, certains agents insatisfaits peuvent alors se comporter de manière malhonnête – en propagant de fausses informations, usurpant l'identité d'un autre ou interceptant des communications – afin d'en tirer profit. Cela n'aurait peut-être que peu d'importance si ces comportements n'étaient pas au détriment des autres agents. Pour garantir à ces derniers qu'ils peuvent interagir sans risque, il est important de définir des stratégies de défense et, en particulier, être capable de détecter et raisonner sur les informations transmises par les autres agents qui peuvent parfois être mensongères. Pour traiter ces questions, nous proposons en section 1 un système logique permettant de raisonner sur la sincérité des agents, en section 2 un mécanisme permettant d'évaluer la crédibilité des agents dans un système de réputation et enfin, en section 3, une caractérisation des manipulation dans les jeux hédoniques. Nous concluons ce chapitre par un bilan de l'animation et l'encadrement scientifique réalisés autour de ce travail.

1 Sincérité d'un discours

Dans le contexte du raisonnement sur la confiance, la plupart [1] des approches reposent sur des logiques modales [Dundua et Uridia, 2010, Herzig *et al.*, 2010, Singh, 2011, Smith *et al.*, 2011]. Nous avons vu au chapitre 2 section 3.2 que ces travaux s'intéressaient essentiellement dans la confiance en la fiabilité des actions d'un agent et les quelques travaux qui s'intéressent non pas aux actions mais aux discours des agents se concentrent eux-aussi sur la question de la fiabilité [Liau, 2003, Demolombe, 2004, Dastani *et al.*, 2004]. Dans le contexte de notre travail sur l'honnêteté et en nous fondant sur la définition *free of deceit, truthful and sincere*, nous proposons ici une logique modale permettant d'exprimer la *confiance en la sincérité* accordée par un agent a_i à propos d'un énoncé ϕ d'un autre agent a_j. La caractéristique principale de cette logique est de lier une modalité de confiance avec les croyances de l'agent cible – un agent est sincère s'il croit ce qu'il énonce – et d'en faire une confiance non transitive : ce n'est pas parce qu'un agent a_i a confiance en la sincérité d'un agent a_j lorsque ce dernier énonce sa confiance en la sincérité d'un agent a_k que l'agent a_i doit avoir confiance dans la sincérité de a_k.

1.1 Une logique normale de la confiance

La logique que nous proposons – appelée système TB – repose sur deux modalités : une modalité de croyance B_i où $B_i\phi$ signifie que l'agent a_i croit que ϕ est vraie, et une modalité de confiance en la sincérité $T^s_{i,j}$ où $T^s_{i,j}\phi$ signifie que l'agent a_i a confiance en la sincérité de a_j à propos de la proposition ϕ. Le système TB s'appuie sur un langage $\mathcal{L}_{T,B}$ avec $\mathcal{P} = \{a, b, c, \ldots\}$ un ensemble de symboles propositionnels, $\mathcal{N} = \{a_1 \ldots a_n\}$ un ensemble d'agents tel que $a_i, a_j, a_k \in \mathcal{N}$ et $p \in \mathcal{P}$ une variable propositionnelle. Nous considérons la règle BNF suivante :

$$\psi ::= p \mid \neg\psi \mid \psi \wedge \psi \mid \psi \vee \psi \mid \psi \Rightarrow \psi \mid T^s_{i,j}\psi \mid B_i\psi$$

Remarquons que $B_i\phi$ n'est pas équivalent à $T^s_{i,i}\phi$ car cette dernière formule signifie que l'agent a_i a confiance en sa sincérité pour ϕ. De plus, contrairement à [Liau, 2003] et [Demolombe, 2004] qui modélisent une confiance dispositionnelle, nous ne considérons pas de modalité explicite d'acquisition de l'information ou de communication. Par exemple, Liau considère qu'un agent a confiance dans un autre s'il peut croire l'énoncé futur de ce dernier. Notre modalité $T^s_{i,j}$ est une modalité de confiance occurente signifiant qu'elle ne prend son sens que dans l'instant présent : si un agent a confiance en la sincérité d'un autre alors cela signifie qu'il croit que ce dernier croit ce qu'il énonce. Nous considérons que lorsqu'un agent a_i a confiance dans la sincérité de a_j alors il a déjà acquis les informations lui permettant de déduire si a_j est sincère ou non.

Nous définissons un cadre de Kripke $\mathcal{C} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}^s_{i,j}\}_{i,j\in\mathcal{N}})$ associé à $\mathcal{L}_{T,B}$ où :

1. Bien entendu, d'autres approches ont été très étudiées comme l'utilisation des logiques floues [Demolombe et Liau, 2001, Falcone *et al.*, 2002, Wang et Huang, 2007, Kant et Bharadwaj, 2013].

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

— $\{\mathcal{B}_i\}_{i\in\mathcal{N}}$ est un ensemble de relations binaires telles que :

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

— $\{\mathcal{T}^s_{i,j}\}_{i,j\in\mathcal{N}}$ est un ensemble de relations binaires telles que :

$$\forall a_i, a_j \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{T}^s_{i,j}(w) := \{v \in \mathcal{W} | w\mathcal{T}^s_{i,j} v\}$$

Notre modèle de Kripke est défini comme étant $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{T}^s_{i,j}\}_{i,j\in\mathcal{N}}, \iota)$ avec $\iota : \mathcal{P} \to 2^{\mathcal{W}}$ une fonction d'interprétation. Pour chaque monde $w \in \mathcal{W}$, pour chaque $\phi, \psi \in \mathcal{L}_{T,B}$ et pour chaque $p \in \mathcal{P}$:

1. $w \models \top$
2. $w \not\models \bot$
3. $w \models p$ si, et seulement si, $w \in i(p)$
4. $w \models \neg\phi$ si, et seulement si, $w \not\models \phi$
5. $w \models \phi \vee \psi$ si, et seulement si, $w \models \phi$ ou $w \models \psi$
6. $w \models \phi \wedge \psi$ si, et seulement si, $w \models \phi$ et $w \models \psi$
7. $w \models \phi \Rightarrow \psi$ si, et seulement si, $w \models \neg\phi$ ou $w \models \psi$
8. $w \models B_i\phi$ si, et seulement si, $\forall v \in \mathcal{W} : w\mathcal{B}_i v, v \models \phi$
9. $w \models T^s_{i,j}\phi$ si, et seulement si, $\forall v \in \mathcal{W} : w\mathcal{T}^s_{i,j} v, v \models \phi$

Remarquons que B_i est une modalité $\Box$ classique comme dans [Liau, 2003, Demolombe, 2004, Herzig *et al.*, 2010]. Concernant la modalité de confiance, nous considérons une relation d'accessibilité pour chaque paire d'agents $(a_i, a_j) \in \mathcal{N}^2$. Cette relation exprime le fait qu'un agent a_i a confiance dans a_j à propos d'une proposition ϕ dans un monde possible $w \in \mathcal{W}$ si, et seulement si, ϕ est vraie dans chaque monde accessbile depuis w par la relation $\mathcal{T}^s_{i,j}$. Ainsi, notre cadre de Kripke $\mathcal{C}$ est tel que pour chaque $a_i, a_j \in \mathcal{N}$:

1. $\forall w \in \mathcal{W}, \exists v \in \mathcal{W} : w\mathcal{T}^s_{i,j} v$
2. $\forall w, u, v \in \mathcal{W} : w\mathcal{B}_i u \wedge u\mathcal{T}^s_{i,j} v \Rightarrow w\mathcal{T}^s_{i,j} v$
3. $\forall w, u, v \in \mathcal{W} : w\mathcal{B}_i u \wedge w\mathcal{T}^s_{i,j} v \Rightarrow u\mathcal{T}^s_{i,j} v$
4. $\forall w, u, v \in \mathcal{W} : w\mathcal{B}_i u \wedge u\mathcal{B}_j v \Rightarrow w\mathcal{T}^s_{i,j} v$
5. $\mathcal{B}_i$ est sérielle, transitive et euclidienne.

La notion de sincérité n'est pas liée à celle de fiabilité. En effet, un agent peut être sincère dans un énoncé et croire ce qu'il dit alors que cet énoncé est manifestement faux. Cela ne contredit pas la sincérité. Il y a toutefois contradiction lorsqu'un agent énonce une chose et son contraire, et il n'est plus possible de faire confiance en la sincérité de l'agent. La propriété (1) le représente en exprimant le fait qu'il y ait toujours un monde

accessible par $\mathcal{T}^s_{i,j}$ depuis n'importe quel monde. De plus, un agent est « conscient » de la confiance qu'il accorde à un autre, ce qui est représenté par la propriété (2) – si un agent a_i a confiance en un agent a_j alors a_i croit qu'il a confiance en a_j – et la propriété (3) – si un agent a_i n'a pas confiance en un agent a_j alors a_i croit qu'il n'a pas confiance en a_j. La propriété (4) signifie qu'un agent sincère ne communique que des informations qu'il croit vraies et donc, lorsqu'un agent a_i a confiance en un agent a_j pour ϕ alors a_i croit que l'agent a_j croit ϕ. Enfin, les propriétés (5) sont les propriétés usuelles des modalités doxastiques.

L'axiomatique du système TB utilise les tautologies et les règles d'inférence classiques [2] du calcul propositionnel (**Nec**, **Sub**, **MP**), l'axiome **K** de la logique modale et un axiome de cohérence entre les confiances (**D**). À cela s'ajoutent trois axiomes représentant les interactions entre la croyance et la confiance ($\mathbf{4_{T,B}}$, $\mathbf{5_{T,B}}$, **S**).

Remarquons en premier lieu que [Liau, 2003] ne considère pas l'axiome **K** car il utilise une sémantique minimale exprimant une forme caractère irrationel de confiance, ce qui fait que $T^r_{i,j}p \wedge T^r_{i,j}(p \Rightarrow q)$, $T^r_{i,j}q$ ne peut pas être déduit. Toutefois, en se plaçant dans le cadre de systèmes artificiels conçus pour une application précise, il n'y a pas de raison de ne considérer les agents comme rationels. Ainsi, avoir confiance en a_j pour p et pour $p \Rightarrow q$ implique que a_i devrait avoir confiance en a_j pour q, que ce soit dans le contexte de la fiabilité ou de la sincérité. Notre modalité de confiance satisfait donc l'axiome **K** :

$$\vdash T^s_{i,j}(p \Rightarrow q) \Rightarrow T^s_{i,j}p \Rightarrow T^s_{i,j}q \quad (K)$$

Nous désirons aussi exprimer le fait que si un agent a_i a confiance en la sincérité d'un agent a_j pour une proposition donnée, a_i ne peut pas avoir confiance en a_j pour l'opposé car un agent sincère doit avoir un discours cohérent. Cependant, ceci ne peut être généralisé aux discours contradictoires entre agents. En effet, si un agent a_i a confiance en la sincérité de a_j pour p, rien n'empêche a_i d'avoir confiance en la sincérité d'un autre agent a_k pour $\neg p$. Bien que l'un des deux se trompe (et donc n'est pas fiable [3]), cela ne remet pas en cause leur sincérité et ce n'est donc pas une incohérence. Ainsi, nous considérons uniquement l'axiome **D** tel que :

$$\vdash T^s_{i,j}p \Rightarrow \neg T^s_{i,j}\neg p \quad (D)$$

Il y a également un lien entre la confiance et la croyance : un agent est « conscient » de la confiance qu'il accorde à un autre, représenté par les axiomes $\mathbf{4_{T,B}}$ et $\mathbf{5_{T,B}}$. Si un

2. Pour rappel, la nécessitation **Nec** signifie que si une formule ϕ est un théorème ($\vdash \phi$) alors n'importe quel agent i peut avoir confiance en la sincérité d'un autre agent j à propos de ce théorème ($\vdash T^s_{i,j}\phi$) et croit ϕ ($\vdash B_i\phi$). La substitution **Sub** signifie que si nous pouvons substituons uniformément n'importe quelle formule valide à un symbole propositionel dans un théorème alors la formule résultante est aussi un théorème. Le *modus ponens* **MP** signifie que si la formule $\vdash \phi$ est un théorème et que la formule $\vdash \phi \Rightarrow \psi$ est aussi un théorème alors la formule $\vdash \psi$ est prouvée.

3. Dans le cas de la fiabilité au contraire, il n'est pas possible de faire confiance à deux sources contradictoires [Liau, 2003].

agent a_i a confiance en la sincérité de a_j à propos de p, alors a_i crois qu'il a confiance en la sincérité de a_j sur p. De plus, nous considérons aussi une forme d'*introspection négative*. De manière intéressante, notre système nous permet de déduire les réciproques de ces deux axiomes.

$$\vdash T^s_{i,j}p \Rightarrow B_i T^s_{i,j}p \quad (4_{T,B})$$

$$\vdash \neg T^s_{i,j}p \Rightarrow B_i \neg T^s_{i,j}p \quad (5_{T,B})$$

Remarquons que nous ne considérons pas d'axiome d'incohérence entre la confiance et la croyance d'un même agent. En effet, si un agent croit que quelque chose est vrai, cela ne l'empêche en rien d'avoir confiance en la sincérité d'un autre agent à propos du fait que cette chose soit fausse. Encore une fois, il s'agit là d'une spécificité de la sincérité par rapport à la fiabilité. Enfin, nous considérons un dernier axiome, l'axiome **S** ou *axiome de sincérité*, qui exprime le fait que si un agent a_i a confiance dans la sincérité d'un autre agent a_j pour p, alors a_i croit que a_j croit p.

$$\vdash T^s_{i,j}p \Rightarrow B_i B_j p \quad (S)$$

Il est important de noter que nous ne considérons pas la réciproque de cet axiome. En effet, l'axiome **S** exprime le sens de la confiance en la sincérité et non pas la sincérité en elle-même. Un agent peut se tromper sur les croyances qu'il a envers les états mentaux d'un autre agent et faire confiance à un agent qui n'est en réalité pas sincère.

Au vu de ces axiomes et des propriétés des relations d'accessibilité de notre cadre de Kripke, **le système TB est cohérent et complet**. Nous ne donnons pas dans ce mémoire le détail des preuves qui sont des preuves standards (par exemple la preuve de complétude est une preuve à la [Henkin, 1949] où le lemme de Lindenbaum est utilisé pour construire le modèle canonique). Nous renvoyons le lecteur à [Leturc et Bonnet, 2018a] pour les preuves complètes et les rappels sur les éléments de méthodologie.

1.2 Propriétés de la confiance en la sincérité

Nous présentons ici quelques propriétés intéressantes du système TB. Comme il s'agit d'une logique normale, un agent a_i ne peut pas avoir confiance en la sincérité d'un discours contradictoire d'un agent a_j car cela conduirait à faire confiance à toutes les propositions de a_j.

Proposition 4.1 (Distributivité de la confiance en la sincérité)
Soit $a_i, a_j \in \mathcal{N}$.

1. $\vdash T^s_{i,j}\phi \wedge T^s_{i,j}\psi \equiv T^s_{i,j}(\phi \wedge \psi) \quad (\wedge_T)$
2. $\vdash (T^s_{i,j}\phi \vee T^s_{i,j}\psi) \Rightarrow T^s_{i,j}(\phi \vee \psi) \quad (\vee_T)$

Comme $T^s_{i,j}$ est une modalité normale, la proposition ci-dessus se déduit immédiatement [Chellas, 1980]. De plus, comme énoncé en section 1.1, nous avons :

Proposition 4.2 (Reciproques des axiomes $\mathbf{4_{T,B}}$ et $\mathbf{5_{T,B}}$)
Pour tous les agents $a_i, a_j \in \mathcal{N}$,

1. $\vdash B_i T^s_{i,j} p \Rightarrow T^s_{i,j} p \quad (C4_{T,B})$
2. $\vdash B_i \neg T^s_{i,j} p \Rightarrow \neg T^s_{i,j} p \quad (C5_{T,B})$

Démonstration 4.2
Soit $a_i, a_j \in \mathcal{N}$. Nous prouvons la première propriété :

1. $\vdash \neg T^s_{i,j} p \Rightarrow B_i \neg T^s_{i,j} p \quad (5_{T,B})$
2. $\vdash B_i \neg T^s_{i,j} p \Rightarrow \neg B_i T^s_{i,j} p \quad (D_B)$
3. $\vdash (\neg T^s_{i,j} p \Rightarrow (B_i \neg T^s_{i,j} p \Rightarrow \neg B_i T^s_{i,j} p))$
4. $\vdash (\neg T^s_{i,j} p \Rightarrow (B_i \neg T^s_{i,j} p \Rightarrow \neg B_i T^s_{i,j} p)) \Rightarrow$
   $((\neg T^s_{i,j} p \Rightarrow B_i \neg T^s_{i,j} p) \Rightarrow (\neg T^s_{i,j} p \Rightarrow \neg B_i T^s_{i,j} p))$
5. $\vdash \neg T^s_{i,j} p \Rightarrow \neg B_i T^s_{i,j} p$
6. $\vdash (\neg T^s_{i,j} p \Rightarrow \neg B_i T^s_{i,j} p) \Rightarrow (B_i T^s_{i,j} p \Rightarrow T^s_{i,j} p)$
7. $\vdash B_i T^s_{i,j} p \Rightarrow T^s_{i,j} p$

Nous prouvons la seconde propriété :

1. $\vdash B_i \neg T^s_{i,j} p \Rightarrow \neg B_i T_{i,j} p \quad (D_B)$
2. $\vdash T^s_{i,j} p \Rightarrow B_i T^s_{i,j} p \quad (4_{T,B})$
3. $\vdash (T^s_{i,j} p \Rightarrow B_i T^s_{i,j} p) \Rightarrow (\neg B_i T^s_{i,j} p \Rightarrow \neg T^s_{i,j} p)$
4. $\vdash \neg B_i T^s_{i,j} p \Rightarrow \neg T^s_{i,j} p$
5. $\vdash (B_i \neg T^s_{i,j} p \Rightarrow (\neg B_i T^s_{i,j} p \Rightarrow \neg T_{i,j} p)$
6. $\vdash (B_i \neg T^s_{i,j} p \Rightarrow (\neg B_i T^s_{i,j} p \Rightarrow \neg T_{i,j} p) \Rightarrow$
   $((B_i \neg T^s_{i,j} p) \Rightarrow (\neg B_i T^s_{i,j} p)) \Rightarrow ((B_i \neg T^s_{i,j} p) \Rightarrow (\neg T_{i,j} p))$
7. $\vdash B_i \neg T_{i,j} p \Rightarrow \neg T_{i,j} p$

□

Ces propriétés expriment le fait que lorsque les agents croient faire confiance alors c'est qu'ils font confiance et que lorsqu'ils croient ne pas faire confiance, alors c'est qu'ils ne font pas confiance. Enfin, nous avons une propriété de rationalité faible de la confiance en la sincérité, exprimant le fait qu'un agent a_i qui croit qu'un autre agent a_j croit une proposition ϕ ne peut pas avoir confiance en la sincérité de a_j pour $\neg\phi$. Cependant, rien ne l'oblige pour autant à accorder sa confiance à a_j.

Proposition 4.3 (Rationalité faible de la confiance en la sincérité)
Pour tous les agents $a_i, a_j \in \mathcal{N}$, $\vdash B_i B_j \phi \Rightarrow \neg T^s_{i,j} \neg \phi$.

Démonstration 4.3
Soit $a_i, a_j \in \mathcal{N}$.

1. $\vdash B_j\phi \Rightarrow \neg B_j\neg\phi$
2. $\vdash B_i(B_j\phi \Rightarrow \neg B_j\neg\phi)$
3. $\vdash B_i(B_j\phi \Rightarrow \neg B_j\neg\phi) \Rightarrow (B_iB_j\phi \Rightarrow B_i\neg B_j\neg\phi)$
4. $\vdash B_iB_j\phi \Rightarrow B_i\neg B_j\neg\phi$
5. $\vdash B_i\neg B_j\neg\phi \Rightarrow \neg B_iB_j\neg\phi) \Rightarrow \neg B_iB_j\neg\phi$
6. $\vdash T^s_{i,j}\neg\phi \Rightarrow B_iB_j\neg\phi$
7. $\vdash (T^s_{i,j}\neg\phi \Rightarrow B_iB_j\neg\phi) \Rightarrow (\neg B_iB_j\neg\phi \Rightarrow \neg T^s_{i,j}\neg\phi)$
8. $\vdash \neg B_iB_j\neg\phi \Rightarrow \neg T^s_{i,j}\neg\phi$
9. $\vdash (B_iB_j\phi \Rightarrow B_i\neg B_j\neg\phi \Rightarrow \neg B_iB_j\neg\phi) \Rightarrow ((B_iB_j\phi \Rightarrow B_i\neg B_j\neg\phi) \Rightarrow (B_iB_j\phi \Rightarrow \neg B_iB_j\neg\phi))$
10. $\vdash (B_iB_j\phi \Rightarrow \neg B_iB_j\neg\phi \Rightarrow \neg T^s_{i,j}\neg\phi) \Rightarrow (((B_iB_j\phi \Rightarrow \neg B_iB_j\neg\phi) \Rightarrow (B_iB_j\phi \Rightarrow \neg T^s_{i,j}\neg\phi))$
11. $\vdash B_iB_j\phi \Rightarrow \neg T^s_{i,j}\neg\phi$

□

Enfin, certains travaux ont déjà mis en lumière des arguments pour l'absence de transitivité dans certains aspects de la confiance [Christianson et Harbison, 1997]. La transitivité est plutôt une caractéristique de la confiance en la fiabilité tandis que la confiance en la sincérité n'est pas transitive au sens où nous ne disposons pas de règle d'inférence permettant de déduire que si $T^s_{i,j}T^s_{j,k}\phi$ alors $T^s_{i,k}\phi$. En effet, ce n'est pas parce qu'un agent a_i a confiance en la sincérité d'un agent a_j lorsque ce dernier énonce qu'il a confiance en la sincérité d'un agent a_k que a_i peut faire confiance en la sincérité de a_k : a_j peut être sincère tout en ayant tort. Cependant, la confiance en la sincérité dispose d'une pseudo-transitivité qui permet de déduire des croyances sur la représentation des croyances qu'un autre agent se fait à propos d'un tiers.

Proposition 4.4 (Pseudo-transivité de la confiance en la sincérité)
Pour tous les agents $a_i, a_j, a_k \in \mathcal{N}$, $\vdash T^s_{i,j}T^s_{j,k}\phi \Rightarrow B_iB_jB_k\phi$.

Démonstration 4.4
Soit $a_i, a_j, a_k \in \mathcal{N}$.

1. $\vdash T^s_{i,j}T^s_{j,k}\phi \Rightarrow B_iB_jT^s_{j,k}\phi$
2. $\vdash T^s_{j,k}\phi \Rightarrow B_jB_k\phi$
3. $\vdash B_jT^s_{j,k}\phi \Rightarrow T^s_{j,k}\phi$
4. $\vdash B_i(B_jT^s_{j,k}\phi \Rightarrow T^s_{j,k}\phi)$
5. $\vdash B_i(B_jT^s_{j,k}\phi \Rightarrow T^s_{j,k}\phi) \Rightarrow B_iB_jT^s_{j,k}\phi \Rightarrow B_iT^s_{j,k}\phi$
6. $\vdash B_iB_jT^s_{j,k}\phi \Rightarrow B_iT^s_{j,k}\phi$
7. $\vdash T^s_{i,j}T^s_{j,k}\phi \Rightarrow B_iB_jB_k\phi$

□

1.3 Extension à la confiance partagée

Nous étendons notre notion de confiance en la sincérité à un groupe d'agents afin d'exprimer une *confiance partagée*. Les autres aspects de la confiance collective, comme la confiance réciproque ou la confiance mutuelle sont des perspectiques que nous abordons au chapitre 6. Afin de définir la confiance partagée, nous nous fondons sur [Smith *et al.*, 2011] : il y a confiance partagée signifiant qu'un groupe d'agents accorde sa confiance à un autre groupe d'agents si, et seulement si, tous les agents du premier groupe accordent leur confiance à chaque agent du second groupe. Notons que cette notion de confiance partagée peut être définie différemment dans la littérature. Par exemple, [Herzig *et al.*, 2010] considèrent un prédicat de *réputation* indiquant qu'une *majorité* des agents de I a une confiance dispositionelle envers les agents de J. Formellement,

$$\forall I, J \subseteq \mathcal{N} : Tc_{I,J}\phi \stackrel{\triangle}{=} \bigwedge_{(i,j)\in I\times J} T^s_{i,j}\phi$$

Ceci exprime un consensus au sens que tous les agents de I font confiance à tous les agents de J à propos d'un même énoncé. De plus, nous considérons une notion duale à la confiance partagée, notée $Tc^*_{I,J}$ qui exprime le fait qu'au moins un agent de I a confiance en la sincérité d'un autre agent de J. En effet, si aucun agent de I n'a confiance en la sincérité d'un agent de J pour ϕ alors $\neg Tc^*_{I,J}\phi$.

$$\forall I, J \subseteq \mathcal{N} : Tc^*_{I,J}\phi \stackrel{\triangle}{=} \bigvee_{(i,j)\in I\times J} T^s_{i,j}\phi$$

Cette confiance partagée se comporte comme un système KD

Proposition 4.5 (La confiance partagée est un système KD)
Pour tout $I, J, K \subseteq \mathcal{N}$:

1. $\vdash Tc_{I,J}\phi \wedge Tc_{I,J}\psi \equiv Tc_{I,J}(\phi \wedge \psi)$
2. $\vdash (Tc_{I,J}\phi \vee Tc_{I,J}\psi) \Rightarrow Tc_{I,J}(\phi \vee \psi)$
3. $\vdash (Tc_{I,J}\phi \wedge Tc_{I,J}(\phi \Rightarrow \psi)) \Rightarrow Tc_{I,J}\psi$
4. $\vdash Tc_{I,J}\phi \Rightarrow \neg Tc^*_{I,J}\neg\phi$
5. $\vdash Tc_{I,J}\phi \Rightarrow \neg Tc_{I,J}\neg\phi$

Démonstration 4.5
Soit $I, J, K \subseteq \mathcal{N}$,

(1) $\vdash \bigwedge_{(i,j)\in I\times J}(T^s_{i,j}\phi \wedge T^s_{i,j}\psi) \equiv \bigwedge_{(i,j)\in I\times J} T^s_{i,j}(\phi \wedge \psi)$

(2) $\vdash \bigwedge_{(i,j)\in I\times J}(T^s_{i,j}\phi \vee T^s_{i,j}\psi) \Rightarrow \bigwedge_{(i,j)\in I\times J} T^s_{i,j}(\phi \vee \psi)$

(3) est obtenu par :

— $\{Tc_{I,J}\phi \wedge Tc_{I,J}(\phi \Rightarrow \psi)\} \vdash \bigwedge_{(i,j)\in I\times J}(T^s_{i,j}\phi \wedge (T^s_{i,j}(\phi \Rightarrow \psi)))$

— $\{Tc_{I,J}\phi \wedge Tc_{I,J}(\phi \Rightarrow \psi)\} \vdash \bigwedge_{(i,j)\in I\times J} T^s_{i,j}\psi$

Par conséquent $\vdash (Tc_{I,J}\phi \wedge Tc_{I,J}(\phi \Rightarrow \psi)) \Rightarrow Tc_{I,J}\psi$.

(4) est obtenu par :

— $\{Tc_{I,J}\phi\} \vdash \bigwedge_{(i,j)\in I\times J} (T^s_{i,j}\phi \wedge (T^s_{i,j}\phi \Rightarrow \neg T^s_{i,j}\neg\phi))$

— $\{Tc_{I,J}\phi\} \vdash \bigwedge_{(i,j)\in I\times J} \neg T^s_{i,j}\neg\phi$

— $\{Tc_{I,J}\phi\} \vdash \neg \bigvee_{(i,j)\in I\times J} T^s_{i,j}\neg\phi$

Par conséquent, $\vdash Tc_{I,J}\phi \Rightarrow \neg Tc^*_{I,J}\neg\phi$.

(5) est obtenu par :

— $\{Tc_{I,J}\phi\} \vdash \bigwedge_{(i,j)\in I\times J} (T^s_{i,j}\phi \wedge (T^s_{i,j}\phi \Rightarrow \neg T^s_{i,j}\neg\phi))$

— $\{Tc_{I,J}\phi\} \vdash \bigwedge_{(i,j)\in I\times J} \neg T^s_{i,j}\neg\phi$

— $\{Tc_{I,J}\phi\} \vdash \bigwedge_{(i,j)\in I\times J} \neg T^s_{i,j}\neg\phi \Rightarrow \bigvee_{(i,j)\in I\times J} \neg T^s_{i,j}\neg\phi$

— $\{Tc_{I,J}\phi\} \vdash \bigvee_{(i,j)\in I\times J} \neg T^s_{i,j}\neg\phi \Rightarrow \neg \bigwedge_{(i,j)\in I\times J} T^s_{i,j}\neg\phi$

— $\{Tc_{I,J}\phi\} \vdash \neg \bigwedge_{(i,j)\in I\times J} T^s_{i,j}\neg\phi$

Par conséquent, $\vdash Tc_{I,J}\phi \Rightarrow \neg Tc_{I,J}\neg\phi$. □

Enfin, l'axiomatique de la confiance en la sincérité est la même au niveau collectif.

Proposition 4.6 (La confiance partagée implique des croyances partagées) *Pour tout* $I, J, K \subseteq \mathcal{N}$,

$$(1) \vdash Tc_{I,J}\phi \Rightarrow \bigwedge_{(i,j)\in I\times J} B_iB_j\phi$$

$$(2) \vdash Tc_{I,J}Tc_{J,K}\phi \Rightarrow \bigwedge_{(i,j,k)\in I\times J\times K} B_iB_jB_k\phi$$

Démonstration 4.6
Pour tout $I, J, K \subseteq \mathcal{N}$

(1) est obtenu par :

— $\{Tc_{I,J}\phi\} \vdash \bigwedge_{(i,j)\in I\times J} (T^s_{i,j}\phi \wedge (T^s_{i,j}\phi \Rightarrow B_iB_j\phi))$

— $\{Tc_{I,J}\phi\} \vdash \bigwedge_{(i,j)\in I\times J} B_iB_j\phi$

Par conséquent, $\vdash Tc_{I,J}\phi \Rightarrow \bigwedge_{(i,j)\in I\times J} B_iB_j\phi$.

(2) est obtenu par :

— $\{Tc_{I,J}Tc_{J,K}\phi\} \vdash \bigwedge\limits_{(i,j)\in I\times J}(T^s_{i,j} \bigwedge\limits_{k\in K} T_{j,k}\phi \wedge(T^s_{i,j}T^s_{j,k}\phi \Rightarrow B_iB_jB_k\phi))$

— $\{Tc_{I,J}Tc_{J,K}\phi\} \vdash \bigwedge\limits_{(i,j,k)\in I\times J\times K} B_iB_jB_k\phi$

Par conséquent, $\vdash Tc_{I,J}Tc_{J,K}\phi \Rightarrow \bigwedge_{(i,j,k)\in I\times J\times K} B_iB_jB_k\phi$. □

Remarquons que ces preuves reposent sur la distributivité de la confiance en la sincérité et parce que $\forall k \in \mathcal{N}, \vdash B_k(p\wedge q) \equiv B_kp\wedge B_kq$. Ainsi, si deux groupes d'agents ont confiance en la sincérité de l'autre groupe, alors chaque agent de I croit que chaque autre agent de J croit ce qu'ils énoncent collectivement.

2 Crédibilité des discours

Si le travail précédent caractérise la confiance en la sincérité, il fait l'impasse sur la manière dont cette confiance est construite et sur son usage. Dans cette section, nous abordons ces deux questions par le prisme de la crédibilité dans les systèmes de réputation qui, parce qu'ils permettent de raisonner sur les statistiques des interactions, sont bien adaptés pour traiter de cette question.

2.1 Une notion de crédibilité

Dans toute la suite, nous nous fondons sur le modèle générique de système de réputation présenté au chapitre 3. Pour rappel, nous considérons les observations résultant des interactions entre agents comme des variables aléatoires issues de fonctions de distribution de probabilité de paramètres inconnus, mais indépendantes de l'agent qui a interagit. Ainsi, avec suffisament d'observations, les estimations faites par chaque agent doivent converger. Ainsi, comme [Malik et Bouguettaya, 2009], nous considérons qu'un témoignages est crédible s'il est similaire aux observations des autres agents et nous proposons d'utiliser la divergence de Kullback-Leibler [Kullback, 1997] pour comparer les estimations des fonctions de distribution de probabilité, et ce afin de définir si un témoignage est crédible ou non. Contrairement aux approches qui pondèrent les témoignages, notre mesure de crédibilité est associée aux témoignages et non pas aux agents. De plus, contrairement aux approches qui se fondent sur un seuil donné *a priori*, nous proposons une notion de seuil dynamique en fonction des connaissances des agents.

L'expertise de a_k pour le service s_x correspond à l'espérance de gain moyen lorsqu'un agent a_i lui demande ce service. À partir de ses observations directes $O_{i,k,x}$, l'agent a_i peut calculer le gain moyen qu'il a reçu lors de ses interactions avec l'agent a_k. Notons par $\mu_{i,k,x}$ ce gain moyen et par $\sigma_{i,k,x}$ l'écart-type. Bien que la qualité des services fournis

par a_k ne suit pas nécessairement une loi normale, l'agent a_i peut l'approximer par la loi $\mathcal{N}(\mu_{i,k,x}, \sigma^2_{i,k,x})$. De même, grâce aux témoignages qu'il a reçus, a_i peut calculer $\mu_{i,j,k,x}$ et $\sigma_{i,j,k,x}$ le gain moyen et l'écart-type fondé sur les témoignages de a_j et ainsi obtenir l'approximation $\mathcal{N}(\mu_{i,j,k,x}, \sigma^2_{i,j,k,x})$. Ainsi, sous l'hypothèse que la qualité des services fournis est indépendante de l'agent recevant le service, deux agents doivent obtenir les mêmes estimations pour un grand nombre d'observations.

Hypothèse 4.7
Si $O_{i,k,x}$ et $F_{i,j,k,x}$ sont des observations du service s_x fourni par a_k alors ces observations proviennent de la même fonction de distribution de probabilité. Ainsi, pour $n = |O_{i,k,x}|$ et $m = |F_{i,j,k,x}|$:

$$\lim_{n,m\to\infty} \mathcal{N}(\mu_{i,k,x}, \sigma^2_{i,k,x}) = \mathcal{N}(\mu_{i,j,k,x}, \sigma^2_{i,j,k,x})$$

Si $F_{i,j,k,x}$ est un faux témoignage alors :

$$\lim_{n,m\to\infty} \mathcal{N}(\mu_{i,k,x}, \sigma^2_{i,k,x}) \neq \mathcal{N}(\mu_{i,j,k,x}, \sigma^2_{i,j,k,x})$$

Notons que cette supposition n'a de sens que si les observations des agents sont sans erreurs. Dans le cas contraire, l'erreur d'observation peut être considérée comme du bruit et si le bruit d'un témoignage est trop important, ce dernier fausse l'estimation de l'expertise. Par ailleurs, comme les agents ne disposent que d'un nombre fini d'observations, leurs estimations diffèrent nécessairement. Ainsi, une mesure de crédibilité des témoignages nécessite de prendre en compte ces deux points. Pour cela, nous proposons d'utiliser la divergence de Kullback-Leibler pour mesurer la différence entre deux témoignages.

Définition 4.8
La divergence de Kullback-Leibler *entre les observations de a_i et les témoignages de a_j vis-à-vis de $\varepsilon_{k,x}$ est :*

$$D_{i,j,k,x} = D_{KL}(\mathcal{N}(\mu_{i,k,x}, \sigma^2_{i,k,x}) || \mathcal{N}(\mu_{i,j,k,x}, \sigma^2_{i,j,k,x}))$$

où :

$$D_{KL}(f||g) = \int f(x) \log \frac{f(x)}{g(x)} d(x)$$

Si les témoignages fournis par l'agent a_j sont similaires aux observations de l'agent a_i, alors $D_{i,j,k,x} \simeq 0$. Inversement, si $D_{i,j,k,x}$ est supérieure à un seuil δ, cela signifie que l'agent a_i et l'agent a_j n'ont pas la même estimation de $\varepsilon_{k,x}$. Cela peut être dû à plusieurs facteurs : soit les agents n'ont pas suffisament d'observations pour avoir une bonne estimation de $\varepsilon_{k,x}$, soit ils évaluent la qualité des services sur des critères différents ($v_i \neq v_j$), soit les témoignages de a_j sont faux. Dans le premier cas, après quelques interactions supplémentaires, $D_{i,j,k,x}$ tendra vers 0. Dans les deux autres cas, cela signifie que l'agent a_i ne peut pas considérer comme crédibles les témoignages de l'agent a_j car ils sont soit faux, soit inutiles.

Pour fixer le seuil δ à partir duquel a_i considère comme non crédibles des témoignages, nous proposons d'utiliser l'erreur type de l'estimateur. Nous considérons ici l'erreur type de la moyenne (SEM $= \sigma_{i,k,x}/\sqrt{n}$) qui correspond à la confiance de a_i dans son estimation de $\mu_{i,k,x}$. L'approximation de la fonction de distribution de probabilité par une loi normale permet à l'agent a_i de déterminer avec une confiance de 95 % que la moyenne réelle des gains espérés se trouve dans l'intervalle :

$$[\mu_{i,k,x} - \frac{1.96 \times \sigma_{i,k,x}}{\sqrt{n}}, \mu_{i,k,x} + \frac{1.96 \times \sigma_{i,k,x}}{\sqrt{n}}]$$

Ainsi, l'agent a_i peut utiliser sa propre SEM pour fixer δ et calculer si les témoignages de a_j sont crédibles. Remarquons qu'un agent a_i peut n'avoir eu aucune interaction avec l'agent à propos duquel il reçoit un témoignage et ne peut donc calculer $D_{i,j,k,x}$. Dans ce cas, l'agent peut adopter soit une attitude *optimiste* et considérer que le témoignage est toujours crédible, soit une attitude *pessimiste* et considérer que le témoignage n'est jamais crédible. Dans la suite de cette section, nous considérons des agents optimistes [4].

Définition 4.9
Soit $F_{i,j,k,x}$ le témoignage que a_j a fourni à a_i vis-à-vis de $\varepsilon_{k,x}$. $F_{i,j,k,x}$ est KL-crédible *si $O_{i,k,x} = \emptyset$ ou $D_{i,j,k,x} \leq \delta$, où :*

$$\delta = D_{KL}(\mathcal{N}(\mu_{i,k,x}, \sigma^2_{i,k,x})||\mathcal{N}(\mu_{i,k,x} + \frac{1.96 \times \sigma_{i,k,x}}{\sqrt{n}}, \sigma^2_{i,k,x}))$$

Utiliser la divergence de Kullback-Leibler comme mesure de crédibilité et l'erreur type de la moyenne pour fixer dynamiquement le seuil présentent plusieurs avantages.

1. Comme la divergence de Kullback-Leibler est fortement liée à l'entropie, un témoignage divergent apporte de nouvelles informations utiles lorsque l'agent ne dispose que de peu d'observations. À l'inverse, plus l'agent dispose d'informations, moins un nouveau témoignage est supposé apporter une information utile. Comme l'erreur type de la moyenne dépend du nombre d'observations, plus l'agent en dispose, moins un témoignage divergent est supposé crédible, et inversement. Ainsi, cette notion de crédibilité est dynamique car elle peut être remise en cause au cours du temps au fur et à mesure que l'agent obtient de nouvelles informations.
2. L'assymétrie de la divergence de Kullback-Leibler nous permet de représenter le fait que si l'agent a_i considère comme non crédible le témoignage d'un agent a_j, l'agent a_j peut quant à lui considérer le témoignage de l'agent a_i comme crédible car lui-même ne dispose pas du même nombre d'observations. Enfin, la prise en compte de l'erreur type de la moyenne dans la définition du seuil de crédibilité permet à un agent de considérer que ses observations sont en partie imparfaites.

4. Le cas pessimiste se traite en définissant « $F_{i,j,k,x}$ est KL-crédible si $O_{i,k,x} \neq \emptyset$ et $D_{i,j,k,x} \leq \delta$ ».

3. La divergence de Kullback-leibler rend la mesure de crédibilité robuste. En effet, bien qu'une connaissance de δ permette à un agent malveillant de construire un témoignage qui pourrait être accepté comme crédible, un tel faux témoignage sera alors peu divergent des observations de l'agent manipulé. Ainsi, il sera nécessaire d'avoir un grand nombre de faux témoignages pour affecter la décision de l'agent. Remarquons que ceci s'applique aussi aux témoignages d'agents honnêtes : ils doivent aussi être nombreux pour affecter la décision de l'agent.

Dans toute la suite, nous notons par $KL_i(F_{i,j,k,x})$ (resp. $\neg KL_i(F_{i,j,k,x})$) si le témoignage $F_{i,j,k,x}$ est KL-crédible (resp. non KL-crédible) du point de vue de a_i.

2.2 Filtrer les témoignages non crédibles

Afin de diminuer l'influence des faux témoignages dans le système de réputation, nous proposons d'introduire dans les fonctions de réputation un mécanisme de filtrage des témoignages jugés non crédibles.

Définition 4.10
La fonction de filtrage *de l'agent a_i est la fonction $\phi_i(\mathcal{F}_i)$ qui retourne l'ensemble des témoignages que a_i considère comme crédibles.*

Nous proposons ici d'utiliser dans une fonction de réputation uniquement les témoignages retournés par la fonction de filtrage.

Définition 4.11
Soit un service $s_x \in S$ et deux agents $a_i \in N$ et $a_k \in N_x$. La réputation crédible *de a_k pour le service s_x fondée sur la fonction de réputation f_i et la fonction de filtrage ϕ_i est définie par $f_i(a_k, s_x, \phi_i(\mathcal{F}_i))$.*

Bien qu'il existe de nombreuses fonctions de filtrage possibles, nous nous concentrons uniquement sur trois d'elles : la première utilise trivialement la mesure de crédibilité que nous avons définie précédemment et les deux autres ont pour objectif de tenir compte de l'incertitude de l'agent, respectivement en généralisant ses observations ou en faisant appel aux observations d'autres agents. En premier lieu, une méthode intuitive pour filtrer les témoignages est de ne considérer que les témoignages KL-crédibles (définition 4.9) du point de vue de l'agent a_i.

Définition 4.12
Soit un service $s_x \in S$ et deux agents $a_i \in N$ et $a_k \in N_x$. La fonction de KL-filtrage *est la fonction ϕ_i définie par :*

$$\phi_i(\mathcal{F}_i) = \{F_{i,j,k,x} \in \mathcal{F}_i | KL_i(F_{i,j,k,x})\}$$

Avec cette approche, un témoignage de a_j est considéré comme crédible ou non indépendamment de la crédibilité de ses autres témoignages. Or, si un agent n'a pas toujours les observations lui permettant de juger correctement un témoignage, il peut en avoir pour juger un autre témoignage provenant du même agent. Si nous faisons l'hypothèse qu'un agent mentant sur un témoignage a une forte probabilité de mentir sur un autre, nous pouvons considérer que si a_j n'est pas crédible sur un sous-ensemble de ses témoignages alors aucun de ses témoignages n'est crédible.

Définition 4.13
Soient deux agents $a_i, a_j \in N$. L'agent a_j est k-crédible *si :*

$$\forall a_{k'} \in N, s_x \in S : |\{F_{i,j,k',x} \in \mathcal{F}_i | \neg KL_i(F_{i,j,k',x})\}| \leq k$$

Dans toute la suite, nous notons par $KL_i(N) \subseteq N$ l'ensemble des agents considérés comme k-crédibles par a_i. Nous pouvons ainsi définir une fonction de filtrage plus drastique qui rejette tous les témoignages provenant des agents non k-crédibles.

Définition 4.14
La fonction de filtrage par k fautes *est la fonction ϕ_i telle que :*

$$\phi_i(\mathcal{F}_i) = \{F_{i,j,k',x} \in \mathcal{F}_i | a_j \in KL_i(N) \wedge KL_i(F_{i,j,k',x})\}$$

Remarquons que même si l'agent a_j est k-crédible, le sous-ensemble de ses témoignages qui ne sont pas KL-crédibles sont tout de même filtrés. Ainsi, le filtrage par k fautes est une généralisation de KL-filtrage. En effet, plus k est proche de 0, moins un agent accepte de témoignages crédibles car l'agent qui les fournit ne l'est pas. Inversement, plus k est grand, plus le filtrage par k fautes est proche du KL-filtrage.

Les deux fonctions de filtrage précédentes sont fondées sur les observations de l'agent a_i. La troisième fonction de filtrage que nous proposons permet d'utiliser les témoignages des autres agents pour déterminer si un témoignage est crédible, en s'inspirant d'une procédure de stochocratie. En politique, la stochocratie désigne un État dont le gouvernement est sélectionné aléatoirement. Les membres d'un tel gouvernement sont ainsi considérés comme moins sensibles à des manipulations [Delannoi et Dowlen, 2010]. Dans notre contexte, nous proposons d'utiliser la stochocratie afin de juger si un témoignage est crédible : la fonction de filtrage par k-stochocratie accepte un témoignage si, parmi un sous-ensemble de k agents tirés aléatoirement uniformément dans N, une majorité d'entre eux le juge comme KL-crédible.

Définition 4.15
Le témoignage $F_{i,j,k',x}$ est dit crédible par k-stochocratie *si, pour un sous-ensemble $N' \subseteq N \setminus \{a_j, a_{k'}\}$ de k agents tirés aléatoirement uniformément au moins $\lceil k/2 \rceil$ agents de N' jugent $F_{j,k',x}$ comme KL-crédible.*

Dans la suite, nous notons L_i l'ensemble des témoignages considérés comme crédibles par k-stochocratie par l'agent a_i. Nous pouvons ainsi définir la fonction de filtrage qui rejette tous les témoignages qui ne sont pas crédibles par k-stochocratie.

Définition 4.16
La fonction de filtrage par k-stochocratie *est la fonction* ϕ_i *où :*

$$\phi_i(\mathcal{F}_i) = \{F_{i,j,k',x} \in \mathcal{F}_i | L_i(F_{i,j,k',x})\}$$

Notons que, dans la fonction de filtrage par k-stochocratie, les observations de l'agent sont elles aussi soumises au processus de filtrage. Ainsi, si a_i a un SEM important, ses observations peuvent ne pas être prises en compte lors du calcul de la réputation. Ceci permet de tenir compte d'une possible incertitude sur les observations. Nous pouvons aussi nous interroger sur l'hypothèse implicite qui rend cette fonction efficace : seule une minorité d'agents juges peuvent être malveillants. En effet, comme certains des agents malveillants peuvent être sélectionnés parmi l'ensemble des juges, il est possible que ceux-ci amènent un faux témoignage à être considéré comme crédible. Cependant, le processus de k-stochocratie ne rend crédible un faux témoignage que si et seulement si une majorité des juges le considère comme KL-crédible. Un tel cas n'arrive que si les agents honnêtes sélectionnés ont fourni peu de témoignages ou si la majorité des k agents sélectionnés appartiennent à la même coalition malveillante. La probabilité qu'au moins $\lceil k/2 \rceil$ de ces mauvais juges soient sélectionnés suit une loi hypergéométrique. Ainsi, un faux témoignage $F_{i,j,k',x}$ sera jugé par k-stochocratie comme crédible avec une probabilité p s'il existe l autres agents $a_z \in N \setminus \{a_j, a_{k'}\}$ tels que $KL_z(F_{i,j,k',x})$ et que :

$$\sum_{K=\lceil k/2 \rceil}^{k} \binom{l}{K} \binom{|N| - 2 - l}{k - K} \geq p \binom{|N| - 2}{k}$$

Par exemple, considérons un système avec $|N| = 100$, $l = 20$ et un agent honnête utilisant la 10-stochocratie. La probabilité qu'un faux témoignage soit jugé comme crédible est de $0,0278$. Notons que plus k est petit, plus la probabilité qu'un faux témoignage soit jugé crédible est importante. En revanche, un k plus grand implique un temps de calcul plus important.

2.3 Influence de la crédibilité

Pour évaluer l'efficacité des fonctions de filtrage, nous considérons deux protocoles expérimentaux. Le premier est le même que celui décrit au chapitre 3 section 4 : 100 agents interagissant durant 200 pas de temps dont 10 agents malveillants effectuant une attaque oscillante. Dans le contexte de notre application, il s'agit d'une phase d'initialisation puisque aucun agent n'a de connaissance sur les autres. Notre second protocole correspond à une phase de fonctionnement nominal : 100 agents dont 10 malveillants ont déjà interagit durant 100 pas de temps lorsque 20 nouveaux agents honnêtes rejoignent le système. Ces

nouveaux agents n'ont donc aucune connaissance *a priori* sur les autres agents et utilisent les témoignages qu'ils reçoivent pour calculer les valeurs de réputation. Nous mesurons alors le regret moyen de ces 20 agents durant les 200 pas de temps suivant.

Nous comparons trois des fonctions de réputation définies au chapitre 3 : l'estimation collective qui fait la moyenne des témoignages reçus, puis BetaReputation et FlowTrust, sans et avec nos trois fonctions de filtrage (KL-filtrage, filtrage par 10 fautes et 10-stochocratie [5]). Nous avons mené des expérimentations avec EigenTrust mais, comme il a été prouvé par [Cheng et Friedman, 2006], il suffit d'un unique faux témoignage pour manipuler le système et le regret des agents est identique avec ou sans filtrage. Afin de mettre en difficulté notre approche, nous ne présentons ici que les résultats avec la politique de sélection UCB qui est la plus sensible aux attaques oscillantes. Enfin, nos résultats sont comparés à la fonction d'estimation personnelle.

Nous considérons trois métriques : le regret (définition 3.17), le rappel et la précision [Bramer *et al.*, 2007]. Ces mesures nous permettent de déterminer si l'utilisation de la divergence de Kullback-Leibler permet d'évaluer correctement la crédibilité des témoignages. Le rappel est la proportion de faux témoignages filtrés parmi tous les témoignages reçus et la précision est la proportion de faux témoignages filtrés parmi l'ensemble de tous les témoignages filtrés.

La figure 4.1 montre le regret des agents selon les fonctions de filtrage utilisées sur les différentes fonctions de réputation. Les gains obtenus sont donnés sur le tableau 4.1 où les valeurs correspondent à la réduction de regret apportée par les fonctions de filtrage. Comme nous l'avons vu précédement (figure 3.3.1), l'estimation collective est très manipulable. Cependant, un filtrage permet de réduire fortement l'influence des faux témoignages et ainsi diminuer le regret. Remarquons sur la figure 4.1.2 que la 10-stochocratie est beaucoup plus performante en fonctionnement nominal qu'en phase d'initialisation. En effet, cette fonction utilise les témoignages des autres agents pour détecter les faux témoignages. Or, en phase nominale, les nouveaux agents se reposent sur les observations précises des agents ayant déjà interagi. Dans les autres cas, les agents doivent obtenir plusieurs observations avant d'avoir une estimation correcte de l'expertise des fournisseurs.

Les figures 4.1.3 et 4.1.4 montrent que les fonctions de filtrage permettent de détecter les promotions et ainsi obtenir un regret plus faible sur FlowTrust. En effet, seules les promotions sont efficaces sur FlowTrust. Une fois celles-ci détectées, les agents ayant les meilleures réputations sont nécessairement ceux ayant une bonne expertise. Comme précédement, la 10-stochocratie a de meilleurs résultats en phase nominale.

BetaReputation étant peu sensible aux manipulations, les fonctions de filtrage apporte naturellement un gain plus faible, voir même une augmentation du regret en phase nominale. Cependant ce résultat provient d'une moyenne sur l'ensemble de l'expérimentation.

5. Le paramètre $k = 10$ a été fixé par des expérimentations non présentées ici.

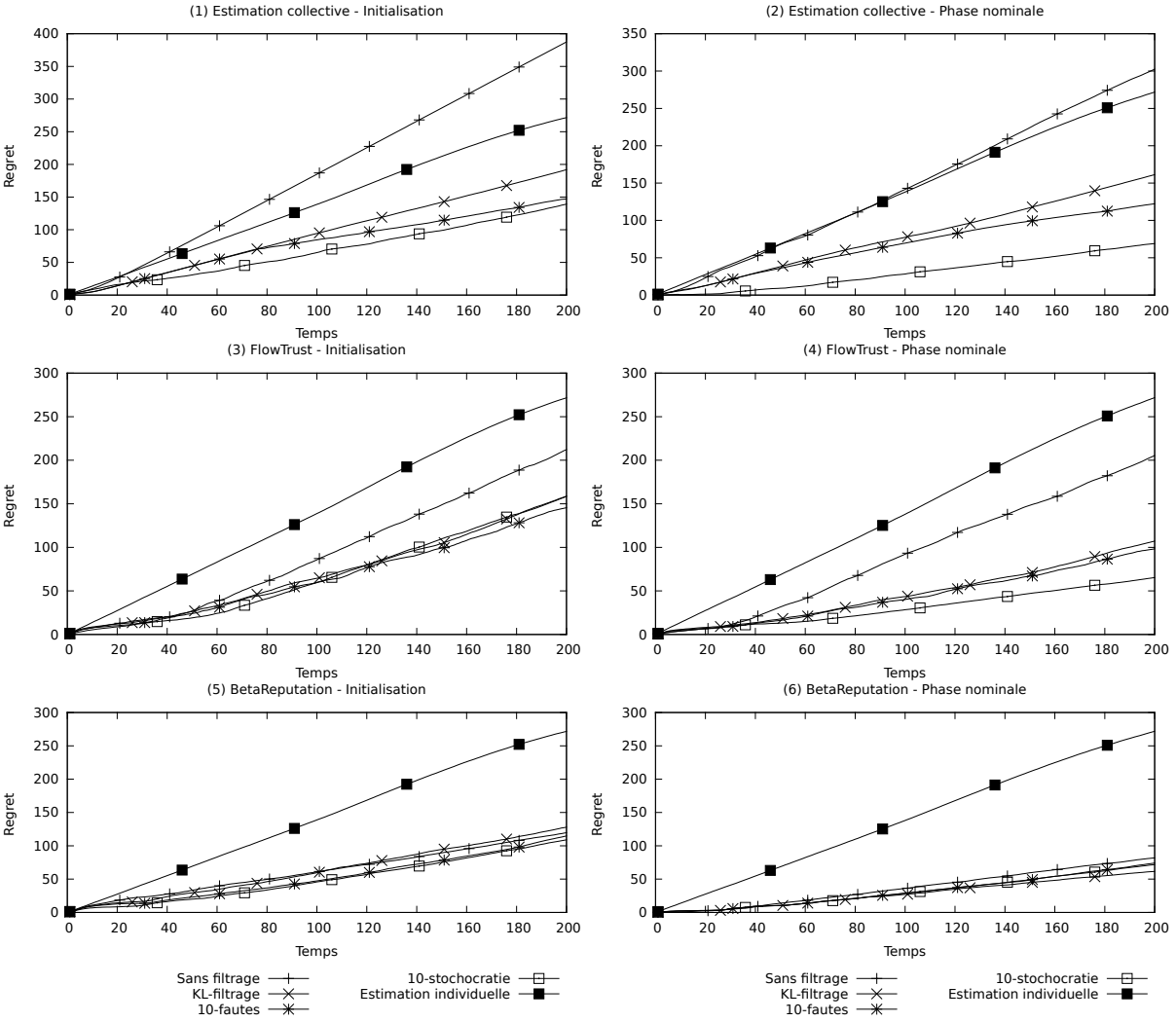

FIGURE 4.1 – Évolution du regret selon les différentes fonctions de filtrage

Dans les 40 premiers pas de temps, il y a effectivement une augmentation du regret. Après cela, les fonctions de filtrage obtiennent un regret inférieur. Il serait alors intéressant d'étudier si un agent malveillant ne pourrait pas profiter de cette sensibilité pour construire une manipulation efficace.

La figure 4.2 présente les rappels et précisions des différentes fonctions de filtrage sur les trois fonctions de réputation que sont l'estimation collective, FlowTrust et BetaReputation en phase d'initialisation.

Le KL-filtrage permet de détecter 50 % des faux témoignages pour l'estimation collective (figure 4.2.1), 65 % pour BetaReputation (figure 4.2.5) et seulement 20 % sur

Phase	Système de réputation	KL-filtrage	10-fautes	10-stochocratie
Initialisation	Estimation collective	$0,49$	$0,55$	$0,59$
	FlowTrust	$0,18$	$0,25$	$0,29$
	BetaReputation	$0,04$	$0,2$	$0,27$
Nominale	Estimation collective	$0,43$	$0,49$	$0,83$
	FlowTrust	$0,37$	$0,43$	$0,56$
	BetaReputation	$-0,05$	$-0,13$	$-0,11$
	(entre $t = 1$ et $t = 40$)	$-1,24$	$-1,44$	$-1,32$
	(entre $t = 41$ et $t = 200$)	$0,24$	$0,18$	$0,17$

TABLE 4.1 – Gains apportés par les fonctions de filtrage

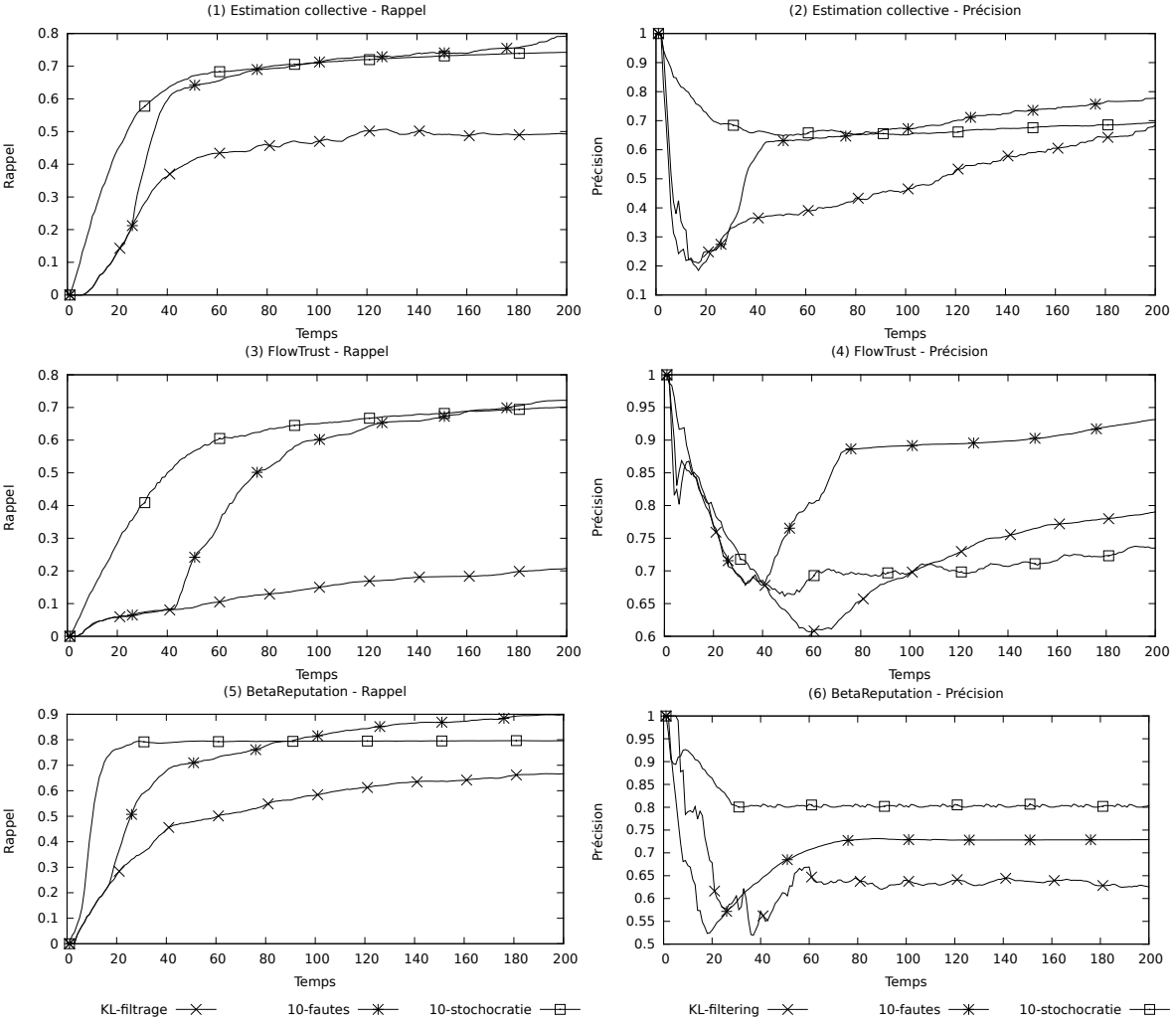

FIGURE 4.2 – Rappel et précision des fonctions des filtrages

FlowTrust (figure 4.2.3). Dans les deux premiers cas, les agents sont sensibles aux promotions et interagissent avec les agents malveillants dans les premiers pas de temps. Après quelques interactions, la divergence entre leurs observations et les faux témoignages est suffisante pour détecter les promotions. Les agents peuvent alors interagir avec les agents honnêtes même si ceux-ci sont diffamés, ce qui permet alors de détecter les diffamations [6]. Comme FlowTrust est insensible aux diffamations, le KL-filtrage ne détecte que les promotions et les diffamations vis-à-vis des meilleurs fournisseurs de services : les agents ne disposent alors pas de suffisament d'observations pour détecter les autres diffamations.

Le filtrage par 10 fautes obtient un rappel identique à celui du KL-filtrage dans les premiers pas de temps. Après cela, la majorité des faux témoignages est soudainement détectée car, les agents malveillants n'étant plus crédibles sur certains témoignages, aucun de leurs témoignages ne l'est. La 10-stochocratie permet de détecter sur les trois fonctions de réputation environ 70 % des faux témoignages. En effet, les agents n'ont pas besoin d'avoir des observations pour déterminer si un témoignage est crédible.

Les figures 4.2.2, 4.2.4 et 4.2.6 montrent que certains vrais témoignages sont considérés comme non crédibles par les fonctions de filtrage. Il s'avère que ces témoignages considérés comme non crédibles portent sur les agents ayant une mauvaise expertise. Comme les agents ont peu d'intérêt à interagir avec eux, ils n'ont que peu d'observations et une SEM importante. Ainsi, les témoignages sont rejetés car trop divergents des rares observations.

De manière générale, la précision décroit rapidement lors des premiers pas de temps avant de remonter progressivement. En effet, les agents débutent avec une SEM importante puis explorent petit à petit à l'aide d'UCB. La précision du filtrage par 10 fautes suit initialement la même décroissance avant de soudainement remonter lors de la détection massive de faux témoignages.

Ces résultats nous permettent de confirmer que nos fonctions de filtrage sont efficaces pour détecter les faux témoignages sans faire un trop grand nombre d'erreurs. Le KL-filtrage et le filtrage par 10 fautes nécessitent quelques observations initiales pour être efficaces. À l'inverse, la 10-stochocratie utilise les témoignages des autres agents afin de décider de la crédibilité de chaque témoignage, réduisant ainsi le besoin en observations directes. Dans les trois cas, l'utilisation de la divergence de Kullback-Leibler entre les observations d'un agent et les témoignages reçus est une mesure efficace de crédibilité. Les agents peuvent donc l'utiliser pour filtrer les faux témoignages et ainsi augmenter la robustesse de leurs fonctions de réputation.

Notons que l'efficacité des fonctions de filtrage dépendent de la fonction de réputation. Il est donc intéressant d'étudier quels paramètres du système influent sur l'efficacité de chaque fonction de filtrage, afin de généraliser ces résultats et déterminer quelle fonction de filtrage est la plus adaptée.

6. Notons que les diffamations qui ne sont pas détectées par le KL-filtrage correspondent en fait à des diffamation envers des agents honnêtes ayant une faible expertise.

3 Robustesse des jeux hédoniques aux manipulations

Classiquement, les systèmes de réputation s'intéressent aux interactions deux-à-deux entre les agents. Toutefois, de manière plus générale, les agents peuvent avoir besoin d'interagir avec plus d'un autre agent. Dans ce cadre, traditionnellement étudié dans le contexte des jeux de coalitions, qu'en est-il de l'honnêteté des agents et sous quelles conditions ces jeux peuvent-ils être robustes à la présence d'agents malhonnêtes ? Nous présentons ici une étude de la robustesse des jeux hédoniques à une forme très générale de manipulation : les attaques Sybil.

3.1 Un modèle de manipulations rationnelles

Afin d'étudier les conditions sous lesquelles un jeu hédonique est robuste aux manipulations, nous nous fondons sur les définitions données au chapitre 2 section 2.2.

Exemple 4.17
L'exemple qui suit sera utilisé comme exemple récurrent dans le reste de cette section. Nous considérons un jeu dont les caractéristiques sont présentées en figure 4.3. Pour des raisons de compacité d'écriture, nous omettons les indices dans les relations de préférence et notons $13m$ *pour la coalition* $\{a_1, a_3, m\}$*. Ici, a signifie « agent honnête », m « agent malhonnête » et s « agent Sybil » (une fausse identité de m comme expliqué en page 89). Pour* a_1*, la coalition* 12 *est préférée à* $13m$ *qui est préférée à son tour à* 13 *et à* $12m$ *(*a_1 *est indifférent à celles-ci). Enfin,* a_1 *préfère la coalition singleton* 1 *à* $1m$ *et à* 123*.*

a_1	$12 \succ 13m \succ 13 \sim 12m \succ 123m \sim 1 \succ 1m \sim 123$
a_2	$12 \sim 23m \succ 123 \sim 2m \succ 12m \succ 23 \succ 123m \sim 2$
a_3	$13 \sim 23m \succ 3m \succ 123 \succ 23 \succ 123m \sim 3 \succ 13m$
m	$1m \succ 2m \succ 3m \succ m \succ 12m \succ 13m \sim 23m \succ 123m$
NS_{HG}	$\Pi_1 = \{12, 3m\}, \Pi_2 = \{13, 2m\}$
UR_{HG}	$\Pi_3 = \{1, 23m\}, \Pi_4 = \{12m, 3\}, \Pi_5 = \{123m\}$

FIGURE 4.3 – Exemple de jeu hédonique avec quatre agents a_1, a_2, a_3, m

Dans la suite, nous considérons comme concept de solution la stabilité au sens de Nash. Ce concept de solution peut sembler restrictif [7] mais capture les cas où les agents sont pleinement autonomes : ils ne peuvent pas être contraints de quitter ou rester dans une coalition. Dans la suite, nous nous référons à des *partitions stables* pour stables au sens de Nash, et l'ensemble de ces partitions stables est noté NS_{HG}. Pour rappel, une

7. Le lecteur intéressé pourra trouver dans [Vallée, 2015, Vallée *et al.*, 2014c] des résultats où nous relâchons cette hypothèse et étudions d'autres concepts de solution.

partition est stable si aucun agent ne désire changer unilatéralement de coalition au sein de cette partition. Ceci permet aux agents de toujours rejoindre la coalition singleton s'ils le désirent. C'est pourquoi, nous considérons des préférences sous forme de listes rationnelles : toutes les coalitions moins préférées que la coalition singleton n'ont pas besoin d'être représentées.

Définition 4.18
Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. Une partition Π de N est stable au sens de Nash *si, et seulement si : $\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\}, C \cup \{a_i\} \succ_i C_i^{\Pi}$.*

Exemple 4.19
Reprenons l'exemple 4.17. Les deux partitions stables au sens de Nash du jeu sont indiquées sur la ligne NS_{HG} de la figure 4.3.

Comme vu au chapitre 2 section 2.2, il y plusieurs manières de calculer l'issue du jeu à partir de NS_{HG}. Pour des besoins de généralité, nous faisons l'hypothèse suivante.

Hypothèse 4.20
L'issue du jeu G est tiré uniformément parmi les partitions de NS_{HG}.

Afin d'étudier la robustesse de ces jeux aux manipulations, nous considérons des agents malhonnêtes capables de réaliser des *attaques Sybil* [Douceur, 2002]. Une attaque Sybil consiste à apparaître dans le système sous plusieurs fausses identitées, chacune exprimant stratégiquement des préférences qui lui sont propres. Une fois le jeu résolu, l'agent malhonnête peut choisir quelle identité il endossera réellement (en quittant le système pour les autres ou en simulant une défaillance) et, donc, quelle coalition il pourra rejoindre. Cette attaque est une généralisation des manipulations classiques car nous ne faisons pas d'hypothèse ni sur le nombre d'agents Sybil, ni sur la connaissance qu'a l'agent malhonnête sur le jeu (il peut ne pas connaître le nombre d'agents, ni le profil de préférences). Une manipulation classique peut être représentée par une attaque Sybil avec aucun agent Sybil et simplement l'agent malhonnête qui ment sur ses préférences.

Définition 4.21
Soient $HG = \langle N, \succeq \rangle$ un jeu et $m \in N$ un agent. Une attaque Sybil *sur HG par m est définie par un ensemble de nouveaux agents $\{s_1, \ldots, s_k\}$, appelés* agents Sybil*, une relation de préférence $\succeq'_m$ pour m et une relation de préférence $\succeq'_{s_i}$ pour chaque agent Sybil s_i.*

Comme de nouveaux agents sont introduits dans le système par la manipulation, nous devons faire des hypothèses sur le nouveau profil de préférences (noté $\succeq'_i$) de chaque agent honnête a_i puisque de nouvelles coalitions peuvent être formées. En premier lieu, nous considérons l'indépendance des alternatives non-pertinentes [Arrow, 1963] qui impose que si un agent préfère C_1 à C_2, l'arrivée d'un nouvel agent ne modifie pas cette préférence. La seconde hypothèse modélise une acceptation *a priori* des agents honnêtes pour les agents inconnus : un agent honnête accepte au plus un agent inconnu dans sa coalition.

Hypothèse 4.22 (Indépendance des alternatives non-pertinentes)

$\forall C_1, C_2 \subseteq N, \forall a_i \in C_1 \cap C_2 : \; C_1 \succeq_i C_2 \Leftrightarrow C_1 \succeq'_i C_2$

Hypothèse 4.23 (Bénéfice du doute)

$\forall C \subseteq N, \forall a_i \in C, \forall u \notin N : \; C \sim'_i C \cup \{u\}$

Si l'hypothèse 4.22 est une hypothèse classique, l'hypothèse 4.23 peut paraître très favorable aux agents malhonnêtes, même si elle ne permet pas à plusieurs agents inconnus de rejoindre une coalition. En effet, elle ne s'applique qu'aux $C \subseteq N$ et $C \cup \{u\} \not\subseteq N$. Cependant, elle est intéressante pour deux raisons. Premièrement, si un jeu est robuste dans un contexte favorable, il le sera d'autant plus sous des hypothèses moins favorables. Deuxièmement, l'hypothèse 4.23 est nécessaire au bon fonctionnement d'un système ouvert car il convient de permettre aux nouveaux agents de coopérer avec ceux qui sont déjà présents. Notons que nous relâchons cette hypothèse dans la section 3.3.

Exemple 4.24
Si un nouvel agent u rejoint le jeu de la figure 4.3, alors $\succeq'_1$ devient :

$$12 \sim'_1 12u \succ'_1 13m \sim'_1 13mu \succ'_1 13 \sim'_1 13u \sim'_1 12m \ldots 123 \sim'_1 123u$$

Ainsi,

Définition 4.25
Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique avec $N = \{a_1, \ldots, a_n, m\}$. Un jeu HG' résulte *d'une attaque Sybil $(\{s_1, \ldots, s_k\}, \succeq'_m, (\succeq'_{s_1}, \ldots, \succeq'_{s_k}))$ sur HG par m, si il est de la forme $HG' = \langle N \cup \{s_1, \ldots s_k\}, (\succeq'_1, \ldots, \succeq'_n, \succeq'_m, \succeq'_{s_1}, \ldots, \succeq'_{s_k}) \rangle$ où, pour $i = 1, \ldots, n$, $\succeq'_i$ satisfait les hypothèses 4.22 et 4.23.*

Nous nous intéressons aux agents malhonnêtes *rationnels* au sens où ils ne réalisent une attaque que si, et seulement si, ils préfèrent l'issue du jeu résultant à l'issue du jeu initial. Au regard de l'hypothèse 4.20, une manipulation n'est *efficace* que si elle augmente la proportion de partitions stables *satisfaisantes* où satisfaisante est définie par rapport à une *coalition cible* C_θ, représentant la coalition minimalement préférée que l'agent m désire rejoindre. Cette coalition C_θ est une entrée qui modélise l'intention de l'agent. Par exemple, si C_θ est la coalition maximalement préférée de m alors m désire augmenter ses chances d'être précisement dans cette coalition ; si C_θ est la première coalition préférée à la coalition singleton $\{m\}$ alors m désire simplement augmenter ses chances de ne pas être seul.

Définition 4.26
Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. Une partition Π de N est satisfaisante *pour m relativement à une coalition cible C_θ si $\Pi \in NS_{HG}$ et $C^\Pi_m \succeq_m C_\theta$ est vérifié. L'ensemble de toutes les partitions stables satisfaisantes est noté NS^θ_{HG}.*

Exemple 4.27
Sur la figure 4.3 et pour $C_\theta = 1m$, aucune partition stable au sens de Nash n'est satisfaisante pour m. Pour $C_\theta = 3m$, les deux partitions le sont.

Cependant, dans un jeu HG' résultant d'une manipulation, m est présent sous plusieurs identitées. Intuitivement, si m désire rejoindre une coalition C, il sera autant satisfait s'il le fait avec sa véritable identité m ou avec l'une de ses fausses identités $s_1, \ldots, s_k$. Nous redéfinissons la notion de satisfaction pour HG' comme suit.

Définition 4.28
Soit $HG' = \langle N \cup \{s_1, \ldots, s_k\}, \succeq' \rangle$ un jeu résultant d'une manipulation. Une partition Π' est satisfaisante *relativement à une coalition cible C_θ si $\Pi' \in NS_{HG'}$ et que soit $C_m^{\Pi'} \succeq_m C_\theta$, soit $\exists s_i \in \{s_1, \ldots, s_k\}, C_{s_i}^{\Pi'} \cup \{m\} \setminus \{s_i\} \succeq_m C_\theta$.*

Il faut bien remarquer que la satisfaction de m dans le jeu résultant HG' s'appuie sur ses *préférences initiales* $\succeq_m$ dans le jeu HG. Ainsi, une coalition contenant plusieurs identités de m ne peut pas être satisfaisante. Ceci représente le fait que m ne peut pas agir simultanéement avec toutes ses identités pour participer correctement aux coalitions. Il doit faire défection, c'est-à-dire quitter le système, une fois les coalitions formées et nous faisons l'hypothèses que ces défections n'affectent pas les autres coalitions (elles arrivent par exemple suffisament tard pour que le jeu ne soit pas réinitialisé).

Définition 4.29
Soit HG un jeu hédonique, m un agent malhonnête et HG' un jeu résultant d'une manipulation sur HG par m. Soit C_θ la coalition cible de m et r_θ^{HG} la proportion $|NS_{HG}^\theta|/|NS_{HG}|$ (avec par convention $r_\theta^{HG} = 0$ si $|NS_{HG}| = 0$). La manipulation est efficace *relativement à C_θ si $r_\theta^{HG'} > r_\theta^{HG}$.*

Remarquons que si C_θ est la coalition singleton $\{m\}$, alors toutes les partitions stables sont satisfaisantes. Remarquons également que si toutes les partitions stables sont satisfaisantes alors $r_\theta^{HG} = 1$ et aucune manipulation ne peut être (strictement) efficace.

3.2 Caractérisation formelle des manipulations

La première manipulation que nous considérons est une **attaque Sybil constructive** : l'agent malhonnête manipule le jeu en créant de nouvelles partitions stables qui vont modifier la proportion de partitions satisfaisantes. Pour cela, au sein de chaque partition non stable, nous pouvons distinguer les agents qui ne désirent pas changer de coalition et ceux qui le désirent. Ces derniers sont dit *responsables* de l'instabilité de la partition.

Définition 4.30
Soit HG un jeu hédonique, a_i un agent et Π une partition non stable. L'agent a_i est responsible *de l'instabilité de Π s'il existe une coalition $C \in \Pi$ telle que $C \cup \{a_i\} \succ_i C_i^\Pi$. Une telle coalition est dite* attractive *pour l'agent a_i.*

Notons UR_{HG} l'ensemble de toutes les partitions qui sont instables de par la seule responsabilité de l'agent malhonnête et UR^{θ}_{HG} les partitions de UR_{HG} qui contiennent une coalition C satisfaisante et attractive pour pour m, c'est-à-dire $C \cup \{m\} \succ_m C^{\Pi}_m$ et $C \cup \{m\} \succeq_m C_\theta$).

Exemple 4.31
La ligne UR_{HG} de la figure 4.3 indique les partitions instables dont l'agent m est le seul responsable. Pour $C_\theta = 1m$ ou $2m$, $UR^{\theta}_{HG} = \{\Pi_3\}$, et pour $C_\theta = 3m$, $UR^{\theta}_{HG} = \{\Pi_3, \Pi_4\}$.

La [8] manipulation constructive s'appuie sur le fait que l'agent malhonnête est l'unique responsable de l'instabilité de certaines partitions et que les coalitions attractives dans ces dernières sont pourtant satisfaisante pour lui. L'agent malhonnête peut alors manipuler le jeu en état indifférent à toutes les coalitions et en introduisant une unique identité qui exprime ses préférences initiales pour tirer partie de l'hypothèse 4.23.

Définition 4.32
Soit $HG = \langle \{a_1, \ldots, a_n, m\}, \succeq \rangle$ un jeu hédonique. La manipulation constructive *de HG par m est une manipulation impliquant un seul agent Sybil s, dans laquelle m exprime la relation de préférence $\succeq'_m := \succeq^{indif}_m$ où $C_1 \sim^{indif}_m C_2$ pour toutes $C_1, C_2 \ni m$, et s la relation de préférence $\succeq'_s := \succeq_m [m/s]$ où $\succeq_m [m/s]$ est la relation $\succeq_m$ qui substitue s à m.*

Remarquons que l'agent Sybil exprime le fait qu'il ne *désire pas* rejoindre la coalition de m (puisque m est remplacé par s dans $\succeq'_s$).

Exemple 4.33
Sur la figure 4.3, la manipulation constructive introduit un agent Sybil s avec $1s \succ'_s 2s \succ'_s 3s \succ'_s s$. Les partitions stables au sens de Nash avant et après la manipulation sont représentées en figure 4.4.

La question que nous nous posons est désormais la suivantes : sous quelles conditions sur le jeu initial la manipulation constructive est-elle efficace ? Examinons en premier lieu sous quelles conditions un agent peut désirer changer de coalition dans le jeu résultant de la manipulation. Trivialement, m ne désire jamais changer de coalition puisqu'il est indifférent à toutes. Fixons un jeu $HG = \langle N, \succeq \rangle$, un agent malhonnête $m \in N$ et une partition Π. Notons HG' le jeu résultant de la manipulation constructive de HG par m, et $\Pi' = \Pi[s \to C_0]$ la partition de HG' obtenue à partir de Π lorsque l'agent Sybil s rejoint une $C_0 \in \Pi \cup \{\emptyset\}$, c'est-à-dire $\Pi' = \Pi \setminus \{C_0\} \cup \{C_0 \cup \{s\}\}$.

8. Le terme « la » est un abus de langage pour référer à cette manipulation car il existe bien évidemment de nombreuses autres manipulations constructives, tout comme cela est le cas pour la manipulation destructive présentée page 96. Toutefois, nous montrons en section 3.3 que les manipulations que nous considérons présentent une forme de canonicité qui justifie cet abus de langage.

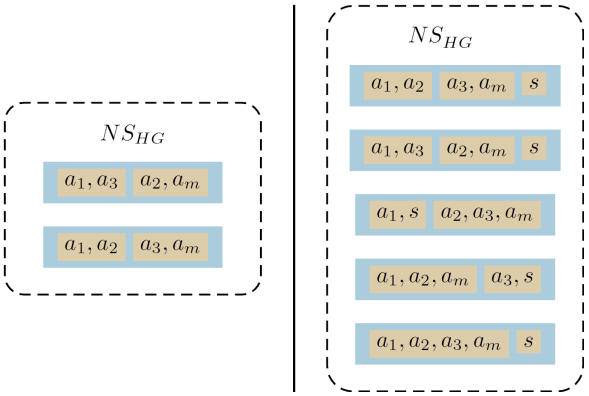

FIGURE 4.4 – Partitions stables avant puis après la manipulation constructrive

Lemme 4.34
Un agent honnête a ne désire changer de coalition dans Π' que si, et seulement si, il désire changer de coalition dans Π.

Démonstration 4.34
Par définition, a désire changer de coalition dans Π' si, et seulement si, il existe $C' \in \Pi'$ telle que $C' \cup \{a\} \succ'_a C_a^{\Pi'}$. Selon l'hypothèse 4.23, nous avons $C' \cup \{a\} \sim'_a C' \cup \{a\} \setminus \{s\}$ et $C_a^{\Pi'} \sim'_a C_a^{\Pi'} \setminus \{s\}$. Ainsi, $C' \cup \{a\} \succ'_a C_a^{\Pi'}$ est équivalent à $C' \cup \{a\} \setminus \{s\} \succ'_a C_a^{\Pi'} \setminus \{s\}$. Cependant, $C_a^{\Pi'} \setminus \{s\}$ est précisément C_a^{Π} et $C' \setminus \{s\}$ est dans Π. Donc, a désire rejoindre $C' \in \Pi'$ si, et seulement si, il désire rejoindre $C' \setminus \{s\} \in \Pi$. □

Lemme 4.35
L'agent Sybil s ne désire changer de coalition in Π' que si, et seulement si, $m \in C_0$ ou s'il existe $C \in \Pi$ telle que $m \notin C$ et $C \cup \{m\} \succ_m C_0 \cup \{m\}$.

Démonstration 4.35
($\Rightarrow$) Supposons que s désire quitter $C_0 \cup \{s\}$ et rejoindre $C' \in \Pi'$. Supposons alors que $m \notin C_0$ et qu'il existe $C \in \Pi$ telle que $m \notin C$ et $C \cup \{m\} \succ_m C_0 \cup \{m\}$. Comme s désire changer de coalition, nous avons $C' \cup \{s\} \succ'_s C_0 \cup \{s\}$. Or, par définition de $\succeq'_s$, nous avons $m \notin C'$ et $C' \cup \{m\} \succ_m C_0 \cup \{m\}$, c'est-à-dire comme indiqué dans le lemme.
($\Leftarrow$) Si $m \in C_0$, alors s désire changer de coalition dans Π', pour au minimum rejoindre $\{s\}$). S'il existe $C \in \Pi$ telle que $m \notin C$ et $C \cup \{m\} \succ_m C_0 \cup \{m\}$, alors par définition de $\succeq'_s$, s désire rejoindre la coalition $C \cup \{s\}$ de Π'. □

Les lemmes 4.34 et 4.35 nous permettent de déduire le corollaire suivant.

Corollaire 4.36
Une partition $\Pi' = \Pi[s \to C_0]$ est stable dans HG' si, et seulement si, $m \notin C_0$, $C_0 \cup \{m\}$ est maximalement préférée par m dans Π, et soit (1) Π est stable, soit (2) m est l'unique responsable de l'instabilité de Π.

Exemple 4.37
Sur la figure 4.3, pour $C_\theta = 1m$, $\Pi'_3 = \Pi_3[s \to 1] = \{1s, 23m\}$ *est satisfaisante et* $\Pi'_1 = \Pi_1[s \to 12] = \{12s, 3m\}$ *est stable mais non satisfaisante.*

Nous pouvons maintenant donner les conditions exactes sous lesquelles la manipulation constructive est efficace dans un jeu HG. Trivialement, si $NS^\theta_{HG} = NS_{HG}$, l'agent malhonnête est déjà pleinement satisfait et la manipulation ne peut être (strictement) efficace. Il reste deux autres cas : $NS^\theta_{HG} = \emptyset$ et $NS^\theta_{HG} \neq \emptyset$.

Proposition 4.38
Supposons $NS^\theta_{HG} = \emptyset$. *La manipulation constructive est efficace sur* HG *si, et seulement si :*

- *soit* $NS^\theta_{HG} = \emptyset$ *et* $UR^\theta_{HG} \neq \emptyset$,
- *soit* $NS^\theta_{HG} \neq \emptyset$ *et* $|UR^\theta_{HG}|/|UR_{HG}| > |NS^\theta_{HG}|/|NS_{HG}|$

Démonstration 4.38

1. *Supposons* $NS^\theta_{HG} = \emptyset$. *Par définition de* r^{HG}_θ, *si la manipulation est efficace, alors* HG' *a au moins une partition satisfaisante* $\Pi' = \Pi[s \to C_0]$. *Du corollaire 4.36, il s'ensuit que* $\Pi \in NS_{HG}$ *ou* $\Pi \in UR_{HG}$. *Comme* Π' *est satisfaisante, nous avons soit* $C_0 \cup \{m\} \succeq_m C_\theta$, *soit* $C^\Pi_m \succeq_m C_\theta$. *Dans les deux cas, comme* $NS^\theta_{HG} = \emptyset$, *nous avons* $\Pi \notin NS_{HG}$, *et* $\Pi \in UR_{HG}$. *Par conséquent,* $\Pi \in UR^\theta_{HG}$. *L'autre sens de l'implication se démontre de manière similaire.*
2. *Supposons* $NS^\theta_{HG} \neq \emptyset$ *et comparons les proportions de partitions satisfaisantes dans* HG *et* HG'. *Trivialement,* $r^{HG}_\theta = |NS^\theta_{HG}|/|NS_{HG}|$. *De plus, selon le corollaire 4.36 et comme* $\succeq_m$ *est un ordre total, il s'en suit que* $|NS_{HG'}| = |NS_{HG}| + |UR_{HG}|$. *Comptons combien d'entre elles sont satisfaisantes. Considérons en premier lieu une partition stable* $\Pi \in NS_{HG}$. *Comme précédemment, il y a donc une partition stable* Π' *de* HG' *de la forme* $\Pi[s \to C_0]$. *Comme* Π *est stable,* $C_0 \cup \{m\} \not\succeq_m C^\Pi_m$ *et donc* Π' *est satisfaisante dans* HG' *si, et seulement si,* Π *est satisfaisante dans* HG. *Considérons maintenant une partition non-stable* $\Pi \notin NS_{HG}$. *Du corollaire 4.36, il s'en suit que* $\Pi' = \Pi[s \to C_0]$ *est stable dans* HG' *si, et seulement si,* m *est l'unique responsable de la non-stabilité de* Π *et* HG, *et que* $C_0 \cup \{m\}$ *est maximalement préférée selon* $\succeq_m$. *Ainsi,* Π' *est stable* HG' *si, et seulement si,* $C_0 \cup \{m\} \succeq C_\theta$, *c'est-à-dire si* $\Pi \in UR^\theta_{HG}$. *Par conséquent, le nombre de partitions satisfaisantes dans* HG' *est* $|NS^\theta_{HG}| + |UR^\theta_{HG}|$, *et la manipulation n'est efficace que si, et seulement si :*

$$\frac{|NS^\theta_{HG}| + |UR^\theta_{HG}|}{|NS_{HG}| + |UR_{HG}|} > \frac{|NS^\theta_{HG}|}{|NS_{HG}|} \text{ , } i.e., \frac{|UR^\theta_{HG}|}{|UR_{HG}|} > \frac{|NS^\theta_{HG}|}{|NS_{HG}|}$$

□

Exemple 4.39
Sur la figure 4.3, la manipulation constructive est efficace pour $C_\theta = 1m$ *car* $NS^\theta_{HG} = \emptyset$ *et* $UR^\theta_{HG} = \{\Pi_3\}$. *Cependant, elle n'est pas efficace pour* $C_\theta = 2m$ *car* $NS^\theta_{HG} = \{\Pi_2\}$ *et* $UR^\theta_{HG} = \{\Pi_3\}$, *et donc* $r^{HG}_\theta = 1/2$ *et* $r^{HG'}_\theta = 2/5$.

Proposition 4.40
Soit un jeu HG avec des préférences représentées par RIRLC, un agent m et une coalition C_θ, décider si la manipulation constructive est efficace sur HG pour m relativement à la coalition cible C_θ est un problème NP-dur.

Démonstration 4.40
Nous faisons une réduction depuis un problème consistant à décider si un jeu HG_0 avec des préférences sous forme RIRLC possède au moins une partition stable au sens de Nash, qui est un problème NP-complet [Ballester, 2004]. À partir de HG_0, nous construisons un jeu HG tel que $NS^\theta_{HG} = \emptyset$ et $UR^\theta_{HG} \neq \emptyset$ si, et seulement si, HG_0 possède une partition stable. De la proposition 4.38, il s'en suit que la manipulation constructive est efficace sur HG si, et seulement si, HG_0 possède une partition stable.

Soit $HG_0 = \langle N_0, \succeq_0 \rangle$ with $N_0 = \{a_1, \ldots, a_n\}$. Le jeu HG est défini à partir de HG_0 en y ajoutant deux nouveaux agents, a et m avec les relations de préférence suivantes : $\{a, m\} \succ_m \{m\}$, $\{a\} \succ_a C$ pour toutes les coalitions $C \neq \{a\}$, et $\succeq_i$ construit à partir de $(\succeq_0)_i$ à l'aide des hypothèses 4.22 and 4.23.

Intuitivement, a désire être dans sa coalition singleton et m désire rejoindre a. Les autres agents sont indifférents vis-à-vis d'eux, et conservent leurs préférences de HG_0 dans les autres cas. Trivialement, HG peut être construit en temps polynomial en la taille de HG_0. Enfin, fixons $C_\theta = \{a, m\}$ la coalition cible. Aucune partition Π n'est stable dans HG car si a n'est pas dans $\{a\}$, il désire la rejoindre tandis que si il est dans $\{a\}$, alors m désire le rejoindre.

Supposons qu'il existe une partition stable Π_0 dans HG_0 telle que $\Pi = \Pi_0 \cup \{\{a\}, \{m\}\}$ dans HG. Alors, m est trivialement l'unique responsable de la non-stabilité de Π. De plus, dans Π, la coalition attractive pour m est satisfaisante. Ainsi, $\Pi \in UR^\theta_{HG}$. Dualement, si toutes les partitions Π_0 de HG_0 sont non stables, alors comme $a_1, \ldots, a_n$ sont indifférents à a, m, toutes les partitions impliquant a, m ne sont pas stables elles-aussi. Par conséquent, HG n'a pas de partition stable. Donc $UR^\theta_{HG} \neq \emptyset$ si, et seulement si, HG_0 possède une partition stable, comme nous le désirions. □

Bien que cette manipulation constructive est indépendante des préférences des agents honnêtes, décider si cette manipulation est efficace, en plus d'être difficile à calculer, nécessite de les connaître. Ceci est d'autant plus important qu'une manipulation constructive non efficace peut (strictement) empirer la situation de m (voir l'exemple 4.39).

Nous nous intéressons maintenant à une seconde forme de manipulation, une **attaque Sybil destructive** dans le sens où elle rend instables des partitions stables non satisfaisantes pour l'agent malhonnête. Elle repose sur le fait que, dans le cas de la stabilité au sens de Nash, un unique veto d'un agent permet de refuser une coalition, et donc de rendre une partition instable. Cette manipulation va donc s'appuyer sur une unique fausse identité qui va poser son veto sur toutes les partitions qui ne satisfons par l'agent malhonnête.

Définition 4.41
Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. La manipulation destructive *de HG' par m utilise un unique agent Sybil s, où $\succeq'_m := \succeq_m$, et $\succeq'_s$ est définie telle que, pour chaque $C \subseteq N$, si $m \in C$ et $C \not\succeq_m C_\theta$ alors $C \cup \{s\} \succ'_s \{s\}$, sinon $\{s\} \succ'_s C \cup \{s\}$.*

Le point essentiel est que $\{s\} \succ'_s C_\theta \cup \{s\}$ tandis que les préférences relatives entre les coalitions est arbitraire. Informellement, l'agent Sybil désire rejoindre toutes les coalitions qui contiennent m et qui ne sont pas péférées à C_θ. Comme m ne désire pas être avec s, toutes les partitions contenant ces coalitions deviennent instables.

Exemple 4.42
Sur la figure 4.3, pour $C_\theta = 2m$, les préférences de s satisfont :

$$3ms, ms, 12ms, 13ms, 23ms, 123ms \succ'_s s$$

Les partitions stables au sens de Nash avant et après la manipulation destructive sont représentées en figure 4.5.

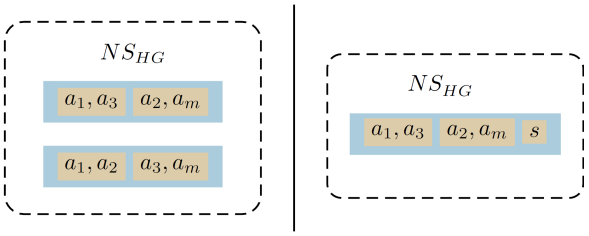

FIGURE 4.5 – Partitions stables avant puis après la manipulation destructive

Lemme 4.43
Soit un jeu HG et un agent malhonnête m. Soit HG' le jeu résultant de la manipulation destructive de HG par m.

1. *HG' a une partition satisfaisante si, et seulement si, il y en a une dans HG,*
2. *toutes les partitions stables de HG' sont satisfaisantes pour m.*

Démonstration 4.43
1. *($\Rightarrow$) Supposons que $\Pi' = \Pi[s \to C_0]$ est satisfaisante dans HG'. Si m est dans une coalition satisfaisante de Π', alors Π est satisfaisante dans HG. Dans le cas contraire, seule $C_0 \cup \{s\}$ est satisfaisante dans Π', mais par définition de $\succeq'_s$, s désire alors rejoindre la $C_m^{\Pi'}$, ce qui contredit la stabilité de Π'.*
   ($\Leftarrow$) Si Π est satisfaisante dans HG, alors $\Pi \cup \{\{s\}\}$ est satisfaisante dans HG'.
2. *Soit Π' une partition de HG' stable mais non satisfaisante. Par définition de $\succeq'_s$, s est alors dans la même coalition que m. Cependant, m préfère alors rejoindre sa coalition singleton $\{m\}$, ce qui contredit l'hypothèse de stabilité de Π'.*

□

De manière intéressante, lorsque la manipulation destructive est efficace, elle l'est *complètement* : toutes les partitions stables sont aussi satisfaisantes. Par conséquent, trivialement, la manipulation est efficace s'il existe au moins une partition satisfaisante et une partition non satisfaisante (sans ce dernier point, l'agent malhonnête est déjà satisfait).

Proposition 4.44
La manipulation destructive est efficace *sur HG si, et seulement si, HG a au moins une partition satisfaisante et au moins une partition stable non satisfaisante.*

Exemple 4.45
Sur la figure 4.3, la manipulation destructive est efficace pour $C_\theta = 2m$ *car* Π_2 *est satisfaisante tandis que* Π_1 *ne l'est pas : la seule partition stable de* HG' *est donc* $\Pi_2[s \to \emptyset] = \{13, 2m, s\}$. *En revanche, la manipulation destructive n'est pas efficace ni pour* $C_\theta = 3m$ *(*m *est déjà satisfait de HG), ni pour* $C_\theta = 1m$ *(*$r_\theta^{HG} = 0$*).*

Comme pour la manipulation constructive, il est difficile de décider sur la manipulation destructive est efficace et cela requière des connaissances sur le jeu HG. Toutefois, contrairement à la manipulation constructive, la manipulation destructive ne peut jamais strictement dégrader la situation de m (car seules des partitions non satisfaisantes sont rendues instables).

Proposition 4.46
Soit un jeu HG avec des préférences représentées par RIRLC, un agent m *et une coalition* C_θ*, décider si la manipulation destructive est efficace sur HG pour* m *relativement à la coalition cible* C_θ *est un problème NP-dur.*

Démonstration 4.46
La preuve est similaire à celle de la proposition 4.40. Soit $HG_0 = \langle N_0, \succeq_0 \rangle$*, nous construisons un HG avec à la fois une partition stable et satisfaisante, et une partition stable et non satisfaisante, si, et seulement si,* HG_0 *a une partition stable. Le jeu HG est construit à partir de* HG_0 *en ajoutant trois agents,* a*,* a' *et* m *avec les relations de préférence* $\{a, a', m\} \succ_a \{a\}$ *pour* $a \in \{a, a', m\}$ *et, pour tous les* a_i*,* $\succeq_i$ *est construit à partir de* $(\succeq_0)_i$ *selon les hypothèses 4.22 et 4.23. Intuitivement,* a*,* a' *et* m *désirent être ensemble ou séparés tandis que les autres agents sont indifférents. Enfin, fixons* $C_\theta = \{a, a', m\}$.

Soit Π *une partition de HG.* Π *n'est pas stable si* a *ou* a' *ou* m *est avec un agent* a_i *(car dans ce cas, il préfère être seul). La partition n'est pas stable non plus si exactement deux d'entre eux sont ensemble. Dans les deux cas restants, soit chacun est dans sa coalition singleton, soit ils sont ensemble. Nous pouvons voir que* Π *est stable si, et seulement si, la partition* $\Pi \setminus \{\{a\}, \{a'\}, \{m\}, \{a, a', m\}\}$ *est stable dans* HG_0*. De plus, bien que les deux partitions sont stables, seule celle contenant la coalition* $\{a, a', m\}$ *est satisfaisante pour* m*, comme désiré.* □

3.3 Robustesse pour le cas de la stabilité au sens de Nash

Dans cette section, nous montrons que les deux manipulations présentées précédemment suffisent à étudier la robustesse de jeux hédoniques dans le cas de la stabilité au sens de Nash. En effet, ces deux manipulations présentent une forme de canonicité dans le sens où si un jeu n'est pas manipulable (efficacement) par la manipulation constructive ou la manipulation destructive, alors il n'est manipulable par aucune attaque utilisant au plus une fausse identité.

Proposition 4.47
Soit HG un jeu hédonique avec pour concept de solution la stabilité au sens de Nash, m un agent malhonnête et C_θ une coalition cible pour m. Si ni la manipulation constructive, ni la manipulation destructive ne sont efficaces sur HG, alors aucune attaque Sybil utilisant au plus une fausse identité n'est efficace sur HG.

Démonstration 4.47
Supposons qu'il existe une manipulation efficace M mais que la manipulation destructive ne l'est pas. Nous montrons alors que la manipulation constructive est efficace.

Comme la manipulation destructive n'est pas efficace, la proposition 4.44 implique que soit toutes les partitions stables de HG sont satisfaisantes, soit aucune d'entre elles ne l'est (satisfaisante). Dans le premier cas, M ne peut pas être efficace car l'agent malhonnête est déjà satisfait, ce qui contredit l'hypothèse. Ainsi, HG n'a aucune partition stable satisfaisante.

Notons HG' le jeu résultant de la manipulation M, et s l'agent Sybil utilisé dans M. Comme M est efficace, il existe alors au moins une partition Π' stable et satisfaisante dans HG'. Soit Π la partition $\{C' \setminus \{s\} \mid C' \in \Pi'\}$. Nous montrons alors que soit Π est satisfaisante dans HG, ce qui conduit à une contradiction, soit $\Pi \in UR^{\theta}_{HG}$. Remarquons en premier lieu qu'aucun agent honnête a_i ne désire changer de coalition dans Π car, dans le cas contraire, selon l'hypothèse 4.23, a_i désire aussi changer de coalition dans Π', ce qui contredit la stabilité de Π'. En ce qui concerne m, nous distinguons deux cas.

1. *Supposons qu'aucune coalition de Π' n'est préférée par m à celle dans laquelle il est. Plus précisement, pour toute coalition $C \in \Pi'$, $C \setminus \{s\} \cup \{m\} \not\succeq_m C^{\Pi'}_m \setminus \{s\}$. Dans ce cas, m ne désire pas changer de coalition dans Π. Donc, Π est stable. De plus, comme m est dans sa coalition préférée dans Π' et comme Π' est satisfaisante, alors Π est satisfaisante aussi, ce qui est une contradition.*
2. *C'est pourquoi, il y a une coalition $C \in \Pi'$ telle que $C \setminus \{s\} \cup \{m\} \succeq_m C^{\Pi'}_m \setminus \{s\}$, et m désire la rejoindre. De plus, comme Π' est satisfaisante, C est aussi satisfaisante et, donc, $\Pi \in UR^{\theta}_{HG}$. Comme HG ne dispose d'aucune partition satisfaisante ($NS^{\theta}_{HG} = \emptyset$), alors selon la proposition 4.38 la manipulation constructive est efficace.*

□

La proposition 4.47 caractérise ainsi complètement les conditions sous lesquelles un jeu hédonique avec pour concept de solution la stabilité au sens de Nash est manipulable par une attaque Sybil impliquant au plus une fausse identité. De plus, ce résultat peut être étendu aux manipulations impliquant plus d'un agent Sybil à condition d'étendre l'hypothèse 4.23 à un groupe d'agents, c'est-à-dire faire l'hypothèse que les agents honnêtes sont indifférents à un nombre arbitraire d'agents inconnus. Ceci est particulièrement intéressant car cela signifie que, sous cette hypothèse, utiliser plusieurs fausses identités n'aide pas plus dans le cas général qu'en utiliser une seule.

Il est aussi intéressant de relâcher l'hypothèse 4.23. Nous pouvons le faire en consiérant à la place une forme de *sous-additivité faible* : les agents honnêtes préfèrent que les agents inconnus ne les rejoignent pas tout en maintenant leurs préférences sur les sous-ensembles d'agents connus.

Hypothèse 4.48 (Sous-additivité faible)
$\forall C_1, C_2 \subseteq N$, $\forall a_i \in N$ *tel que* $C_1 \succeq_i C_2, \forall u \notin N$, *nous avons* $C_1 \succeq'_i C_1 \cup \{u\} \succeq'_i C_2$.

Le lemme 4.34 de la manipulation constructive et la caractérisation de la proposition 4.47 sont fondés sur le fait qu'un agent honnête a désire rejoindre la coalition $C' \in \Pi'$ si, et seulement si, il désire rejoindre $C' \setminus \{s\} \in \Pi$. Comme cela est toujours vrai sous l'hypothèse 4.48 et que les autres résultats n'utilisent pas l'hypothèse 4.23, tous nos résultats sur la stabilité au sens de Nash restent vrais en relâchant l'hypothèse. Cela est aussi vrai si nous considérons une hypothèse duale de *super-additivité faible* – $C_1 \cup \{u\} \succeq'_i C_1 \succeq'_i C_2 \cup \{u\} \succeq'_i C_2$ – bien que de manière intuitive cette super-additivité aurait semblé être au bénéfice des agents malhonnêtes.

Enfin, si décider de l'efficacié d'une manipulation sur un jeu hédonique reste un problème NP-dur, cela ne signifie pas qu'en moyenne ce problème soit difficile. Par exemple, [Conitzer et Sandholm, 2006] ont montré l'existence d'une procédure permettant de manipuler une règle de vote lorsqu'elle satisfait l'axiome de monotonie faible et que le vote d'agents malhonnêtes en coalition peut faire gagner un candidat parmi deux. Cependant, ils ont également montré empiriquement que les jeux satisfaisant ces conditions sont fréquents. Ainsi, si certaines règles de votes sont dites robustes aux manipulations, car il est NP-difficile de décider d'une manipulation efficace, en pratique ce problème de décision est souvent facile. C'est pourquoi, pour montrer que la stabilité au sens de Nash permet de garantir une robustesse aux manipulations, nous montrons ici que les jeux hédoniques satisfaisant les conditions nécessaires à la mise en œuvre d'une manipulation efficace sont rares. Pour cela, nous estimons empiriquement la probabilité d'existence des jeux hédoniques efficacement manipulables par au moins un agent du système, soit par la manipulation constructive, soit par la manipulation destructive.

Pour cela, nous générons k jeux hédoniques HG où le profil de préférences des n agents est tiré aléatoirement uniformément. Pour chacun de ces jeux hédoniques, nous considérons tour à tour chaque agent a_i et *calculons s'il existe une coalition cible pour*

l'agent qui n'est pas sa coalition singleton et pour laquelle il peut réaliser une manipulation soit constructive, soit destructive efficace. Notons que dans certains cas, ces manipulations sont toutes les deux efficaces. Afin que la confiance en nos résultats soit suffisante, nous fixons pour nos simulations $k = 10\,000$ et considérons le pourcentage de jeux manipulables par au moins un agent. Dans nos simulations, nous faisons varier n entre 3 et 10 agents. Notons que nous ne considérons pas le cas où $n = 2$ car soit les deux agents désirent coopérer et forment la grande coalition, soit l'un des deux agents ne désire pas coopérer et ils forment les coalitions singletons.

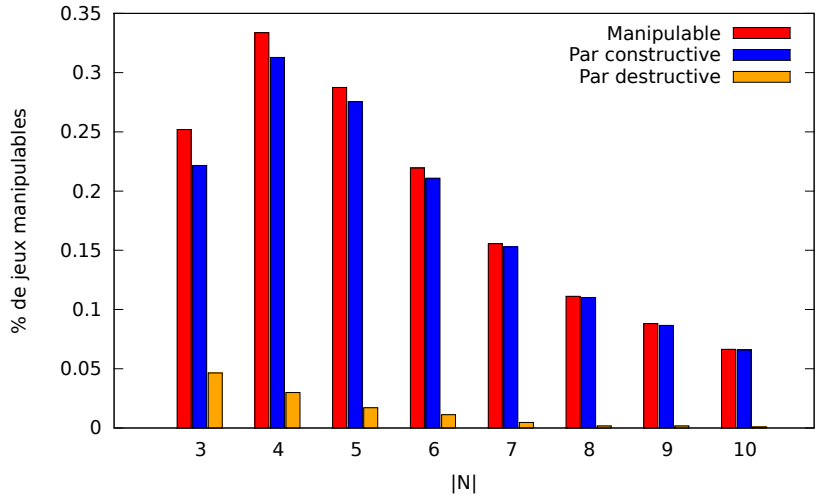

FIGURE 4.6 – Taux de jeux hédoniques manipulables en fonction du nombre d'agents

La figure 4.6 donne le pourcentage de jeux hédoniques où la manipulation constructive ou destructive est k-rationnelle pour au moins un agent $a_i \in N$.

Il est intéressant de constater que, bien qu'en théorie la manipulation destructive est toujours pleinement efficace, les conditions pour qu'elle le soit sont rarement satisfaites en pratique. Par exemple, seuls $1,71$ % des jeux à 5 agents sont manipulables par une manipulation destructive. Ceci est dû aux faits que plus n est important, moins il existe de partitions stables (puisque pour une partition donnée, il est fréquent qu'au moins un agent désire changer de coalition).

La manipulation constructive est plus souvent rationnelle. Par exemple, elle est rationnelle dans environ 11 % des jeux hédoniques à 8 agents. Cependant, comme pour la manipulation destructive, plus n est important, moins il existe de jeux hédoniques manipulables. Ceci s'explique par le fait que plus il y a d'agents participant aux jeux, moins il existe de structures de coalitions où un agent désirant manipuler le jeu est l'unique responsable de la non-stabilité.

Remarquons la présence d'un cas particulier. En effet, le pourcentage de jeux manipulables augmente en passant de jeux à 3 agents à des jeux à 4 agents. Cette augmentation est due au fait qu'il n'existe que 5 structures de coalitions possibles dans les jeux à 3 agents et qu'il est fréquent que la solution soit individuellement optimale pour chaque agent, et qu'aucun n'ait besoin de mettre en œuvre une manipulation. Quoi qu'il en soit, ces simulations mettent en lumière le fait que le nombre de jeux hédoniques manipulables par au moins un agent est relativement faible.

Bilan et animation scientifique

Nous avons présenté dans ce chapitre un exemple de travaux réalisés dans le cadre de l'axe de recherche sur l'honnêteté des agents autonomes. Cet axe s'est construit à partir de travaux préliminaires menés en 2011 dans le cadre du stage de master de Sami Hajlaoui. Ce stage consistait à modéliser des attaques Sybil afin de simuler des comportements de resquillage sur les réseaux pair-à-pair. Ce premier travail, couplé avec le fait que les questions de fiabilité mènent naturellement aux questions d'honnêteté, nous a conduit à développer pleinement cet axe.

Nous avons alors abordé l'honnêteté au travers de la modélisation de la sincérité et la caractérisation de manipulations. Le travail présenté en section 1 a été réalisé dans le cadre de la thèse de Christopher Leturc, débutée en 2016 et co-encadrée avec Bruno Zanuttini (Université de Caen Normandie). Cette proposition a fait l'objet d'une publication internationale [Leturc et Bonnet, 2018a] et de deux publications nationales [Leturc et Bonnet, 2018b, Leturc et Bonnet, 2017]. Les travaux présentés en sections 2 et 3 font suite au stage de Thibaut Vallée en 2012 qui a débouché sur le co-encadrement de sa thèse (2012 – 2015) avec François Bourdon (Université de Caen Normandie). Ces travaux ont aussi fait l'objet de publications internationales [Vallée et Bonnet, 2015, Vallée *et al.*, 2014c] et nationales [Vallée *et al.*, 2015, Vallée *et al.*, 2013].

Nous avons vu au chapitre 1 que l'honnêteté est une valeur morale représentant la qualité d'un agent à agir conformément à une convention pour dire la vérité et faire ce qu'il se doit. Toutefois, un agent malhonnête n'est pas nécessairement malveillant et peut avoir de « bonnes » raisons, liées au contexte, de se comporter ainsi. Travailler sur l'honnêteté nous a alors amené à nous questionner de façon plus générale sur la modélisation de valeurs morales, et de leur respect ou non en fonction d'un contexte, c'est-à-dire en fonction d'une éthique.

Chapitre 5

Troisième axe : représentation de l'éthique

Sommaire

Nous avons vu au chapitre 1 que l'introduction croissante d'agents autonomes artificiels dans certains domaines applicatifs soulevait des questions éthiques. Par exemple, dans le domaine de l'aide à la décision médicale, il est désirable que les agents respectent le code de déontologie médicale. De manière plus prospective, il pourrait être désirable de disposer de voitures autonomes faisant preuve de civilité. Se pose alors la question de concevoir des agents autonomes exhibant des comportements qualifiés d'éthiques, sous-tendus par des valeurs et des principes éthiques et moraux. Pour traiter cette question, nous proposons en section 1 un modèle individuel de jugement éthique fondé sur une architecture BDI. Nous étendons ce modèle en section 2 pour juger les autres agents et construire une notion de confiance en l'éthique d'autrui. Enfin, en section 3, nous nous plaçons dans un contexte collectif et proposons un modèle de jeux hédoniques permettant aux agents d'exprimer une éthique du processus de construction de coalitions. Nous concluons ce chapitre par un bilan de l'animation et l'encadrement scientifique réalisés autour de ce travail.

1 Un modèle de jugement éthique

Comme expliqué au chapitre 1 section 3.3, nous distinguons l'éthique, la morale et les valeurs sur lesquelles ces deux notions s'appuient. De plus, l'éthique est le résultat d'une procédure, une concialiation entre les désirs, la morale, les capacités de l'agent au regard de la situation dans laquelle il se trouve. Pour prendre ces dimensions en compte, nous proposons un modèle de jugement éthique – noté EJP – intégré dans une architecture BDI qui utilise des connaissances sur l'évaluation de situation, la morale et l'éthique. Ce modèle est structuré en quatre sous-modèles comme illustré en figure 5.1 : un modèle de *reconnaissance de situation*, un modèle d'*évaluation*, un modèle *moral* et un modèle *éthique*[1]. Pour des raisons de simplicité, nous considérons ici des raisonnements éthiques à court terme, ne portant que sur un comportement qui se résume à des actions immédiates, et se fondant uniquement sur des états mentaux en faisant abstraction des spécificités de l'implémentation.

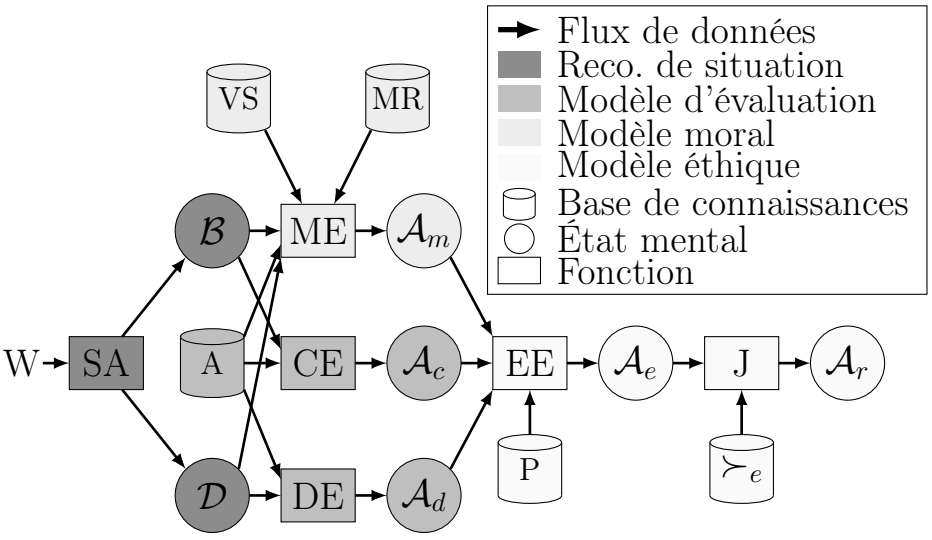

FIGURE 5.1 – Modèle de jugement éthique

Définition 5.1 (Modèle de jugement éthique)
Un modèle de jugement éthique (ou Ethical Judgment Process) EJP est défini comme une composition d'une reconnaissance de situation (AP), un modèle d'évaluation (EP), un modèle moral (GP), un modèle éthique (RP) et une ontologie $\mathcal{O}$ ($\mathcal{O} = \mathcal{O}_v \cup \mathcal{O}_m$) de valeurs morales ($\mathcal{O}_v$) et valuations morales ($\mathcal{O}_m$). Ce modèle de jugement éthique produit une

1. Notons que la dénomination de ces deux modèles est un raccourci par rapport à leur fonction exacte qui est dédiée respectivement à l'évaluation de la moralité (modèle moral) et à l'évaluation du respect de l'éthique (modèle éthique) des actions considérées. Il ne s'agit nullement de prétendre que l'un est un modèle fonctionnant de manière morale et l'autre de manière éthique.

évaluation des actions pour l'état courant du monde W en tenant compte de considérations morales et éthiques.

$$EJP = \langle AP, EP, GP, RP, \mathcal{O} \rangle$$

Ce modèle doit être considéré comme un modèle générique composé de fonctions abstraites, états mentaux et bases de connaissances. Ces fonctions peuvent être implémentées de diverses manières. Par exemple, les valuations morales de $\mathcal{O}$ peuvent prendre la forme d'un ensemble d'éléments discrets tel que { bien, mal } ou continu comme un degré de moralité. Ce modèle de jugement peut être intégré comme nouveau composant du mécanisme de décision d'un agent BDI pour qu'un agent puisse décider de son comportement mais aussi comme composant permettant de juger du comportement des autres agents. Afin de mettre en place cette double utilisation, nous indiçons chacun des ensembles de données entrant dans le modèle par l'agent a_i. Notons que l'ontologie $\mathcal{O}$ de valeurs et de valuations morales n'est pas indicée par a_i car nous considérons qu'elle est commune à l'ensemble des agents du système. Cela permet aux agents d'employer les mêmes noms de valeurs et d'exprimer la moralité de leurs actions sur une même échelle de valuations.

1.1 Reconnaissance de situation et évaluation

Dans ces modèle, l'agent commence par évaluer l'état du monde, c'est-à-dire produire des croyances et désirs en appliquant un modèle de reconnaissance de situation à partir de l'état de l'environnement dans lequel l'agent est situé (l'environnement inclut également les autres agents du système).

Définition 5.2 (Modèle de reconnaissance de situation)
Le modèle de reconnaissance de situation (ou Awareness Process) AP génère l'ensemble des croyances qui décrivent l'état courant du monde W et l'ensemble des désirs qui décrivent les buts de l'agent. Il est défini comme :

$$AP = \langle \mathcal{B}_{a_i}, \mathcal{D}_{a_i}, SA \rangle$$

où $\mathcal{B}_{a_i}$ est l'ensemble des croyances de l'agent a_i sur W parmi l'ensemble B_{a_i} de ses croyances possibles, et $\mathcal{D}_{a_i}$ ses désirs à partir de W parmi l'ensemble D_{a_i} de ses désirs possibles, générés par la fonction SA :

$$SA : W \rightarrow \mathcal{B}_{a_i} \cup \mathcal{D}_{a_i}$$

À partir d'un ensemble de croyances $\mathcal{B}_{a_i}$ et d'un ensemble de désirs $\mathcal{D}_{a_i}$ (a_i comme expliqué ci-dessus peut désigner l'agent effectuant le modèle de jugement – dans ce cas il s'agit de ses propres croyances et désirs – ou un autre agent – dans ce cas il s'agit de la représentation que l'agent effectuant le jugement a sur les croyances et désirs de a_i) un agent exécute le modèle d'évaluation EP pour établir les actions désirables $\mathcal{A}_d$ d'une part (c'est-à-dire les actions qui permettent de satisfaire un désir) et les actions exécutables

$\mathcal{A}_c$ d'autre part (c'est-à-dire les actions pouvant être effectuées dans l'état courant du monde). Ces actions sont déduites par raisonnement sur les conditions et conséquences des actions décrites dans A_{a_i}, c'est-à-dire les actions à disposition de l'agent effectuant le jugement si a_i représente cet agent, ou la représentation des actions qu'un autre agent a_j peut réaliser.

Définition 5.3 (Modèle d'évaluation)
Le modèle d'évaluation (ou Evaluation Process) EP produit les ensembles d'actions désirables et d'actions exécutables à partir des ensembles de désirs et croyances. Il est défini comme :

$$EP = \langle A_{a_i}, \mathcal{A}_{d_{a_i}}, \mathcal{A}_{c_{a_i}}, DE, CE \rangle$$

où A_{a_i} est un ensemble d'actions qu'il s'agit de juger (chacune étant décrite comme une paire de conditions et conséquences portant sur les croyances et les désirs), $\mathcal{A}_{d_{a_i}} \subseteq A_{a_i}$ et $\mathcal{A}_{c_{a_i}} \subseteq A_{a_i}$ sont respectivement l'ensemble des actions désirables et exécutables, DE et l'évaluation de capacités CE sont des fonctions telles que :

$$DE : 2^{D_{a_i}} \times 2^{A_{a_i}} \rightarrow 2^{A_{a_i}}$$

$$CE : 2^{B_{a_i}} \times 2^{A_{a_i}} \rightarrow 2^{A_{a_i}}$$

L'évaluation de désirabilité est la capacité à déduire les actions pertinentes à effectuer au regard des désirs et des connaissances sur les conditions et conséquences des actions. Ainsi, une action α est évaluée comme étant désirable si l'agent désire la réalisation de α ou la réalisation de ses conséquences (et inversement, elle peut être indésirable s'il désire que ces éléments ne se réalisent pas). L'action peut être désirable et indésirable simultanément si sa réalisation ou les conséquences de sa réalisation sont évaluées différemment. Notons ici qu'il est possible d'envisager que les conséquences d'une action α puissent être désirables en raison de connaissances sur une autre action α', désirable, et dont les conditions sont des conséquences de α. Les conditions d'une action α permettent également de savoir si l'action est exécutable dans le contexte courant décrit par l'ensemble des croyances. Par la suite, nous désignons par CK_{a_i} l'union des croyances $\mathcal{B}_{a_i}$ et désirs $\mathcal{D}_{a_i}$. Il s'agit des connaissances contextuelles (ou *Contextual Knowledge*) d'un agent a_i sur l'état du monde.

Maintenant que nous avons défini les modèle de reconnaissance de situation et d'évaluation, nous pouvons aborder les modèle au cœur du modèle de jugement : le modèle moral qui emploie les règles morales et valeurs morales, le modèle éthique qui emploie les principes éthiques.

1.2 Supports de valeurs, règles morales et principes éthiques

Le modèle moral est défini comme suit :

Définition 5.4 (Modèle moral)
Le modèle moral (ou Goodness Process) GP identifie les actions morales à partir des croyances, désirs et connaissances sur les actions d'un agent a_i ainsi que de ses valeurs et règles morales. Il est défini comme :

$$GP = \langle VS_{a_i}, MR_{a_i}, \mathcal{A}_{m_{a_i}}, ME \rangle$$

où VS_{a_i} est la base de connaissances du support de valeurs de l'agent a_i, MR_{a_i} est sa base de connaissances de règles morales, $\mathcal{A}_{m_{a_i}} \subseteq A_{a_i}$ est l'ensemble de ses actions morales[2]*. La fonction d'évaluation morale ME est :*

$$ME : 2^{D_{a_i}} \times 2^{B_{a_i}} \times 2^{A_{a_i}} \times 2^{VS_{a_i}} \times 2^{MR_{a_i}} \rightarrow 2^{A_{a_i}}$$

La base de connaissance du support de valeur associe un ensemble fini de valeurs morales à des combinaisons d'actions et de situations. L'exécution d'une action dans une situation donnée promeut alors une valeur correspondante. Nous pouvons considérer diverses disjonctions pour une même valeur morale : par exemple l'honnêteté peut être définie comme « Ne pas dire quelque chose d'incompatible avec mes croyances » (car c'est mentir sciemment) ou comme « Dire ce que je crois lorsque je crois qu'un autre agent croit le contraire » (pour éviter les mensonges par omission).

Définition 5.5 (Support de valeur)
Un support de valeur est un couple $\langle s, v \rangle \in VS_{a_i}$ où $v \in \mathcal{O}_v$ est une valeur morale et $s = \langle \alpha, w \rangle$ est le support de cette valeur morale avec $\alpha \in A_{a_i}$, $w \subset B_{a_i} \cup D_{a_i}$.

Exemple 5.6
Les supports de la générosité comme « donner à tout agent pauvre » et de l'honnêteté comme « ne pas dire quelque chose d'incompatible avec mes croyances » peuvent être représentés par :

$$\langle \langle \texttt{give}(a), \{\texttt{poor}(a)\} \rangle, \texttt{generosity} \rangle$$

$$\langle \langle \texttt{tell}(a, \phi), \{\phi\} \rangle, \texttt{honesty} \rangle$$

où a représente un agent et $\texttt{poor}(a)$ (respectivement ϕ) est une croyance représentant le contexte d'exécution de l'action $\texttt{give}(a)$ (respectivement $\texttt{tell}(a, \phi)$) supportant la valeur $\texttt{generosity}$ (respectivement $\texttt{honesty}$).

En plus des valeurs morales, nous représentons des règles morales. Une règle morale décrit l'association d'une valuation morale (par exemple parmi un ensemble tel que {moral, amoral, immoral}) à des actions ou valeurs morales dans une situation. Ici « amoral » est un élément de référence de cet ensemble permettant de préciser que la morale est indifférente à une action (ce qui est différent de l'inexistence de connaissances sur la moralité d'une action).

2. Notons que $A_{m_{a_i}} \nsubseteq A_{d_{a_i}} \cup A_{c_{a_i}}$ car une action peut être morale en soi, même si elle n'est pas désirée ou réalisable (ex : sauver le monde).

Définition 5.7 (Règle morale)
Une règle morale est un n-uplet $\langle w, o, m\rangle \in MR_{a_i}$ *où* w *est un état du monde décrit par* $w \subset CK_{a_i}$ *interprété comme une conjonction de croyances et désirs,* $o = \langle \alpha, v\rangle$ *où* $\alpha \in A_{a_i}$ *et* $v \in \mathcal{O}_v$*, et* $m \in \mathcal{O}_m$ *est une valeur morale décrite dans* $\mathcal{O}_m$ *qui qualifie* o *quand* w *est l'état courant.*

Exemple 5.8
Certaines règles morales classiques telles que « tuer un humain est immoral » ou « être honnête avec un menteur est plutôt moral » peuvent être représentées comme :

$$\langle \{\mathit{human}(a)\}, \langle \mathit{kill}(a), \_ \rangle, \mathit{immoral}\rangle$$

$$\langle \{\mathit{liar}(a)\}, \langle \_ , \mathit{honesty}\rangle, \mathit{good}\rangle$$

Une règle peut être plus ou moins spécifique à une situation w ou un objet o. Par exemple « la justice est morale » est plus générale (s'applique à un plus grand nombre de valeurs de w et o) que « juger un meurtrier en prenant compte de sa religion, sa couleur de peau, son origine ethnique ou ses opinions politiques est immoral ». De manière classique, les théories morales peuvent être représentées selon trois approches :

1. une approche **vertueuse** utilise des règles générales s'exprimant sur des valeurs morales : « Il est moral d'être généreux »,
2. une approche **déontologique** est généralement décrite par des règles spécifiques décrivant des devoirs ou des interdits : « Les journalistes doivent refuser toute faveur aux publicitaires, donateurs ou groupes d'intérêt et résister aux pressions internes ou externes qui tenteraient de les influencer » [3],
3. une approche **consequentialiste** utilise des règles générales et spécifiques concernant les états et les conséquences : « Tout médecin doit s'abstenir, même en dehors de l'exercice de sa profession, de tout acte de nature à déconsidérer celle-ci. » [4].

La définition 5.7 laissant la possibilité d'exprimer la moralité d'une action en fonction d'un contexte, de supports de valeurs ou de conséquences, ces trois approches sont compatibles avec le modèle proposé ici. Par la suite, nous ferons référence à ces différentes connaissances (règles morales MR_{a_i}, support de valeurs VS_{a_i} et valeurs $\mathcal{O}_v$) utilisées dans le modèle moral de l'agent a_i sous l'appellation de connaissance du bien, notée GK_{a_i}

À partir de l'ensemble des actions possibles, désirables et morales, nous pouvons introduire le modèle éthique qui a pour but de déterminer les actions justes. Comme mis en lumière par le psychologue Jonathan Haidt, un agent peut utiliser plusieurs *principes éthiques* pour concilier ces ensembles d'actions [Haidt, 2001].

3. Extrait de [of Professional Journalists, 2014], section « Act Independently ».
4. Code de déontologie médicale, article 31.

Définition 5.9 (Modèle éthique)
Un modèle éthique (ou Rightness Process) RP produit les actions justes selon une représentation donnée de l'éthique. Il est défini comme :

$$RP = \langle P_{a_i}, \succ_{e_{a_i}}, \mathcal{A}_{r_{a_i}}, EE, J \rangle$$

où P_{a_i} est la base de connaissances sur les principes éthiques de l'agent a_i, $\succ_{e_{a_i}} \subseteq P_{a_i} \times P_{a_i}$ est un ensemble de relations de préférences représentant un ordre total sur ces principes. Les deux fonctions sont EE (évaluation éthique) et J (jugement) permettent respectivement de construire l'ensemble $\mathcal{A}_{e_{a_i}}$ des actions éthiques, c'est-à-dire conformes aux principes éthiques et l'ensemble $\mathcal{A}_{r_{a_i}} \subseteq A_{a_i}$ des actions justes à partir de $\mathcal{A}_{e_{a_i}}$ et des préférences :

$$EE : 2^{\mathcal{A}_{d_{a_i}}} \times 2^{\mathcal{A}_{c_{a_i}}} \times 2^{\mathcal{A}_{m_{a_i}}} \times 2^{P_{a_i}} \rightarrow 2^{\mathcal{A}_{a_i}} \quad où \quad \mathcal{A}_{e_{a_i}} = A_{a_i} \times P_{a_i} \times \{\bot, \top\}$$

$$J : 2^{\mathcal{A}_{e_{a_i}}} \times 2^{\succ_{e_{a_i}}} \rightarrow 2^{\mathcal{A}_{r_{a_i}}}$$

Nous représentons chaque principe éthique par une fonction – inspirée d'une théorie philosophique – qui estime s'il est juste ou non d'effectuer une action dans une situation donnée au regard de cette théorie. La fonction d'évaluation éthique EE renvoie alors l'évaluation de toutes les actions désirables ($\mathcal{A}_{d_{a_i}}$), réalisables ($\mathcal{A}_{c_{a_i}}$) ou morales ($\mathcal{A}_{m_{a_i}}$) étant donné l'ensemble P_{a_i} des principes éthiques connus.

Définition 5.10 (Principe éthique)
Un principe éthique $p \in P_{a_i}$ est une fonction décrivant la justesse d'une action évaluée en termes de capacités, désirs et moralité dans une situation donnée :

$$p : 2^{A_{a_i}} \times 2^{\mathcal{B}_{a_i}} \times 2^{\mathcal{D}_{a_i}} \times 2^{MR_{a_i}} \times 2^{V_{a_i}} \rightarrow \{\top, \bot\}$$

Exemple 5.11
Par exemple, considérons trois agents dans la situation suivante inspirée de la critique de Benjamin Constant de l'Impératif Catégorique de Kant [Constant et Kant, 2003]. Un agent A est caché chez un agent B pour échapper à un agent C, et C vient demander à B où se trouve A pour le tuer. Les règles morales de B sont mr_1 : « mettre autrui en danger est immoral » et mr_2 : « mentir est immoral ». B sait qu'il sera tué à la place de A s'il refuse de répondre. B désire éviter tout problème avec C. B connait la vérité et doit choisir l'une des trois actions suivantes : dire la vérité à C (satisfaisant ainsi mr_2 et son désir), mentir (satisfaisant mr_1 et son désir) ou refuser de répondre (satisfaisant les deux règles morales mais pas son désir). B connait deux principes éthiques (implémentés en P comme fonctions) : P1 pour lequel une action est juste si elle est possible, motivée par au moins une règle morale ou un désir et P2 pour lequel une action est juste si elle est possible et n'enfreint aucune règle morale. L'évaluation de l'éthique de B renvoie les n-uplets donnés par la table 5.1 où chaque ligne représente une action et chaque colonne représente un principe éthique.

Principe / Action	P1	P2
dire la vérité	$\top$	$\bot$
mentir	$\top$	$\bot$
refuser	$\top$	$\top$

TABLE 5.1 – Évaluation éthique des actions pour le dilemme de Benjamin Constant

Étant donné un ensemble d'actions issues de l'évaluation éthique $\mathcal{E}$, le jugement J est la dernière étape qui sélectionne l'action juste à effectuer, au regard d'un ensemble de préférences éthiques (définissant un ordre total sur les principes éthiques). Pour poursuivre notre exemple, supposons que les préférences éthiques de l'agent B sont $P1 \succ_e P2$ et que J utilise une règle de bris d'égalité basée sur l'ordre lexicographique. Ici le principe préféré, P1 considère que chacune des actions est éthique. Cependant, « refuser de répondre » est l'action juste car elle satisfait également P2 à l'inverse de « mentir » ou « dire la vérité ». Notons que ce jugement pourrait faire apparaître un dilemme entre « dire la vérité » et « refuser de répondre » en l'absence de règle de bris d'égalité (c'est-à-dire la seule prise en considération du principe préféré).

Par la suite, nous ferons référence à ces différentes connaissances (c'est-à-dire principes éthiques P_{a_i} et préférences éthiques $\succ_{e_{a_i}}$) utilisées dans le modèle éthique de l'agent a_i sous l'appellation de connaissance du juste, notée RK_{a_i}.

1.3 Typologie des jugements

Nous avons illustré au travers les exemples précédents que notre modèle peut permettre à un agent a_i de juger de l'action la plus éthique à effectuer au regard de ses propres connaissances CK_{a_i}, GK_{a_i} et RK_{a_i}. Toutefois, ce modèle peut aussi être employé pour juger le comportement d'un autre agent de manière plus ou moins informée en se projetant à la place de l'agent jugé a_j. Dans un processus de jugement éthique EJP tel que défini dans la section précédente, les états mentaux des éléments de CK_{a_i}, GK_{a_i} et RK_{a_i} peuvent être échangés entre agents (même si nous n'abordons pas dans ce mémoire les modalités de ces échanges). Comme discuté au début de la section précédente, l'ontologie $\mathcal{O}$ est considérée comme une connaissance commune, même si nous pouvons envisager dans des travaux futurs la coexistence de plusieurs ontologies.

Nous distinguons quatre catégories de jugement : (1) le *jugement pour la décision* dans lequel l'agent juge de ses propres actions pour décider celle qui doit être réalisée ; (2) le *jugement aveugle* dans lequel l'agent juge a_j n'a d'autre information sur l'agent jugé a_i que son comportement observé ; (3) le *jugement partiellement informé* lorsque le juge a_j dispose d'informations partielles sur les connaissances de l'agent jugé a_i ; (4) le *juge-*

ment parfaitement informé lorsque l'agent juge a_j dispose de la totalité des informations existantes sur l'agent jugé a_i.

Dans tous ces types de jugement, l'agent juge raisonne sur ses propres états mentaux à défaut de disposer de ceux de l'agent jugé. Ce type de jugement peut être comparé chez l'humain à la théorie de l'esprit (la faculté pour un humain à se représenter les états mentaux d'un autre). Ainsi, l'agent juge peut utiliser son propre processus de jugement éthique EJP en substituant autant que possible les états mentaux de l'autre agent aux siens afin de comparer $\mathcal{A}_r$ et $\mathcal{A}_m$ au comportement observé chez l'autre agent. Si l'action effectuée se trouve dans $\mathcal{A}_r$, l'agent juge peut supposer que l'agent jugé agit conformément à son éthique et, respectivement, si elle se trouve dans $\mathcal{A}_m$, elle est conforme à sa morale.

Jugement pour la décision

Un premier usage du jugement consiste à l'intégrer dans le modèle de décision d'un agent autonome. Pour ce faire, l'agent doit être conçu de manière à ce que seules les actions jugées éthiques puissent être décidées par l'agent. Ainsi, l'agent présenterait un comportement qu'il juge éthique à tout moment. Notons toutefois que si ses connaissances ($A_{a_i}, CK_{a_i}, GK_{a_i}$ ou RK_{a_i}) évoluent au cours du temps, ce jugement pourra être contredit par un nouveau jugement.

Jugement éthique aveugle

Un second type de jugement peut être effectué sur un autre agent sans aucune information sur la morale ou l'éthique de l'agent jugé (par exemple dans le cas d'une impossibilité de communiquer). L'agent juge a_j utilise alors sa propre évaluation de la situation ($\mathcal{B}_{a_j}$ and $\mathcal{D}_{a_j}$) [5], sa propre théorie du bien $\langle MR_{a_j}, VS_{a_j}\rangle$ et théorie du juste $\langle P_{a_j}, \succ_{e,a_j}\rangle$ afin d'évaluer le comportement de l'agent jugé a_i. C'est un jugement *a priori* et a_i est jugé comme ayant effectué une action injuste ou immorale si $\alpha_{a_i} \notin \mathcal{A}_{r,a_j}$ ou $\alpha_{a_i} \notin \mathcal{A}_{m,a_j}$.

Jugement éthique partiellement informé

Le troisième type de jugement tient compte d'une information partielle sur l'agent jugé s'il est capable de l'acquérir (par perception ou communication). Trois types de jugement éthique partiel sont considérés, en disposant respectivement (i) de la connaissance contextuelle CK_{a_j}, (ii) de la connaissance du bien GK_{a_j} et A_{a_j} ou (iii) de la connaissance du juste RK_{a_j} de l'agent jugé. Remarquons que, dans le second cas, A_{a_i} est nécessaire car, à l'inverse des principes éthiques, les règles morales peuvent porter directement sur des actions spécifiques.

1. **Jugement considérant la situation.** Si l'agent juge a_j connait les croyances $\mathcal{B}_{a_i}$ et désirs $\mathcal{D}_{a_i}$ de l'agent jugé a_i, a_j peut se placer dans la position de a_i et juger

5. Nous utilisons une notation indicée pour désigner l'agent concerné par l'information.

de l'action α effectuée par a_i en vérifiant si elle fait partie de $\mathcal{A}_{r,a_j}$, en utilisant ses propres théories du bien et du juste. Premièrement, a_j est capable d'évaluer la moralité d'α en générant $\mathcal{A}_{m,a_i}$ à partir de A_{a_i} et qualifier la moralité du comportement de a_i (c'est-à-dire si α est ou non dans $\mathcal{A}_{m,a_i}$). L'agent a_j peut aller plus loin en générant $\mathcal{A}_{r,a_i}$ à partir de $\mathcal{A}_{m,a_i}$ pour vérifier si α est conforme à la théorie du juste (c'est-à-dire fait partie de $\mathcal{A}_{r,a_i}$).

2. **Jugement considérant la théorie du bien.** Si l'agent juge est capable d'obtenir les règles morales et valeurs de l'agent jugé, il est possible d'évaluer l'action dans une situation (partagée ou non), au regard de ces règles. Dans la simple perspective d'une évaluation de la morale de l'agent jugé, l'agent juge peut comparer leurs théories du bien en vérifiant si les valeurs morales et règles morales de l'agent jugé sont consistantes avec sa propre théorie du bien (c'est-à-dire s'il a les mêmes définitions que a_j ou au moins qu'il n'y a pas de contradictions). Dans la perspective d'un jugement moral, l'agent juge peut évaluer la moralité d'une action donnée du point de vue de l'agent jugé. Cette forme de jugement prend tout son intérêt par exemple lorsque les agents sont tenus à des devoirs moraux différents (en raison d'un rôle ou d'une responsabilité particulière par exemple) comme un humain peut juger un médecin sur la conformité de son comportement vis-à-vis du code de déontologie médicale sans être lui-même un membre du corps médical.

3. **Jugement considérant la théorie du juste.** Nous pouvons également considérer le jugement d'un agent juge capable de raisonner sur les principes et préférences éthiques d'un agent jugé en considérant une situation (partagée ou non) et une théorie du bien (partagée ou non) [6]. Cela permet d'évaluer comment l'agent a_i concilie ses désirs et sa morale dans une situation en comparant l'ensemble des actions justes $\mathcal{A}_{r,a_j}$ et $\mathcal{A}_{r,a_i}$ respectivement générées en utilisant P_{a_j}, $\succ_{e,a_j}$ et P_{a_i}, $\succ_{e,a_i}$. Par exemple, si $\mathcal{A}_{r,a_j} = \mathcal{A}_{r,a_i}$ avec une théorie du bien qui n'est pas partagée, cela montre que les deux théories du juste produisent un même jugement dans ce contexte. Ce jugement peut être utile pour un agent afin d'estimer comment un autre agent peut juger de l'éthique d'une action dans une situation avec une morale donnée.

Jugement pleinement informé

Enfin, l'agent juge peut prendre en considération à la fois la morale et l'éthique de l'agent jugé. Ce type de jugement nécessite la totalité des états mentaux internes et connaissances de l'agent jugé. Un jugement pleinement informé est utile pour vérifier la conformité d'un comportement à une éthique publiquement déclarée.

6. Si la situation et la théorie du bien sont partagées, il s'agit d'un jugement pleinement informé.

2 Jugement et confiance dans les autres agents

Cette section décrit le mécanisme d'agrégation d'informations sur la moralité des actions, puis de jugement éthique progressif de comportements représentés sous la forme d'une séquence d'action. En effet, le jugement tel que présenté dans la section précédente permet d'effectuer un *jugement ponctuel* d'une action dans une situation à un instant donné. Ce section traite de l'information obtenue lors de jugements d'actions successives constituant le *comportement* d'un même agent en ajoutant une dimension temporelle afin de définir une notion de confiance en la moralité ou l'éthique d'un autre agent.

Définition 5.12 (Comportement)
Le comportement $b_{a_j,[t_0,t]}$ d'un agent a_j sur l'intervalle temporel $[t_0, t]$ est l'ensemble des actions α_{a_j} exécutées par a_j entre t_0 et t tel que $0 \leqslant t_0 \leqslant t$.

$$b_{a_j,[t_0,t]} = \{\alpha_{a_j} \in A : \exists t' \in [t_0, t] \text{ tel que } \texttt{done}(\alpha_{a_j}, a_j, t')\}$$

où $\texttt{done}(\alpha, a_k, t)$ est une croyance signifiant que l'agent croit que l'action α a été réalisée par a_j à l'instant t.

Par agrégation de jugements ponctuels successifs, l'agent peut construire de manière incrémentale et cumulative une image de la conformité du comportement de l'agent jugé vis-à-vis d'un ensemble de connaissances employé lors des jugements. Une image peut être calculée avec divers types de jugements agrégés de différentes manières. La période temporelle sur laquelle le comportement de l'agent est jugé pour construire cette image est l'un des paramètres du jugement.

2.1 Images de la moralité et de l'éthique d'un agent

Un agent peut disposer de plusieurs images d'un même comportement construites par jugements progressifs portant sur divers éléments ou ensembles d'éléments du modèle de jugement, de manière aveugle, partiellement ou totalement informés.

Image de la moralité des actions d'un agent

Afin de construire une image de la moralité d'un autre agent, l'agent juge utilise le modèle moral pour évaluer la conformité d'un comportement observé à un ensemble de règles morales ms et classer ainsi chaque action α d'un comportement d'agent $b_{a_j,[t_0,t]}$ au regard de sa conformité à un ensemble de connaissances. La définition d'un tel ensemble de règles morales permet au concepteur de définir une sous-partie de la théorie du bien au regard de laquelle il est pertinent d'évaluer la conformité du comportement d'un autre. Nous présentons dans un premier temps la formalisation de cette construction, puis l'illustrons sur un exemple page 115.

Définition 5.13 (Conformité morale d'une action)
Une action α est dite moralement conforme *au regard des connaissances du contexte (CK_{a_i}) et d'une règle des connaissances du bien (GK_{a_i}) d'un agent a_i à un instant t' – correspondant à la croyance* $\texttt{moral\_conformity}(\alpha, mr, mt, t')$ *– si, et seulement si, la valuation morale associée à α par la fonction d'évaluation morale au regard d'une règle morale $mr \in MR$ est supérieure à un seuil moral $mt \in MV$.*

Remarquons que le prédicat `moral_conformity` ne permet que d'évaluer une seule action au regard d'une seule règle. Pourtant, afin de permettre à un agent de se construire une image de la moralité du comportement d'un autre, il est nécessaire d'évaluer la conformité de ce comportement en évaluant la conformité morale des actions successives qui le composent. Cependant, définir la conformité morale d'un comportement à un ensemble de règles comme la conjonction de la conformité morale de toute action de ce comportement au regard de toute règle morale de cet ensemble serait problématique puisqu'une seule action moralement non conforme suffirait à condamner un comportement, peu importerait le nombre de actions moralement conformes observées. En effet, comme des contradictions peuvent être présentes dans la morale, toute action effectuée dans le cadre d'un dilemme moral se voyant simultanément affectée de valuations supérieures et inférieures au seuil mt rendrait le comportement de l'agent moralement non conforme à l'ensemble de règles.

Afin de permettre à l'agent de raisonner sur la proportion d'actions évaluées, la conformité morale est utilisée pour calculer l'ensemble MC^+ des actions moralement conformes au regard de ms et l'ensemble MC^- des actions moralement non conformes au regard de ms du comportement observé $b_{a_j,[t_0,t]}$ de l'agent jugé a_j :

$$\begin{aligned} MC^+_{b_{a_j,[t_0,t]},ms,mt} = \{&\alpha \in b_{a_j,[t_0,t]} \wedge t' \in [t_0,t] \text{ tel que } \texttt{done}(\alpha, a_j, t') \\ &\wedge \texttt{moral\_conformity}(\alpha, mr, mt, t') \wedge mr \in ms\} \\ MC^-_{b_{a_j,[t_0,t]},ms,mt} = \{&\alpha \in b_{a_j,[t_0,t]} \wedge t' \in [t_0,t] \text{ tel que } \texttt{done}(\alpha, a_j, t') \\ &\wedge \neg\texttt{moral\_conformity}(\alpha, mr, mt, t') \wedge mr \in ms\} \end{aligned}$$

L'ensemble des actions moralement évaluées de l'agent a_j au regard de ms et du seuil mt entre l'instant t_0 et t est noté $MC_{b_{a_j,[t_0,t]},ms,mt}$:

$$MC_{b_{a_j,[t_0,t]},ms,mt} = MC^+_{b_{a_j,[t_0,t]},ms,mt} \cup MC^-_{b_{a_j,[t_0,t]},ms,mt}$$

L'agent juge a ensuite besoin d'une fonction pour agréger les évaluations morales ponctuelles de chaque action du comportement. Lors de cette agrégation, une fonction $weight()$ prend en paramètre une action et donne un nombre réel permettant d'affecter une pondération à certaines actions dans la construction de l'image. Ainsi, l'agent peut, par exemple, accorder plus d'importance aux actions jugées de manière pleinement informée, ou plus récentes. La fonction peut aussi attribuer le même poids à toutes les actions.

Définition 5.14 (Fonction d'agrégation morale)
Une fonction d'agrégation morale $MA : 2^{\mathcal{A}} \rightarrow [0,1]$ *attribue une valeur quantitative représentant le ratio pondéré des actions d'un comportement évaluées moralement conformes par rapport à l'ensemble des actions de ce comportement. Elle est définie telle que :*

$$MA(MC_{b_{a_j,[t_0,t]},ms,mt}) = \frac{\sum_{\alpha \in MC^{+}_{b_{a_j,[t_0,t]},ms,mt}} weight(\alpha)}{\sum_{\alpha \in MC_{b_{a_j,[t_0,t]},ms,mt}} weight(\alpha)}$$

L'agrégation des évaluations de conformité morale permet de construire un ensemble de croyances qualifiant la conformité du comportement d'un agent à un ensemble de règles morales. Le produit de cette agrégation est une image du caractère respectueux du comportement de l'agent observé vis-à-vis d'un ensemble de règles. Cette image est définie de la manière suivante :

Définition 5.15 (Image morale)
Une image morale *d'un agent a_j est une croyance construite par agrégation d'évaluations morales du comportement $b_{a_j,[t_0,t]}$ de cet agent au regard d'un ensemble de règles morales ms, d'une connaissance du contexte CK et d'une connaissance du bien GK. Cette image associe à ce comportement une valuation de conformité $cv \in CV$, où CV est un ensemble ordonné de* valuations de conformité *défini dans l'ontologie. L'image morale qu'un agent a_i se construit par évaluation de la conformité morale du comportement d'un agent a_j au regard de ms et mt entre t_0 et t est notée* $\texttt{moral\_image}(a_i, a_j, ms, mt, cv, t_0, t)$

Cette image qualifiant la conformité du comportement au regard d'un ensemble de règles permettra dans les sections suivantes de tenir compte de cette information dans les interactions entre les agents. Remarquons qu'un agent peut maintenir simultanément plusieurs calculs d'images avec des ensembles moraux, des fonctions de calcul de l'image morale, des seuils moraux ou des périodes de temps différents. Cela permet de caractériser le fait.

Exemple 5.16
Soit un agent a_i qui considère une ensemble moral $ms_1 = \{mr_1, mr_2\}$ où mr_i est une règle morale. Du point de vue sémantique, un comportement dont les actions sont en majorité évaluées positivement par des règles de ms_1 est un comportement promouvant une même valeur. Supposons que a_i discrétise les valuations de conformité morale sur l'ensemble $CV = \{\texttt{improper}, \texttt{neutral}, \texttt{congruent}\}$ associé aux intervalles $\{[0,0,4[, [0,4,0,6[, [0,6,1]\}$. Par simplicité, fixons le seuil moral $mt \in MV$ à $\texttt{neutral}$ *et la fonction $weight()$ à un poids identique de 1 pour toute les actions. Supposons maintenant que a_i observe un agent a_j qui réalise successivement aux instants t et t' les actions α_1 et α_2 telle que α_1 n'est pas morale selon la règle mr_1 et α_2 est morale selon la règle mr_2. Après chaque observation, a_1 met à jour l'image morale de a_j au regard de ms_1 :*

— *Après α_1, $\neg\texttt{moral\_conformity}(\alpha_1, mr_1, \texttt{neutral}, t) \wedge mr_1 \in ms_1$ est vérifié, ce qui permet d'ajouter l'action α_1 à l'ensemble $MC^{-}_{b_{a_j,[t_0,t]},ms_1,\texttt{neutral}}$. Réévaluant l'image*

de a_j, l'agent a_i calcule l'agrégation morale $MA(MC_{b_{a_j,[t_0,t]},ms_1,\texttt{neutral}}) = 0$. L'image morale produite est donc $\texttt{moral\_image}(a_i, a_j, ms_1, \texttt{neutral}, \texttt{improper}, t_0, t)$ indiquant que le comportement observé n'est pas conforme à l'ensemble moral ms_1.

— *Après α_2, $\texttt{moral\_conformity}(\alpha_2, mr_2, \texttt{neutral}, t') \wedge mr_2 \in ms_1$ est vérifié, ce qui permet d'ajouter l'action α_2 à l'ensemble $MC^+_{b_{a_j,[t_0,t']},ms_1,\texttt{neutral}}$. Réévaluant à nouveau l'image de a_j, a_i calcule l'agrégation morale $MA(MC_{b_{a_j,[t_0,t']},ms_1,\texttt{neutral}}) = 0,5$. L'image morale produite est donc $\texttt{moral\_image}(a_i, a_j, ms_1, \texttt{neutral}, \texttt{neutral}, t_0, t')$ illustrant que les évaluations successives des deux actions a conduit a_i à considérer le comportement de a_j comme non-conforme puis comme neutre du point de vue de l'ensemble moral ms_1.*

Image de l'éthique des actions d'un agent

Le jugement d'actions d'un comportement permet d'évaluer leur conformité éthique et classer ainsi chaque action α d'un comportement d'agent $b_{a_j,[t_0,t]}$ au regard du résultat de son jugement en employant un ensemble de connaissances. Comme précédemment, nous présentons dans un premier temps la formalisation et illustrons ensuite sur un exemple page 117.

Définition 5.17 (Conformité éthique)
Une action α est dite éthiquement conforme *au regard des connaissances du contexte (CK_{a_j}), connaissances du bien (GK_{a_j}) et connaissances du juste (RK_{a_j}) d'un agent a_j à un instant t' – correspondant à la croyance $\texttt{ethical\_conformity}(\alpha, t')$ – si, et seulement si, l'action α appartient à l'ensemble des actions justes calculé par la fonction de jugement éthique.*

De manière analogue à la construction de la conformité morale d'un comportement, la conformité éthique est utilisée pour calculer l'ensemble EC^+ des actions éthiquement conformes et l'ensemble EC^- des actions éthiquement non conformes du comportement observé $b_{a_j,[t_0,t]}$ de l'agent jugé a_j entre t_0 et t :

$$EC^+_{b_{a_j,[t_0,t]}} = \{\alpha \in b_{a_j,[t_0,t]} \wedge t' \in [t_0, t] \text{ tel que } \texttt{done}(\alpha, a_j, t') \wedge \texttt{ethical\_conformity}(\alpha, t')\}$$
$$EC^-_{b_{a_j,[t_0,t]}} = \{\alpha \in b_{a_j,[t_0,t]} \wedge t' \in [t_0, t] \text{ tel que } \texttt{done}(\alpha, a_j, t') \wedge \neg\texttt{ethical\_conformity}(\alpha, t')\}$$

L'ensemble des actions jugées de l'agent a_j entre l'instant t_0 et t est noté $EC_{a_j,[t_0,t]}$:

$$EC_{b_{a_j,[t_0,t]}} = EC^+_{b_{a_j,[t_0,t]}} \cup EC^-_{b_{a_j,[t_0,t]}}$$

De manière analogue à la fonction d'agrégation morale, l'agent juge a besoin d'une fonction d'agrégation éthique :

Définition 5.18 (Fonction d'agrégation éthique)
Une fonction d'agrégation éthique $EA : 2^{\mathcal{A}} \rightarrow [0,1]$ *attribue une valeur quantitative représentant le ratio pondéré des actions d'un comportement jugées éthiques par rapport à l'ensemble des actions de ce comportement. Elle est définie telle que :*

$$EA(EC_{b_{a_j,[t_0,t]}}) = \frac{\sum_{\alpha \in EC^{+}_{b_{a_j,[t_0,t]}}} weight(\alpha)}{\sum_{\alpha \in EC_{b_{a_j,[t_0,t]}}} weight(\alpha)}$$

L'agrégation des jugements éthiques permet de construire des croyances qualifiant la conformité du comportement d'un agent à un ensemble de connaissances CK, GK et RK. Le produit de cette agrégation est une image de l'éthique du comportement de l'autre. Cette image est définie de la manière suivante :

Définition 5.19 (Image éthique)
Une image éthique *d'un agent a_j est un jugement agrégé du comportement $b_{a_j,[t_0,t]}$ de cet agent au regard d'une éthique, d'une connaissance du contexte CK, d'une connaissance du bien GK et d'une connaissance du juste RK. Cette image attribue une valuation de conformité $cv \in CV$, où CV est un ensemble ordonné de* valuations de conformité. *L'image éthique qu'un agent a_i se construit par observation du comportement d'un agent a_j entre t_0 et t est notée* $\texttt{ethical\_image}(a_i, a_j, cv, t_0, t)$

Cette image qualifiant le caractère éthique du comportement permettra dans les sections suivantes de tenir compte de cette information dans les interactions entre les agents. Comme pour les images morales, un agent peut maintenir simultanément plusieurs calculs d'images éthiques avec des fonctions de calcul de l'image éthique, des seuils ou des périodes de temps différents. De plus, remarquons qu'il est possible de disposer d'images d'un agent telles qu'il serait jugé conforme par toutes les images morales en n'étant non conforme pour son image éthique. Cela peut se produire par exemple pour des raisons de divergences dans la résolution de dilemmes moraux pour lesquels toute alternative améliore au moins une image morale mais où les divergences de théories du juste de l'agent juge et de l'agent jugé les amènent à ne pas considérer la même action comme juste. À l'inverse, dans des situations dans lesquelles aucune action morale n'est possible, le comportement d'un agent peut amener un agent juge à dégrader les images morales du jugé en améliorant l'image éthique de ce dernier et aboutir à une situation dans laquelle aucune image morale du comportement n'est conforme, mais où l'image éthique le serait.

Exemple 5.20
Reprenons l'exemple 5.16. Supposons que a_i suive un unique principe éthique : une action est juste si elle est possible, et est soutenue par au moins une règle morale ou n'enfreint aucune règle morale. Supposons que a_i discrétise les valuations de conformité éthique sur l'ensemble $CV = \{\texttt{improper}, \texttt{neutral}, \texttt{congruent}\}$ associé aux intervalles $\{[0,0,4[, [0,4,0,6[, [0,6,1]\}$. Après chaque action, a_1 met à jour l'image éthique de a_j :

— *Après* α_1, $\neg$`ethical_conformity`(α_1, t) *est vérifié, ce qui permet d'ajouter l'action* α_1 *à l'ensemble* $EC^-_{b_{a_j,[t_0,t]}}$. *L'agent* a_i *calcule l'agrégation éthique* $EA(EC_{b_{a_j,[t_0,t]}}) = 0$. *L'image éthique est* `ethical_image`$(a_i, a_j,$ `improper`$, t_0, t)$ *indiquant que le comportement observé n'est pas conforme à l'éthique de* a_i.

— *Apès* α_2, `ethical_conformity`(α_2, t') *est vérifié, ce qui permet d'ajouter l'action* α_2 *à l'ensemble* $EC^+_{b_{a_j,[t_0,t']}}$. *L'agent* a_i *calcule l'agrégation éthique* $EA(EC_{b_{a_j,[t_0,t']}}) = 0,5$. *L'image éthique est* `ethical_image`$(a_i, a_j,$ `neutral`$, t_0, t')$ *illustrant que les évaluations successives des deux actions a conduit* a_i *à considérer le comportement de* a_j *comme non-conforme puis comme neutre du point de vue de l'éthique de* a_i.

2.2 Une confiance dans l'éthique des autres agents

La construction d'images du comportement des autres agents vis-à-vis d'éléments de la morale et de l'éthique permet à l'agent juge d'accorder ou non sa confiance à un autre agent. La confiance construite par ce processus peut ensuite être employée pour décrire une manière éthique d'interagir et coopérer avec les autres agents du système. La figure 5.2 représente le mécanisme de construction de la confiance dans sa globalité : le modèle de jugement décrit au chapitre précédent est employé avec un ensemble de connaissances du contexte, de connaissances du bien et de connaissances du juste afin de générer les ensembles d'actions évaluées par l'éthique EC et par la morale MC. Ces ensembles permettent à leur tour la construction d'un ensemble d'images du comportement de l'agent jugé. Nous décrivons à présent l'emploi d'une action permettant à l'agent de construire l'ensemble T des croyances en la confiance qu'il accorde aux autres agents.

Grâce aux images morales et éthiques, un agent peut décider d'accorder sa confiance à un autre ou non. La confiance peut être absolue (une confiance dans la conformité à une éthique du comportement de l'autre) ou relative à un ensemble de règles morales (confiance dans la prudence de l'autre, sa responsabilité, son obéissance à un ensemble de règles de conduite, etc.). Nous définissons deux actions épistémiques internes permettant d'évaluer la possibilité d'établir ces deux types de confiance.

Définition 5.21 (Fonction de confiance morale)
La fonction de confiance morale MTB_{a_i} *permettant d'évaluer si l'agent juge peut accorder sa confiance à l'agent* a_j *pour la conformité de son comportement vis-à-vis de l'ensemble moral* ms *est définie comme :*

$$MTB_{a_i} : Ag \times 2^{ms_{a_i}} \times MV_{a_i} \rightarrow \{\top, \bot\}$$

Définition 5.22 (Fonction de confiance éthique)
La fonction de confiance éthique ETB_{a_i} *permettant d'évaluer si l'agent juge peut accorder sa confiance à l'agent* a_j *pour la conformité de son comportement vis-à-vis du jugement éthique est définie comme :*

$$ETB_{a_i} : Ag \rightarrow \{\top, \bot\}$$

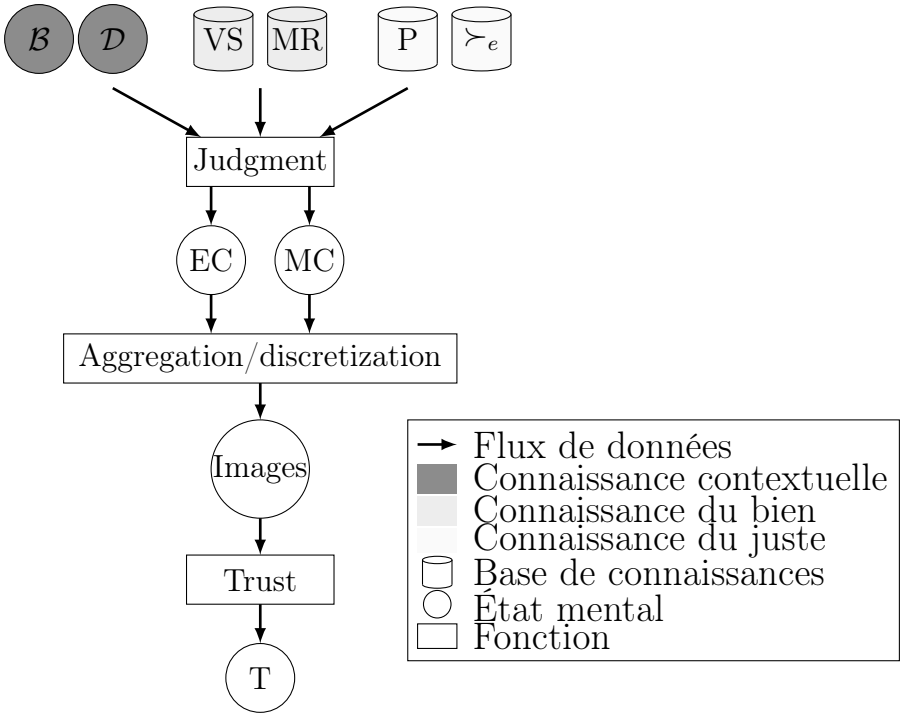

FIGURE 5.2 – Processus de construction de la confiance en l'éthique des agents

Ici, ces fonctions de confiance sont abstraites et doivent être instanciées. Lorsqu'un agent a_i évalue la conformité du comportement d'un autre agent a_j au regard de CK_{a_i}, GK_{a_i} et RK_{a_i} (c'est-à-dire l'image éthique), la fonction de confiance éthique produit une croyance $\texttt{ethical\_trust}(a_j, a_i)$. De même, lorsque l'agent a_i évalue la conformité du comportement de a_j au regard de ms (c'est-à-dire vérifie que la conformité morale de l'image de son comportement par rapport à ms est au moins égale à mt), la fonction de confiance morale produit une croyance $\texttt{moral\_trust}(a_j, a_i, ms, mt)$. L'ensemble de ces croyances de l'agent représentant sa confiance dans les autres agents est noté T.

2.3 Éthique de la confiance

Faire confiance étant une action épistémique, il est possible de décrire la moralité de cette action en fonction du contexte et de juger s'il est juste d'accorder sa confiance à un autre agent. De même, la description de supports de valeurs pour l'action de « faire confiance » permet de définir de nouvelles valeurs décrivant la manière d'accorder sa confiance aux autres agents. Par exemple, l'intransigeance peut être une valeur supportée par l'action d'accorder sa confiance uniquement aux agents dont l'image est au dessus d'un seuil moral ou éthique relativement élevé. À l'inverse, l'indulgence définie comme supportée par l'action consistant à accorder sa confiance à des agents dès lors qu'il existe une image dépassant un seuil moral ou éthique relativement bas. La description de règles

morales peut ensuite décrire la moralité de la confiance. Par exemple, il est possible de définir une règle morale telle que « Il est moral d'être indulgent durant les cinq premières minutes de l'observation de leur comportement » ou bien « Il est immoral de ne pas être intransigeant lorsque la situation est critique ». La moralité de la construction de la confiance en fonction des paramètres de la construction de cette confiance peut ainsi être dépendante de la connaissance que l'agent a du contexte.

Les croyances sur l'image et la confiance peuvent enfin être des éléments de contexte permettant d'exprimer la moralité ou l'éthique d'une action. Autrement dit, la moralité d'une action à l'égard d'un agent peut être conditionnée à la confiance ou l'image que l'agent juge a de l'autre. Premièrement, la confiance éthique et morale peut enrichir la description des règles et valeurs morales. Par exemple, la valeur de *responsabilité* pourrait être supportée lorsque les actions de délégation ne sont confiées qu'à des agents de confiance. Ici, la responsabilité est définie comme la capacité à déléguer des actions sensibles uniquement à des agents appropriés. Deuxièmement, des croyances spécifiques de confiance morale peuvent être employées comme des éléments de règle morale. Par exemple, étant donné une valeur d'honnêteté et ses supports de valeur, un agent peut être doté d'une règle exprimant "Il est immoral de ne pas agir honnêtement à l'encontre de tout agent honnête". Ici, "tout agent honnête" peut être modélisé par l'existence d'une croyance `moral_trust` associant à un agent une confiance morale dans la conformité de son comportement à l'ensemble R des règles définissant la moralité d'un comportement honnête.

Enfin, puisque évaluer et juger les autres constituent des actions, il est également possible d'exprimer et évaluer leur caractère moral ou éthique. Ainsi, la valeur morale de *tolérance* peut être supportée par la construction d'une image des autres avec un seuil peu élevé tant que les ensembles $EC_{a_j,[t_0,t]}$ ou $MC_{a_j,[t_0,t]}$ ne sont pas assez significatifs. Le choix du seuil, des pondérations et la conversion de l'agrégation en niveau de conformité peuvent également permettre de représenter diverses formes de confiance. Une valeur telle que l'*indulgence* peut être supportée par le fait d'accorder toujours une pondération plus faible aux actions les moins récentes. Il est ainsi possible de décrire une morale de la confiance par l'emploi de règles comme "Il est immoral de construire la confiance sans tolérance ni indulgence" [Horsburgh, 1960].

3 Éthique et formation de coalitions

Le modèle précédent porte essentiellement sur l'éthique du comportement individuel de l'agent dans l'accomplissement de ses objectifs. Toutefois, de nombreux domaines applicatifs mettent en présence plusieurs agents qui doivent interagir, décider conjointement et coopérer. Dans ce contexte, un agent doit non seulement tenir compte de critères éthiques au regard de ses objectifs mais aussi sur la manière dont il coopère. Cela peut passer par une éthique de la confiance mais aussi par une éthique de la construction de collectif.

C'est pourquoi, nous nous intéressons dans cette section à la modélisation d'une éthique des vertus – respectant une valeur cardinale – dans le cadre de la formation de coalitions d'agents. Plus précisement, c'est au processus lui-même de formation de ces coalitions au regard de valeurs que les agents désirent respecter que nous nous sommes intéressés. Dans cette section, nous proposons de nous fonder sur des jeux de coalitions hédoniques. Afin de représenter les valeurs cardinales des agents, nous enrichissons ces jeux en proposant des *jeux de déviations* où chaque agent décide de changer de groupe au regard de règles de comportements – appelées *concept de déviations* – qui lui sont propres. Une solution fait consensus lorsqu'aucun agent ne désire changer de coalition. Nous montrons ensuite comment ces règles de déviations peuvent être composées pour représenter une pluralité de valeurs cardinales, en particulier des valeurs de liberté, altruisme et hédonisme, amenant ainsi les agents à suivre une éthique de la vertu dans leur processus de formation de coalitions.

3.1 Des jeux de déviations

Nous avons vu au chapitre 2 section 2 que le problème associé aux jeux hédoniques est de calculer une partition Π de l'ensemble N qui satisfait aux mieux les préférences de chaque agent au regard d'un concept de solution. Chacun de ces concepts de solution représente un comportement spécifique que doivent suivre les agents dans le processus de formation de coalitions. À titre d'exemple, la stabilité au sens de Nash représente des partitions où aucun agent ne désire individuellement rejoindre une autre coalition déjà présente dans cette partition. De plus, tous les concepts de solution canoniques reposent sur une hypothèse forte : tous les agents présentent le même comportement et cherchent à satisfaire le même concept de solution.

Le modèle que nous proposons ici intègre alors ces deux aspects, une généralisation du modèle de [Sung et Dimitrov, 2007] [7] pour représenter des comportements individuels sous forme de déviations et, à la manière de [Vallée et Bonnet, 2017] [8] la prise en compte d'une hétérogénéité des comportements, afin de modéliser une éthique des vertus dans le cadre de la formation de coalitions. Contrairement aux jeux hédoniques classiquement considérés dans la littérature, nous proposons ici un nouveau modèle où la notion de stabilité est définie par l'absence de déviation au regard de conditions propres à chaque agent.

7. [Sung et Dimitrov, 2007] ont proposé de représenter explicitement les comportements en redéfinissant les concepts de solution à partir d'une conjonction de cinq ensembles de déviations, chacun représentant une propriété sur le fait qu'une déviation soit autorisée ou non. Une partition est alors considérée comme stable lorsque aucun agent ne désire dévier.

8. Nous avons proposé dans cet article un modèle de jeux hédonique où chaque agent dispose de son propre concept de solution et avons défini une notion de stabilité faisant consensus entre les agents. Outre des résultats de complexité, nous avons montré que ce type de jeux est une généralisation des jeux hédoniques classiquement étudiés dans la littérature.

Définition 5.23 (Déviation)
Soit $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ *un jeu hédonique et* $\Pi \in \mathcal{P}_N$ *une partition. Une* déviation *est une coalition* $D \subseteq N, D \notin \Pi, D \neq \emptyset$ *telle que l'ensemble des agents de* D *quittent leurs coalitions courantes dans* Π *pour former la nouvelle coalition* D*.*

Nous distinguons deux types de déviations : les déviations *individuelles* et les déviations *collectives*. Une déviation individuelle est une déviation qui nécessite qu'un seul agent a_i quitte sa coalition courante pour rejoindre les autres membres de D, c'est-à-dire $D \setminus \{a_i\} \in \Pi \cup \{\emptyset\}$. À l'inverse, les déviations collectives nécessitent qu'aux moins deux agents distincts quittent leurs coalitions courantes. Dans la suite, nous désignons par $[D \to \Pi]$ l'application de la déviation D sur la partition Π.

Définition 5.24 (Application d'une déviation)
La partition Π' *résultant de* $[D \to \Pi]$ *est telle que :*

- $\forall a_i \in D, C_i(\Pi') = D$
- $\forall a_j \in N : \exists a_i \in D, C_j(\Pi) = C_i(\Pi), C_j(\Pi') = C_j(\Pi) \setminus D$
- $\forall a_k \in N : \nexists a_j \in D, C_j(\Pi) = C_i(\Pi), C_j(\Pi') = C_j(\Pi)$

Étant donnée une partition Π, nous désignons par $AllD_i(\Pi) = \{D \subseteq N, D \notin \Pi : a_i \in D\}$ l'ensemble des déviations qui impliquent l'agent a_i. Plaçons nous maintenant du point de vue d'un agent $a_i \in N$ et considérons une partition $\Pi \in \mathcal{P}_N$. Nous modélisons les déviations que l'agent a_i désirerait voir se réaliser au regard de ses préférences et d'autres critères qui lui sont propres à l'aide de *conditions* (au sens large) qui doivent être satisfaites.

Définition 5.25 (Condition de déviation)
Soit $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ *un jeu hédonique,* $a_i \in N$ *un agent,* $\Pi \in \mathcal{P}_N$ *une partition et* $D \in AllD_i(\Pi)$ *une déviation. Une* conditition de déviation Δ_X *désigne une propriété que doit satisfaire la déviation* D *au regard de l'agent* a_i*, de la partition courante* Π *et du profil de préférence pour que* D *soit désirable pour l'agent* a_i*.*

Dans la suite $\Delta_X(a_i, D, \Pi, HG)$ désigne la fonction booléenne vérifiant si une déviation D satisfait la condition Δ_X du point de vue de l'agent a_i étant donnés la partition Π et le jeu HG. Afin d'illustrer notre propos, nous nous limitons aux conditions ci-dessous. Leur choix est dicté par leur sémantique et par leurs liens avec les concepts de solution classiquement considérés dans la littérature. Nous montrons plus spécifiquement ce lien dans la Section 3.1.

Condition de Rationnalité : $\Delta_R := D \succ_i C_i(\Pi)$ – la déviation D est *rationelle* du point de vue l'agent a_i s'il préfére (strictement) la déviation à sa coalition courante.

Condition d'Acceptation : $\Delta_A := \forall a_j \in D \setminus \{a_i\}, D \succ_j C_j(\Pi)$ – la déviation D est *acceptable* si tous les membres de D préfèrent (strictement) la déviation à leur coalition courante.

Condition de Défection : $\Delta_D := \forall a_k \in N \setminus D : \exists a_j \in D, C_k(\Pi) = C_j(\Pi), C_k(\Pi) \setminus D \succ_k C_k(\Pi)$ – la déviation D est une *défection* si le départ des agents de D est préférable du point de vue des autres membres de leurs coalitions initiales.

Condition d'Optimalité : $\Delta_D := \nexists C \subseteq N : C \succ_i D$ – la déviation D est *optimale* du point de vue de l'agent a_i si elle fait partie de ses coalitions préférées.

Condition de Pareto : $\Delta_{PO} := \exists \Pi' \in \mathcal{P}_N, D \in \Pi' : \forall a_j \in N, C_j(\Pi') \succ_j C_j(\Pi)$ – la déviation D est *Pareto-compatible* s'il existe une partition Π' contenant D, où toutes les coalitions de Π' sont strictement préférées à celles de Π par tous les agents.

Condition d'Individualité : $\Delta_I := D \setminus \{a_i\} \in \Pi \cup \{\emptyset\}$ – la déviation D est *individuelle* si l'agent a_i est le seul agent qui change de coalition lors de la déviation (impliquant que les autres membre de D forment déjà une coalition).

Condition de Collectivité : $\Delta_C := D \setminus \{a_i\} \notin \Pi \cup \{\emptyset\}$ – la déviation D est *collective* si plusieurs agents (dont l'agent a_i) n'appartenaient pas à D avant de la rejoindre.

Remarquons que nous avons ici deux familles de conditions. D'un coté, les conditions Δ_R, Δ_A, Δ_D, Δ_O et Δ_{PO} portent sur la satisfaction des préférences des agents, tandis que les conditions Δ_I et Δ_C portent sur l'identité des agent déviants. Notons par ailleurs que nous n'avons présenté ici que des versions *fortes* des conditions sur les préférences dans le sens où les préférences considérées sont strictes. Nous notons par Δ_X^- les équivalents *affaiblis* usant de préférences non strictes. Par exemple, une déviation D qui satisfait Δ_A^- signifie que les agents de D autres que a_i peuvent également être indifférents au changement de coalition de a_i :

$$\Delta_A^- := \forall a_j \in D \setminus \{a_i\}, D \succeq_j C_j(\Pi)$$

La condition de Pareto diffère des autres conditions de déviation. En effet, cette condition ne compare pas uniquement les coalitions de la partition Π avec les coalitions appartenant à la partition Π' résultant de la déviation $[D \to \Pi]$. Elle compare les coalitions de Π avec toutes les coalitions appartenant à toutes les partitions contenant D. Cette spécificité nous permet de considérer non plus uniquement la déviation D, mais une succession de déviations.

Exemple 5.26
Considérons la partition $\Pi = \{\{a_1, a_3\}, \{a_2, a_4\}\}$ *dans le jeu :*

$$\begin{aligned}
N =& \{a_1, a_2, a_3, a_4\} \\
\succeq_1 =& \{a_1, a_2\} \succ_1 \{a_1, a_3\} \succ_1 \{a_1\} \\
\succeq_2 =& \{a_1, a_2\} \succ_2 \{a_2, a_4\} \succ_2 \{a_2\} \\
\succeq_3 =& \{a_3, a_4\} \succ_3 \{a_1, a_3\} \succ_3 \{a_3\} \\
\succeq_4 =& \{a_3, a_4\} \succ_4 \{a_2, a_4\} \succ_4 \{a_4\}
\end{aligned}$$

Du point de vue de a_1, $\forall D \in AllD_1(\Pi)$, *il existe au moins un agent* $a_j \in N$ *tel que, pour* Π' *résultant de* $[D \to \Pi]$, $C_j(\Pi) \succ_j C_j(\Pi')$. *Le même raisonnement se tient pour les autres agents. Ainsi, quel que soit l'agent ou le groupe d'agents qui effectue une déviation, elle se fait au détriment d'au moins un agent. Par exemple, en considérant la déviation* $D = \{a_1, a_2\}$, *nous avons* $\Pi' = \{\{a_1, a_2\}, \{a_3\}, \{a_4\}\}$ *où* $C_3(\Pi) \succ_3 C_3(\Pi')$. *Cependant, en effectuant cette déviation qui est désavantageuse pour* a_3 *et pour* a_4, *ces derniers peuvent désormais eux-même envisager la déviation* $D_2 = \{a_3, a_4\}$ *avec* $\Pi'' = \{\{a_1, a_2\}, \{a_3\, a_4\}\}$ *qui, elle, satisfait* $\forall a_i \in N, C_i(\Pi'') \succ_i C_i(\Pi)$.

Les conditions de déviation permettent à un agent de définir les règles individuelles permettant de caractériser les déviations qu'il désire réaliser. Trivialement, un agent a_i peut vouloir satisfaire simultanément plusieurs conditions, ou encore qu'au moins l'une soit satisfaite. Par exemple, un agent peut exprimer avec la proposition $\Delta_R \wedge \Delta_A$ le fait de désirer une déviation D si et seulement si D est préférable à la fois pour lui-même mais aussi pour tous les autres agents de D. Ainsi, une agrégation de conditions de déviation est appelée le *concept de déviation* de l'agent a_i.

Définition 5.27 (Concept de déviation)
Soit $a_i \in N$. *Le* concept de déviation $\mathbb{D}_i$ *de l'agent* a_i *est une formule propositionnelle portant sur un ensemble* $\{\Delta_1, \ldots, \Delta_k\}$ *de conditions de déviation. Toute déviation* $D \in AllD_i(\Pi)$ *qui satisfait* $\mathbb{D}_i$ *(noté* $D \vDash \mathbb{D}_i$*) est considérée comme désirable pour l'agent* a_i.

Dans la suite, étant donnés l'agent $a_i \in N$, la partition $\Pi \in \mathcal{P}_N$ et le jeu HG, nous désignons par $\mathbb{D}_i(\Pi, HG)$ l'ensemble des déviations désirables pour l'agent a_i :

$$\mathbb{D}_i(\Pi, HG) = \{D \in AllD_i(\Pi) | D \vDash \mathbb{D}_i\}$$

Exemple 5.28
Considérons un agent a_1 *au concept de déviation* $\mathbb{D}_1 = \Delta_R \wedge \Delta_I$, *signifiant qu'il recherche les déviations individuelles strictement préférées à sa coalition courante. Considérons un autre agent* a_2 *pour qui est désirable toute déviation (individuelle ou collective) telle qu'elle soit strictement préférée par lui-même, par les autres agents déviants et par les agents impactés par la déviation. Ce concept de déviation peut être formalisé comme suit* $\mathbb{D}_2 = (\Delta_I \vee \Delta_C) \wedge \Delta_R \wedge \Delta_A^- \wedge \Delta_D^-$ *qui peut être réduit à* $\mathbb{D}_2$ *à :* $\Delta_R \wedge \Delta_A^- \wedge \Delta_D^-$ *car* $(\Delta_I \vee \Delta_C)$ *est une tautologie.*

Comme nous le montre l'Exemple 5.28, plusieurs agents peuvent suivre des concepts de déviation différents, représentant des agents hétérogènes dans leurs processus de décision vis-à-vis des déviations. Nous pouvons donc définir un nouveau modèle de jeu hédonique – les *jeux hédoniques de déviation* – où chaque agent exprime ses désirs de déviation au regard de son propre concept de déviation.

Définition 5.29 (Jeu hédonique de déviation)
Un jeu hédonique de déviation – *ou* jeu de déviation – *est un triplet* $HGD = \langle N, (\succeq_i)_{a_i \in N}, (\mathbb{D}_i)_{a_i \in N} \rangle$ *où* $N = \{a_1, \ldots, a_n\}$ *est l'ensemble des agents,* $\succeq_i$ *les préférences de l'agent* a_i *vis-à-vis des coalitions et* $\mathbb{D}_i$ *le concept de déviation de l'agent* a_i.

Le problème de partitionnement de ce modèle reste le problème classique des jeux de coalitions hédoniques : trouver une partition $\Pi \in \mathcal{P}_N$ telle qu'aucun agent ne désire dévier. Cependant, contrairement au jeux de coalitions hédoniques canoniques, cette recherche de stabilité passe non pas par la satisfaction de propriétés globales pour tous les agents mais par l'absence de déviation désirable du point de vue d'un agent. Ainsi, du point de vue d'un agent a_i, une partition est *localement stable* lorsqu'il n'existe pas de déviation qui satisfasse son concept de déviation, c'est-à-dire lorsque $\mathbb{D}_i(\Pi, HGD) = \emptyset$. Nous avons donc les deux notions de stabilité suivantes :

Définition 5.30 (Stabilité)
Soit HGD un jeu hédonique de déviation et $\Pi \in \mathcal{P}_N$ *une partition.* Π *est* localement stable *du point de vue l'agent* $a_i \in N$ *si* $\mathbb{D}_i(\Pi, HGD) = \emptyset$*, et* Π *est* collectivement stable *si* $\forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset$*.*

Liens avec les concepts de solutions canoniques

Afin de montrer les liens entre les concepts de solution canoniques et les concepts de déviation que nous proposons, nous allons ici considérer une hypothèse d'homogénéité : tous les agents expriment le même concept de déviation. Nous prouvons ci-après le lien entre stabilité au sens de Nash et un concept de déviation associé. Les preuves pour les autres concepts canoniques sont semblables. Nous présentons ensuite dans le tableau 5.2 les correspondances entre les concepts de solution canoniques et les concepts de déviation.

Propriété 5.31
Soit HGD un jeu de déviation et $\Pi \in \mathcal{P}_N$ *une partition. Si* $\forall a_i \in N, \mathbb{D}_i := \Delta_I \wedge \Delta_R$*, alors l'équivalence suivante est vraie :*

$$\Pi \in NS \iff \forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset$$

Démonstration 5.31
Fixons un jeu de déviation HGD et une partition $\Pi \in \mathcal{P}_N$*. Par définition,* $\Pi \in NS$ *(où NS est l'ensemble des partitions stables au sens de Nash) si :*

$$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi) \tag{5.1}$$

Cette formulation de l'équilibre de Nash est équivalente à :

$$\forall a_i \in N, \nexists C \subseteq N, a_i \in C : C \setminus \{a_i\} \in \Pi \cup \{\emptyset\} \wedge C \succ_i C_i(\Pi) \tag{5.2}$$

Distinguons trois parties dans la formule :

1. $\nexists C \subseteq N, a_i \in C$*, est ici équivalent à* $\nexists C \in AllD_i(\Pi)$ *puisque C doit nécessairement être différente de* $C_i(\Pi)$*,*
2. $C \setminus \{a_i\} \in \Pi \cup \{\emptyset\}$ *est ici équivalent par définition à* $\Delta_I(a_i, C, \Pi, HGD)$*, c'est-à-dire C satisfait la condition de déviation individuelle,*

Concept de solution	**Concept de déviation**
Stabilité au sens de Nash	$\Delta_I \wedge \Delta_R$
Stabilité individuelle	$\Delta_I \wedge \Delta_R \wedge \Delta_A^-$
Stabilité contractuelle de Nash	$\Delta_I \wedge \Delta_R \wedge \Delta_D^-$
Stabilité individuelle contractuelle	$\Delta_I \wedge \Delta_R \wedge \Delta_A^- \wedge \Delta_D^-$
Stabilité au sens du cœur forte	$\Delta_R \wedge \Delta_A$
Stabilité au sens du cœur faible	$\Delta_R \wedge \Delta_A^-$
Optimalité	Δ_O
Pareto-optimalité	$\Delta_R \wedge \Delta_{PO}^-$

TABLE 5.2 – Association entre concepts de solution et concepts de déviation

3. $C \succ_i C_i(\Pi)$ *est ici équivalent par définition à* $\Delta_R(a_i, C, \Pi, HGD)$*, c'est-à-dire* C *satisfait la condition de rationnalité.*

Ainsi, une partition est stable au sens de Nash si, pour aucun agent, il n'existe pas de déviation individuelle vers une coalition déjà existante dans Π *qui soit rationnelle. Nous pouvons alors réécrire la formule 5.2 par :*

$$\forall a_i \in N, \nexists D \in AllD_i(\Pi) : \Delta_I(a_i, D, \Pi, HGD) \wedge \Delta_R(a_i, D, \Pi, HGD) \tag{5.3}$$

Par hypothèse, $\forall a_i, \mathbb{D}_i := \Delta_I(a_i, D, \Pi, HGD) \wedge \Delta_R(a_i, D, \Pi, HGD)$*. La formule 5.3 est alors équivalente à :*

$$\forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset \tag{5.4}$$

Ainsi, par définition, une partition Π *est stable au sens de Nash si le concept de déviation* $\mathbb{D}_i := \Delta_I(a_i, D, \Pi, HGD) \wedge \Delta_R(a_i, D, \Pi, HGD)$ *est vide pour tout les agents.* □

La table 5.2 indique les concepts de déviation $\mathbb{D}_i$ correspondant aux différents concepts de solution canoniques, c'est-à-dire tels que si tous les agents expriment $\mathbb{D}_i$ alors $\Pi \in SC \iff \forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset$. Remarquons que si les concepts de solution canoniques présentent des relations d'inclusion alors ces relations se retrouvent également entre les concepts de déviation. Par exemple, trivialement, pour une partition Π, s'il existe une déviation $D \in AllD_i(\Pi)$ telle que $D \vDash \Delta_I \wedge \Delta_R$ alors D satisfait également le concept de déviation $\Delta_I \wedge \Delta_R \wedge \Delta_A$. Ainsi, une telle partition Π ne satisfait pas les concepts de déviation associés à la stabilité individuelle et à la stabilité au sens de Nash. Nous retrouvons alors le fait que toute partition qui n'est pas individuellement stable ne peut pas être stable au sens de Nash, représenté classiquement par l'inclusion $NS \subseteq IS$. De manière générique, l'inclusion d'un concept de déviation $\mathbb{D}_i^1$ dans un autre concept de déviation $\mathbb{D}_i^2$ – noté $\mathbb{D}_i^1 \subseteq \mathbb{D}_i^2$ – correspond au fait que que toute déviation autorisée par $\mathbb{D}_i^1$ satisfait également les conditions de $\mathbb{D}_i^2$

Définition 5.32 (Concept de déviation inclus)
Un concept de déviation $\mathbb{D}_i^1$ *est* inclus *dans un concept* $\mathbb{D}_i^2$ *si, pour tout jeu hédonique de déviation et pour toute partition* $\Pi \in \mathcal{P}_N$*,* $D \in \mathbb{D}_i^1(\Pi, HGD) \implies D \in \mathbb{D}_i^2(\Pi, HGD)$*.*

Nous pouvons alors facilement déduire certaines de ces relations d'inclusion.

Propriété 5.33
Soit $\mathbb{D}_i^1$ et $\mathbb{D}_i^2$ deux concepts de déviation. Soit A (resp. B) l'ensemble des conditions de déviation qui définissent $\mathbb{D}_i^1$ (resp. $\mathbb{D}_i^2$). Si $B \subseteq A$, le concept de déviation $\mathbb{D}_i^1$ est inclus dans $\mathbb{D}_i^2$.

Démonstration 5.33
Fixons HGD un jeu quelconque et une partition $\Pi \in \mathcal{P}_N$. Soit $\mathbb{D}_i^1$ et $\mathbb{D}_i^2$ deux concepts de déviation. Soit A (resp. B) l'ensemble des conditions de déviation qui définissent $\mathbb{D}_i^1$ (resp. $\mathbb{D}_i^2$). Suposons que $B \subseteq A$ et montrons que nous avons nécessairement l'inclusion $\mathbb{D}_i^1 \subseteq \mathbb{D}_i^2$.

Les concepts de déviations $\mathbb{D}_i^1$ et $\mathbb{D}_i^2$ peuvent être définis par les formes normales conjonctives :

$$\mathbb{D}_i^1 := \bigwedge_{\Delta_X \in A} \Delta_X \text{ et } \mathbb{D}_i^2 := \bigwedge_{\Delta_X \in B} \Delta_X$$

Comme $B \subseteq A$, nous pouvons réécrire $\mathbb{D}_i^1$ sous la forme suivante :

$$(\bigwedge_{\Delta_{X_1} \in B} \Delta_{X_1}) \wedge (\bigwedge_{\Delta_{X_2} \in A \setminus B} \Delta_{X_2})$$

Ainsi,

$$\forall D \in AllD_i(\Pi) : D \vDash (\bigwedge_{\Delta_{X_1} \in B} \Delta_{X_1}) \wedge (\bigwedge_{\Delta_{X_2} \in A \setminus B} \Delta_{X_2}) \implies D \vDash \bigwedge_{\Delta_{X_1} \in B} \Delta_{X_1}$$

Par conséquent,

$$\forall D \in AllD_i(\Pi), D \in \mathbb{D}_i^1(\Pi, HGD) \implies D \in \mathbb{D}_i^2(\Pi, HGD)$$

Nous avons donc nécessairement l'inclusion $\mathbb{D}_i^1 \subseteq \mathbb{D}_i^2$. □

Pour illuster cette propriété, considérons les quatre concepts de déviation suivants :

1. $\mathbb{D}_i^1 := \Delta_I \wedge \Delta_R$ (Nash stabilité)
2. $\mathbb{D}_i^2 := \Delta_I \wedge \Delta_R \wedge \Delta_A$ (Stabilité Individuelle)
3. $\mathbb{D}_i^3 := \Delta_I \wedge \Delta_R \wedge \Delta_A \wedge \Delta_D$ (Stabilité Individuelle contractuelle)
4. $\mathbb{D}_i^4 := \Delta_R \wedge \Delta_A$ (Stabilité du Cœur)

Ici, nous obtenons les relations d'inclusion suivantes : $\mathbb{D}_i^3 \subseteq \mathbb{D}_i^2 \subseteq \mathbb{D}_i^1$ et $\mathbb{D}_i^3 \subseteq \mathbb{D}_i^4$. Nous retrouvons alors les relations d'inclusion entre les concepts de solution canoniques : $NS \subseteq IS \subseteq ICS$ et $CS \subseteq IS \subseteq ICS$.

Ces liens avec les concepts de solution classiquement utilisés dans les jeux de coalitions hédoniques nous amènent à un constat important. Même en nous limitant à 7 conditions

	Δ_I	Δ_C	$\Delta_I \vee \Delta_C$
Δ_R	NS	?	?
$\Delta_R \wedge \Delta_A$	IS	?	CS
$\Delta_R \wedge \Delta_D$	CNS	?	?
$\Delta_R \wedge \Delta_A \wedge \Delta_D$	ICS	?	?
$\Delta_R \wedge \Delta_{PO^-}$	?	?	PO
Δ_O	?	?	O
Δ_A	?	?	?
Δ_D	?	?	?
$\Delta_A \wedge \Delta_D$	?	?	?

TABLE 5.3 – Concepts de déviation non couverts

de déviations (5 portant sur la satisfaction des préférences et 2 sur l'identité des agents), de nombreux cas ne sont pas couverts par les concepts classiques. La table 5.3 met en avant certain [9] de ces manques dans la littérature. Les colonnes donnent les clauses portant sur les conditions d'identité et les lignes les clauses sur les conditions de préférence. Les « ? » représentent des concepts de solution ne correspondant à notre connaissance à aucun concept de solution canonique.

Comme nous l'avons fait remarquer précédemment, l'un des principaux manques vient du fait que tous les concepts de solution considérent la condition de rationnalité. Cependant, il est possible de considérer des agents qui cherchent à maximiser le bien-être social et ce même si la déviation est à leur détriment personnel. L'autre principal manque est l'absence de concepts de solution n'incluant que des déviations collectives. De tel concepts peuvent cependant représenter un agent ne désirant pas être le seul responsable de l'instabilité d'une partition.

3.2 Modéliser la liberté, l'altruisme et l'hédonisme

Dans le processus de formation des coalitions, le choix des agents de rester ou de dévier peut être guidé par une éthique des vertues, représentée une valeur cardinale personnelle. De manière générale, nous proposons de définir pour une valeur v et un agent a_i un concept de déviation $\mathbb{D}_i^v$ tel que toute déviation D qui satisfait $\mathbb{D}_i^v$ est une déviation qui respecte la valeur v. Une partition stable Π représente une répartition des agents telle qu'aucun d'entre eux ne peut changer de coalition sans trahir ses valeurs. Pour illustrer notre propos, nous modélisons trois valeurs en nous fondant sur leur définition dans la littérature : la *liberté*, l'*altruisme* et l'*hédonisme*. Nous proposons ici des concepts minimaux dans le sens où ces concepts sont des conjonctions des conditions qui doivent être minimalement

9. Pour des raisons de lisibilité, nous ne présentons ici qu'un sous-ensemble des concepts de solution manquants.

satisfaites. Cependant, nous ne considérons pas que cette association comme absolue. En effet, tout autre concept de déviation qui satisfait aux moins ces conditions satisfait également la valeur correspondante. Ainsi, pour une même valeur, des agents hétérogènes peuvent y associer des concepts de déviation différents.

Modélisation de la Liberté

La liberté est une valeur qui a été grandement étudiée dans la littérature philosophique et politique. Considérons les quatre définitions (non exhaustives) suivantes :

La Liberté selon John Stuart Mill : Dans [Mill, 1869], deux formes de libertés sont considérées : la « liberté de pensée » et la « liberté d'action ». La *liberté de pensée* représente le fait que tout homme doit pouvoir former son opinion et l'exprimer sans réserve. Mill indique que satisfaire cette liberté est un impératif pour l'intelligence et la nature morale de l'Homme. La *liberté d'action* désigne, elle, le fait que « les hommes soient libres d'agir selon leurs opinions, c'est-à-dire libres de les appliquer à leur vie sans que leurs semblables les en empêchent physiquement ou moralement, tant que leur liberté ne s'exerce qu'à leurs seuls risques et périls. »

La Liberté dans la Constitution : Selon l'article 4 de la Déclaration des Droits de l'Homme et du Citoyen de 1789 [DDHC, 1789], « la liberté consiste à pouvoir faire tout ce qui ne nuit pas à autrui : ainsi, l'exercice des droits naturels de chaque homme n'a de bornes que celles qui assurent aux autres Membres de la Société la jouissance de ces mêmes droits. Ces bornes ne peuvent être déterminées que par la Loi. »

La Liberté selon Montesquieu : « Il est vrai que dans les démocraties le peuple parait faire ce qu'il veux ; mais la liberté politique ne consiste point à faire ce que l'on veut. Dans un État, c'est-à-dire dans une société où il y a des lois, la liberté ne peut consister qu'à pouvoir faire ce que l'on doit vouloir, et à n'être point contraint de faire ce que l'on ne doit pas vouloir. Il faut se mettre dans l'esprit ce que c'est que l'indépendance, et ce que c'est que la liberté. La liberté est le droit de faire tout ce que les lois permettent ; et si un citoyen pouvait faire ce qu'elles défendent, il n'aurait plus de liberté, parce que les autres auraient tout de même ce pouvoir. » [de Montesquieu, 1867] (livre XI, Chapitre III)

La Liberté selon Durkheim : « La vraie liberté individuelle ne consiste donc pas dans la suppression de toute réglementation, mais est le produit d'une réglementation ; car cette égalité n'est pas dans la nature. » [Durkheim, 1893] (Chapitre II)

Dans ces quatres définitions, une même contrainte apparait clairement : l'absence d'atteinte aux autres. Cette contrainte est illustrée par la maxime populaire : « La liberté des uns s'arrête là où commence celle des autres ». En effet, dans le cadre des jeux de déviation, la liberté de pensée telle que définie par Mill correspond au fait que chaque agent est libre d'exprimer des préférences vis-à-vis des coalitions. Il reste donc à satisfaire

la liberté d'action qui consiste à ne pas changer de coalition si cela nuit à un autre agent. Ainsi, un agent est libre de dévier de sa coalition courante si :

1. il ne nuit pas à ceux qu'il rejoint,
2. il ne nuit pas à ceux qu'il quitte.

Ces deux points correspondent respectivement aux formes affaiblies des conditions d'Acceptation (Δ_A^-) et de Défection (Δ_D^-). Remarquons que, bien que Durkheim met en avant une notion de liberté au niveau individuel, la liberté s'applique à tous les agents et il n'y a donc pas de conditions d'identité. De plus, la liberté, tant qu'elle ne nuit pas à autrui, peut nuire à l'agent déviant. Il n'y a donc pas non plus de condition de rationnalité. Ainsi, nous pouvons donc définir la liberté par le concept de déviation $\Delta_A^- \wedge \Delta_D^-$.

Modélisation de l'Altruisme

Considérons maintenant la valeur d'altruisme. S'il y a débat sur l'existence d'actes purement altruistes en s'appuyant sur le fait que tout acte peut être motivé par une forme ou une autre de compensation égoïste [Batson, 2014], cette considération prend sens dans un contexte dynamique où les actions de l'agent à un instant donné influent sur les actions et les croyances des autres agents dans le futur. Par exemple, [Nongaillard et Mathieu, 2011] ont montré que des stratégies altruistes où des agents acceptent des offres désavantageuses pour eux permet d'atteindre plus tard une solution optimale. Cependant, comme les jeux de déviation que nous considérons sont statiques, la question des motivations liées à la mise en œuvre d'un comportement altruiste est hors de notre cadre d'étude. Afin de définir des déviations altruistes, nous considérons les deux définitions suivantes :

L'Altruisme selon Rand : Dans [Rand, 1964], l'altruisme est vu comme la réponse à l'égoïsme : « The ethics of altruism has created the image of the brute, as its answer, in order to make men accept two inhuman tenets : (a) that any concern with one's own interests is evil, regardless of what these interests might be, and(b) that the brute's activities are in fact to one's own interest (which altruism enjoins man to renounce for the sake of his neighbors) ». Plus récemment, [Rand, 2005] a redéfini l'altruisme comme le fait de chercher à satisfaire en premier lieux le bien-être des autres avant son propre intêret : « altruism is the doctrine which demands that man lives for others and places others above self ».

L'Altruisme selon Comte : L'altruisme est le fait de « vivre pour autrui » [Comte, 1852].

Il est important de noter que Rand comme Comte définissent l'altruisme en opposition à l'égoïsme. Si nous considérons l'égoïsme d'un agent comme le fait de satisfaire uniquement ses préférences, alors la stabilité au sens de Nash modélise l'égoïsme. Comme le fait de vouloir satisfaire prioritairement les préférences des autres agents n'est, à notre connaissance, représenté par aucun concept de solution canonique, nous proposons de définir une nouvelle condition de déviation consistant à améliorer la satisfaction des préférences d'au moins un autre agent.

Définition 5.34 (Condition d'altruisme)
Soit $\Pi \in \mathcal{P}_N$ *et* $D \in All D_i(\Pi)$ *une déviation. D satisfait la condition d'altruisme (notée* Δ_{alt}*), si pour* $\Pi' = [D \to \Pi]$,

$$\exists a_j \in N \setminus \{a_i\} : C_j(\Pi') \succ_j C_j(\Pi)$$
$$\wedge \forall a_k \in N \setminus \{a_i\} : C_k(\Pi') \succeq_k C_k(\Pi)$$

La première partie de la condition implique que la déviation doit être profitable pour au moins un agent, la seconde qu'elle ne doit pas être au désavantage d'un tiers. Cette définition de l'altruisme insiste sur le fait qu'il s'agisse avant tout d'un acte personnel que nous pouvons représenter par la condition d'individualité Δ_I. De plus, un acte altruiste peut être soit à l'avantage, soit au désavantage de l'agent qui l'effectue. Durkheim appelait ce dernier cas un *sucide altruiste* [Durkheim, 1897] : un agent commet un suicide altruiste lorsqu'il effectue une déviation qui lui est défavorable pour le bien d'un autre. Nous définissons alors deux concepts de déviation associés à l'altruisme :

Altruisme : $\mathbb{D}_i := \Delta_I \wedge \Delta_{alt}$

Suicide altruiste : $\mathbb{D}_i := \Delta_I \wedge \Delta_{alt} \wedge \neg\Delta_R$

Comme dit précédemment, [Rand, 1964] oppose l'altruisme à l'égoïsme. Si nous considérons un égoïsme modélisé par une stabilité au sens de Nash, cette opposition se retrouve bel et bien lorsque nous considérons le suicide altruiste. En effet, le suicide altruiste implique nécessairement des déviations irrationnelles [10]

Modélisation de l'Hédonisme

L'hédonisme est une valeur morale fondée sur la satisfaction des plaisirs personnels. Si la question de la recherche du plaisir a été fortement discutée, en particulier par les philosophes cyrénaïques et épicuriens, Épicure indiquait que « le plaisir excessif actuel doit être évité s'il conduit à une douleur future ». Plus récement, [Mill, 1889] discutait ainsi que la satisfaction des plaisirs : « pleasure, and freedom from pain, are the only things desirable as ends ; and that all desirable things are desirable either for the pleasure inherent in themselves, or as means to the promotion of pleasure and the prevention of pain ». Dans le deux cas, la satisfaction des plaisirs ne prend de sens que dans l'évitement des douleurs. Ainsi, nous nous fonderons sur la définition de l'hédonisme donnée par [Chamfort, 1857] : « Jouis et fais jouir, *sans faire de mal ni à toi, ni à personne*, voilà je crois, toute la morale ».

D'un côté, un agent hédonique doit chercher à satisfaire ses propres préférences. De l'autre coté, l'agent doit aussi satisfaire les préférences des autres. Ces deux aspects se traduisent respectivement par la satisfaction des conditions de rationalité (Δ_R), d'acceptation (Δ_A) et de défection (Δ_D). Ainsi, à partir de cette définition, l'hédonisme peut être associé au concept de déviation $\mathbb{D}_i := \Delta_R \wedge \Delta_A \wedge \Delta_D$. En terme de concept de solution,

10. Trivialement, le suicide altruiste inclut par définition $\neg\Delta_R$.

cet hédonisme est équivalent à un concept de *stabilité du cœur contractuelle*, concept de solution qui n'existe pas dans la littérature classique. Notons que comme pour la stabilité du cœur, le concept d'hédonisme peut être affaibli en considérant des préférences non strictes. Cet hédonisme faible (que nous définissons par $\mathbb{D}_i := \Delta_R \wedge \Delta_A^- \wedge \Delta_D^-$) signifie que l'agent a_i va chercher à satisfaire ses préférences, sans aller à l'encontre des préférences des autres.

3.3 Propriétés de ces nouveaux concepts de solutions

Modéliser les trois valeurs précédentes à l'aide de concepts de déviation nous permet de définir des solutions à un jeu de coalitions qui ne sont pas couvertes par les concepts de solution classiquement utilisé dans la littérature. En faisant l'hypothèse que les agents désirent respecter les mêmes valeurs, nous pouvons définir les nouveaux concepts de solution suivants :

Stabilité au sens de la Liberté : $\Pi \in \mathcal{P}_N$ est *stable au sens de la Liberté* (noté $\Pi \in LS$) si et seulement si :

$$\forall a_i \in N, \forall C \in N_i : \exists a_j \in N \setminus \{a_i\} : C_j(\Pi) \succ_j C_j([C \to \Pi])$$

Stabilité altruiste : $\Pi \in \mathcal{P}_N$ est *altruistement stable* (noté $\Pi \in AS$) si et seulement si :

$$\begin{aligned} &\forall a_i \in N, \nexists C \in N_i : \exists a_j \in N \setminus \{a_i\} : C_j([C \to \Pi]) \succ_j C_j(\Pi) \\ &\wedge \forall a_k \in N \setminus \{a_i\},\ C_k([C \to \Pi]) \succeq_j C_j(\Pi) \end{aligned}$$

Stabilité hédonique : $\Pi \in \mathcal{P}_N$ est *hédoniquement stable* (noté $\Pi \in HS$) si et seulement si :

$$\begin{aligned} &\forall a_i \in N, \nexists C \in N_i : C \succ_i C_i(\Pi) \wedge \forall a_j \in C, C \succ_j C_j(\Pi) \\ &\wedge \forall a_k \in N \setminus C : (\exists a_j \in C, C_k(\Pi) = C_j(\Pi)), C_k(\Pi) \setminus C \succ_k C_k(\Pi) \end{aligned}$$

La table 5.4 positionne ces trois nouveaux concepts de solution en fonction des concepts de déviation qui leur sont associés.

Il s'agit ici d'un complément du Tableau 5.3 où nos trois concepts de solution correspondent à des situations qui ne sont pas représentées par les concepts de solution canoniques. Étudions quelques propriétés de ces nouveaux concepts de solution en considérant d'un côté l'existence d'une solution qui les satisfait, et de l'autre leurs relations d'inclusion vis-à-vis des concepts de solution canoniques.

Existence des partitions stables au sens de la Liberté

La stabilité au sens de la liberté est un concept de solution où il n'existe pas nécessairement de solution stable.

	Δ_I	$\Delta_I \vee \Delta_C$
Δ_R	Nash-stabilité	?
$\Delta_R \wedge \Delta_A$	Stabilité Individuelle	Stabilité du Cœur
$\Delta_R \wedge \Delta_D$	Stabilité Contractuelle de Nash	?
$\Delta_R \wedge \Delta_A \wedge \Delta_D$	Stabilité Individuelle Contractuelle	**Hédonisme**
$\Delta_R \wedge \Delta_{PO^-}$	?	Pareto-Optimalité
Δ_O	?	Optimalité
Δ_{alt}	**Altruisme**	?
$\neg\Delta_R \wedge \Delta_{alt}$	**Suicide altruiste**	?
$\Delta_A^- \wedge \Delta_D^-$	?	**Liberté**

TABLE 5.4 – Concepts de solution en fonction des concepts de déviation

Propriété 5.35
Il existe des jeux hédoniques HG tel que $LS = \emptyset$.

Intuitivement, la stabilité au sens de la Liberté est un concept de solution pouvant être vide car les agents peuvent désirer réaliser des déviations irrationnelles tant que celles-ci ne mécontentent pas les autres agents.

Démonstration 5.35 (Par l'exemple)
Considérons le jeu de coalitions hédonique HG suivant :

- $N = \{a_1, a_2\}$
- $\{a_1, a_2\} \succ_1 \{a_1\}$
- $\{a_2\} \succ_2 \{a_1, a_2\}$

Dans ce jeu, nous avons deux partitions possibles : $\Pi_1 = \{\{a_1\}, \{a_2\}\}$ *et* $\Pi_2 = \{\{a_1, a_2\}\}$. Π_1 *n'est pas stable au sens de la Liberté puisque* a_2 *peut réaliser la déviation* $D_1 = \{a_1, a_2\}$. *Cette déviation est cependant irrationnelle en terme de satisfaction des préférences pour* a_2. *De même,* Π_2 *n'est pas stable au sens de la Liberté puisque* a_1 *peut réaliser la déviation* $D_2 = \{a_1\}$. *Ainsi, ce jeu HG ne possède pas de partition stable au sens de la Liberté.* □

Existence des partitions hédoniquement stables

La stabilité hédonique est un concept de solution non vide.

Propriété 5.36
Soit $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ *un jeu hédonique. Il existe nécessairement au moins une partition* $\Pi \in \mathcal{P}_N$ *tel que* $\Pi \in HS$.

Démonstration 5.36
Nous allons prouver l'existence d'une partition hédoniquement stable par construction d'un jeu hédonique de déviation où tous les agents considèrent comme concept de déviation

$\mathbb{D}_i := \Delta_R \wedge \Delta_A \wedge \Delta_D$. *Nous montrons qu'il existe nécessairement au moins une partition* $\Pi \in \mathcal{P}_N$ *tel que* $\forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset$, *cette partition étant hédoniquement stable.*

Soit un jeu de déviation $HGD = \langle N, (\succeq_i)_{a_i \in N}, (\mathbb{D}_i)_{a_i \in N} \rangle$ *avec n agents et une première partition* $\Pi_1 = \{\{a_1\}, \ldots, \{a_n\}\}$.

Considérons dans un premier temps l'agent a_1. $\mathbb{D}_1(\Pi_1, HGD) = \emptyset$ *signifie que quelles que soient les déviations que* a_1 *propose, cela est au désavantage d'au moins un autre agent. Notons alors* $\Pi_2 = \Pi_1$.

Supposons maintenant que $D \in \mathbb{D}_1(\Pi_1, HGD) \neq \emptyset$. *Soit* $D^* \in \mathbb{D}_1(\Pi_1, HGD)$ *telle que* $\forall D \in \mathbb{D}_1(\Pi_1, HGD), D^* \succ_1 D$. *Soit* $\Pi_2 = [D^* \to \Pi_1]$. *Par choix de* D^*, *nous avons alors nécessairement* $\mathbb{D}_1(\Pi_2, HGD) = \emptyset$.

Considérons maintenant l'agent a_2 *à partir de la partition* Π_2. *Par construction de* Π_2, *s'il existe* $D \in \mathbb{D}_2(\Pi_2, HGD)$, *alors nécessairement* $a_1 \notin D$. *L'agent* a_2 *peut ainsi effectuer la déviation* $D^{*_2} \in \mathbb{D}_2(\Pi_2, HGD)$ *telle que* $\forall D \in \mathbb{D}_2(\Pi_2, HGD), D^{*_2} \succ_2 D$ *pour passer dans une partition* $\Pi_3 = [D^{*_2} \to \Pi_2]$.

Ainsi en appliquant successivement pour chaque agent les déviations appartenant à $D^{*_i} \in \mathbb{D}_i(\Pi_i, HGD)$, *nous obtenons nécessairement une partition* Π_n *telle que* $\forall a_i \in N, \mathbb{D}_i(\Pi_n, HGD) = \emptyset$, *c'est-à-dire une partition hédoniquement stable.* □

Existence des partitions altruistement stables

Il n'existe pas nécessairement de partition altruistement stable.

Propriété 5.37
Il existe des jeux hédoniques HG tel que $AS = \emptyset$.

Démonstration 5.37 (Par l'exemple)
Reprenons l'exemple de la Preuve 5.35. Π_1 *n'est pas altruistement stable puisque pour satisfaire les préférences de* a_1, a_2 *désire la déviation* $D_1 = \{a_1, a_2\}$. $\Pi_2 = \{\{a_1, a_2\}\}$ *n'est elle non plus pas altruistement stable puisque l'agent* a_1 *désire la déviation* $\{a_1\}$ *pour satisfaire les préférences de* a_2. *Ainsi, ce jeu ne possède pas de partition altruistement stable.* □

De manière intéressante, ce cas illustre des situations où par « politesse » deux personnes se laissent mutuellement l*a priorité*, conduisant à des situations d'interblocage.

Relations d'inclusion des nouveaux concepts

Certaines propriétés intéressantes des concepts de solution sont leurs relations d'inclusion. Pour cela, nous nous fondons sur la Propriété 5.33. Par lisibilité, nous dénotons dans la suite par $\mathbb{D}_{SC}$ le concept de déviation associé au concept de solution SC. Par exemple, $\mathbb{D}_{LS} := \Delta_A^- \wedge \Delta_D^-$. En fin de section, la figure 5.3 résume l'ensemble des relations d'inclusion entre les concepts de solution.

Considérons dans un premier temps le cas de la stabilité au sens de la Liberté et de la stabilité hédonique.

Propriété 5.38
Toute partition $\Pi \in \mathcal{P}_N$ stable au sens de la Liberté est nécessairement hédoniquement stable.

Démonstration 5.38
Nous avons les deux concepts de déviation : $\mathbb{D}_{LS} := \Delta_A^- \wedge \Delta_D^-$ et $\mathbb{D}_{HS} := \Delta_R \wedge \Delta_A \wedge \Delta_D$.

Par définition des formes affaiblies des conditions de déviation, toute déviation D qui satisfait la condition Δ_A (resp. Δ_D) satisfait nécessairement sa forme affaiblie Δ_A^- (resp Δ_D^-). De par la Propriété 5.33, nous avons la relation d'inclusion $\mathbb{D}_{HS} \subseteq \mathbb{D}_{LS}$. Cette relation d'inclusion entre les concepts de déviation se traduit par la relation d'inclusion $LS \subseteq HS$. □

Considérons maintenant le cas de la stabilité hédonique et de la stabilité individuelle contractuelle.

Propriété 5.39
Toute partition $\Pi \in \mathcal{P}_N$ hédoniquement stable est nécessairement individuellement contractuellement stable.

Démonstration 5.39
Nous avons les deux concepts de déviation : $\mathbb{D}_{HS} := \Delta_R \wedge \Delta_A \wedge \Delta_D$ et $\mathbb{D}_{ICS} := \Delta_I \wedge \Delta_R \wedge \Delta_A \wedge \Delta_D$. De par la Propriété 5.33, nous avons la relation d'inclusion $\mathbb{D}_{ICS} \subseteq \mathbb{D}_{HS}$. Cette relation d'inclusion entre les concepts de déviation se traduit par la relation d'inclusion $HS \subseteq ICS$. □

Remarquons que comme $LS \subseteq HS$, nous avons également la relation d'inclusion $LS \subseteq ICS$. De la même manière, la stabilité hédonique est un concept de solution inclus dans les concepts de stabilité individuelle, de stabilité au sens de Nash et de stabilité au sens du cœur (la preuve suit le même principe que précédemment). Enfin, une partition Pareto-optimale est nécessairement hédoniquement stable.

Propriété 5.40
La stabilité hédonique satisfait les relations d'inclusion suivantes : $NS \subseteq IS \subseteq HS$, $CS \subseteq IS \subseteq HS$ et $PO \subseteq HS$.

Nous allons montrer ici uniquement la relation d'inclusion $PO \subseteq HS$.

Démonstration 5.40
Toute partition hédoniquement stable n'est pas nécessairement Pareto-optimale car la Pareto-optimalité considère des successions de déviations, ce que ne fait pas la stabilité hédonique. Nous montrons dans la suite que toute partition Pareto-optimale est nécessairement hédoniquement stable.

Considérons une partition $\Pi \in PO$ et supposons que $\Pi \notin HS$. Par définition de la stabilité hédonique, il existe une déviation D telle que, pour Π' la partition résultante de $[D \to \Pi]$, nous avons $\forall a_i \in N, C_i(\Pi') \subset C_i(\Pi)$. Cela va à l'encontre de la définition de la Pareto-optimalité et donc de notre hypothèse de $\Pi \in PO$. Nous avons donc une contradiction. □

La stabilité au sens de la liberté ne présente pas de relation d'inclusion avec le concept de stabilité individuelle (et par extention avec la stabilité au sens du coeur et la stabilité au sens de Nash).

Propriété 5.41
La stabilité au sens de la Liberté satisfait la relation $IS \not\subseteq LS$.

Démonstration 5.41 (Par l'exemple)
Considérons dans un premier temps le jeu HG_1 avec $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ avec :

- $N = \{a_1, a_2, a_3\}$
- $\{a_1, a_3\} \succ_1 \{a_1, a_2\} \succ_1 \{a_1\}$
- $\{a_1, a_2\} \succ_2 \{a_2\}$
- $\{a_1, a_3\} \succ_3 \{a_3\}$

Soit la partition $\Pi = \{\{a_1, a_2\}, \{a_3\}\}$. Cette partition n'est pas individuellement stable puisque l'agent a_1 peut effectuer la déviation $D = \{a_1, a_3\}$ qui satisfait les conditions Δ_I, Δ_R et Δ_A. Par contre, elle n'est pas stable au sens de la liberté puisqu'il n'existe pas de déviation qui satisfait la condition Δ_D. Nous avons ainsi une partition $\Pi\mathcal{P}_N$ telle que $\Pi \in IS$ et $\Pi \notin LS$.

Considérons maintenant le jeu HG_2 avec $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ avec :

- $N = \{a_1, a_2, a_3\}$
- $\{a_1, a_2, a_3\} \succ_1 \{a_1\}$
- $\{a_2, a_3\} \succ_2 \{a_1, a_2, a_3\} \succ_2 \{a_2\}$
- $\{a_2, a_3\} \succ_3 \{a_1, a_2, a_3\} \succ_3 \{a_3\}$

Soit la partition $\Pi = \{\{a_1, a_2, a_3\}\}$. Cette partition est individuellement stable. Par contre, l'agent a_1 peut réaliser la déviation $D = \{a_1\}$ puisque celle-ci ne s'effectue qu'à ses propres dépends. Nous avons ainsi une partition $\Pi\mathcal{P}_N$ telle que $\Pi \notin IS$ et $\Pi \in LS$. □

Considérons enfin le cas de l'altruisme.

Propriété 5.42
La stabilité altruiste satisfait le relation $LS \subseteq AS$.

Démonstration 5.42
Rappelons que la définition de la condition d'altruisme Δ_{alt} implique nécessairement la satisfaction des deux conditions Δ_A^- et Δ_D^-. Ainsi, nous pouvons écrire $\mathbb{D}_{alt} := \Delta_I \wedge \Delta_{alt} \wedge \Delta_A^- \wedge \Delta_D^-$. En conséquence par la Propriété 5.33, nous avons nécessairement la relation d'inclusion $\mathbb{D}_{AS} \subseteq \mathbb{D}_{LS}$. En termes de concepts de solution, nous avons donc $LS \subseteq AS$.
□

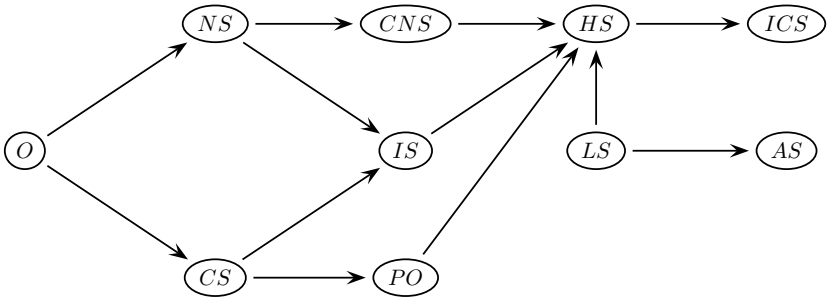

FIGURE 5.3 – Nouvelles relations d'inclusions entre les concepts de solution

Remarquons que comme l'altruisme permet (voire oblige pour le cas du suicide altruiste) les déviations irrationnelles, il n'existe pas de relation d'inclusion entre la stabilité altruiste et les concepts de solution canoniques (dont la Pareto-optimalité), ni entre la stabilité altruiste et la stabilité hédonique.

Enfin, la figure 5.3 résume l'ensemble des relations d'inclusion entre les concepts de solution canoniques et les concepts que nous proposons. Un arc allant du concept A au concept B ($A \longrightarrow B$) signifie que A est inclus dans B.

Bilan et animation scientifique

Nous avons présenté dans ce chapitre un exemple de travaux réalisés dans le cadre de l'axe de recherche sur l'éthique des agents autonomes. Cet axe était en gestation depuis 2008 lorsque nous avons rejoint le groupe de travail « Droits et Devoirs des Agents Autonomes » du GDR I3. À cette occasion, nous avons édité avec Olivier Boissier (École des Mines de Saint-Etienne) et Catherine Tessier (Onera) un numéro spécial de la Revue d'Intelligence Artificielle sur cette thématique en 2010, puis organisé un atelier international « Rights and Duties of Autonomous Agents » à ECAI 2012. Toutefois, cet axe de recherche a pris définitivement corps avec le projet ANR ETHICAA – projet pluridisciplinaire entre informatique et sciences sociales regroupant six partenaires – que nous avons coordonné de 2014 à 2018.

Les travaux présentés en sections 1 et 2 ont été réalisés dans le cadre de la thèse de Nicolas Cointe (2014 – 2017) financée par le projet ETHICAA et co-encadrée avec Olivier Boissier. Ces travaux ont fait l'objet de publications internationales [Cointe *et al.*, 2018, Cointe *et al.*, 2016b, Cointe *et al.*, 2016c] et nationales [Cointe *et al.*, 2017c, Cointe *et al.*, 2017a, Cointe *et al.*, 2016a] et ont été implémentés dans le contexte d'une application de gestion de portefeuille d'actions. Des variantes de ce modèle ont aussi été étudiées lors de l'encadrement des stages de master de Christopher Letruc (2016) et Yohan Bacquey (2017). Le premier visait à définir un cadre d'argumentation formelle permettant de faire du raisonnement pratique en tenant compte de critères éthiques. Le second consistait à modéliser un ensemble de vertus spécifiques aux agents artificiels proposées par la philosophe Kari Gwen Coleman. Enfin, les travaux présentés en section 3 se fondent sur le stage de master de Florent Benavant (2015) que nous avons encadré sur l'éthique dans le contexte des jeux de coalitions à utilité transférable. Ces travaux ont pris leur forme actuelle lors du post-doctorat de Thibaut Vallée (2017 – 2018).

Les questions soulevées par cet axe de recherche – et que nous avons publiées dans [Cointe *et al.*, 2017b, Vallée et Bonnet, 2017, Bonnet *et al.*, 2016, Belloni *et al.*, 2015a, Belloni *et al.*, 2015b] – nous ont permis de développer une animation scientifique importante. Au niveau national, nous avons organisé deux journées « Éthique et Intelligence Artificielle » lors des Plateforme Francophone Intelligence Artificielle en 2015 et 2018 et avons été invité à donner une conférence dans le cadre de l'atelier WACAI 2018. Au niveau international, nous avons été invité en 2015 à participer au séminaire « Normative Human-Robot Interaction » du groupe de recherche « Moral Competence in Computational Architectures for Robots » et à donner un cours à l'école d'été « Responsible Artificial Intelligence » organisée par Virginia Dignum (TU Delft) lors de ECAI 2016. Nous avons également organisé une table ronde « Ethics and autonomous agents » à la conférence CEPE (International Conference on Computer Ethics and Philosophical Enquiry) en 2014 et un atelier « Ethical Design of Intelligent Agents » à ECAI 2016. Enfin, nous avons collaboré en 2018 avec la juriste Maja Brkan (Maastricht University) et avons été invité à présenter ce travail lors du séminaire « Innov-AI-tion Law for Technology 4.0 ».

Enfin, la question de l'éthique de l'intelligence artificielle a été visible du grand public ces dernières années, et nous avons mené des actions de vulgarisation à partir de 2016. Nous avons donné en 2016 et 2018 des conférences grand public dans le cadre de la Fête de la Science, de Pint of Science et d'un cycle de conférence en anthropologie sociale à l'Université de Caen, et sur invitation des associations Relais d'Sciences et Stella Incognita (association pour la promotion de la recherche sur la science-fiction).

Troisième partie

Conclusion

Chapitre 6

Bilan et perspectives de recherche

Sommaire

Ce chapitre de conclusion a pour objectif de finaliser la construction de notre projet de recherche. Après un bilan de ce projet en section 1, nous proposons deux types de perspectives : des perspectives qui visent à étendre les travaux présentés dans ce mémoire et qui sont détaillées en section 2, et des perspectives décrites en section 3 qui viennent compléter les croisements entre nos axes de recherche et nos approches formelles. Ces perspectives prennent respectivement la forme de sujets de master et de thèse qui nous semblent pertinents de proposer.

1 Bilan du projet de recherche

Si l'autonomie est une des caractéristiques principales des agents artificiels, elle ne prend corps que de manière relative dans certaines fonctions de l'agent. Toutefois, l'autonomie, quelle que soit sa forme, implique des problématiques soit au niveau de l'agent, soit au niveau du système. Ces problématiques, comme l'absence de contrôle direct, l'hétérogénéité des agents et l'ouverture des systèmes, induisent à leur tour des besoins en termes de fiabilité, d'honnêteté et d'éthique de la part des agents. Ce sont ces trois besoins qui fondent alors les trois axes de projet de recherche.

Ce dernier se structure autour de l'entrelacement de ces trois axes avec plusieurs approches formelles qui nous semblent pertinentes pour étudier des systèmes d'agents autonomes, que ce soit au niveau de la modélisation individuelle des agents avec les architectures BDI, au niveau de la modélisation des interactions avec les systèmes de réputation

ou au niveau de la prise de décision collective avec les jeux de coalitions. Les chapitres 1 et 2 de ce mémoire nous ont permis d'identifier neuf questions à traiter par le croisement de nos axes de recherche et de ces approches. Les chapitres 3, 4 et 5 ont détaillés les travaux réalisés dans chaque axe en traitant sept de ces neuf questions.

Nous ne répétons pas ici les bilans synthétiques associés à chacun de nos axes de recherche (respectivement pages 67, 102 et 138 pour l'étude de la fiabilité, de l'honnêteté et de l'éthique) et nous invitons le lecteur à se référer à notre curriculum vitae détaillé au chapitre 7. Toutefois, il nous semble pertinent de rappeler quelques éléments d'encadrement et d'animation scientifique saillants pour chaque axe.

— Le chapitre 3 a traité de l'étude de la fiabilité. Nous avons co-encadré une thèse et un stage de master autour de cet axe. Ce stage a été l'occasion d'initier une collaboration au niveau national avec l'INSA Rouen en la personne de Laurent Vercouter, et de réutiliser le modèle de bandits manchots développé au cours de la thèse.

— Le chapitre 4 a traité de l'étude de l'honnêteté. Nous avons co-encadré une thèse et deux stages de master autour de cet axe. Une seconde thèse est en cours de co-encadrement.

— Le chapitre 5 a traité de la représentation de l'éthique. L'animation scientifique et l'encadrement autour de cet axe a été profilique, en particulier en raison de la coordination du projet ANR ETHICAA que nous avons assuré entre six partenaires. Ce projet a été l'occasion pour nous de co-encadrer une thèse, d'encadrer un post-doctorat et trois stages de master. En plus de la collaboration nationale autour du projet ETHICAA et de l'organisation d'ateliers et de journées thématiques, nous avons pu mettre en place des collaborations internationales avec des chercheurs de l'Université de Technologie de Delf et de Maastricht University entre autre. Enfin, l'appétence du grand public pour les questions d'éthique et d'intelligence artificielle nous a conduit à faire de nombreuses interventions de vulgarisation.

Dans le cadre de notre projet de recherche, nous pouvons remarquer que les travaux que nous avons présentés entrelacent deux à deux un axe de recherche et une approche formelle pour traiter ensuite chaque question séparemment. Or, nous avons vu qu'il existe un lien fort entre la fiabilité d'un agent et son honnêteté. Un agent malhonnête peut-il être fiable ? Il existe aussi un lien fort entre honnêteté et éthique car l'honnêteté est une valeur morale. C'est pourquoi la perspective à long terme de notre projet de recherche est d'unifier l'ensemble des approches et problématiques afin d'*établir une approche de l'interaction multi-agent fondée sur des valeurs formellement caractérisées*. Cependant, une telle perspective doit nécessairement s'appuyer sur d'autres, plus précises à plus court terme, en étendant d'une part nos travaux vers des modèles plus généraux, et d'autre part en poussant plus loin l'étude de la fiabilité. Dans ce qui suit, nous passons en revue ces perspectives qui émanent de nos activités de recherche et qui vont nourrir des collaborations, des projets collaboratifs ainsi que la formation à la recherche (encadrement de masters et de thèses).

	Fiabilité	Honnêteté	Éthique
Confiance	Q2	Q5	Q8
Coalitions	Q3	Q6	Q9
Cognition	Q1	Q4	Q7

TABLE 6.1 – Rappel des croisement entre axes de recherche et approches formelles

2 Généralisation au cas par cas de nos travaux

Nous avons illustré nos questions de recherche par des travaux choisis que nous avons réalisés en collaboration avec des doctorants ou des post-doctorants. Bien évidemment, ces travaux n'ont pas la prétention de répondre définitivement à ces questions et il convient de les prolonger par des perspectives. La figure 6.1 rappelle nos différentes questions, sachant que les questions Q1 et Q3 n'ont pas été abordées dans ce mémoire. Pour chacune des sept questions restantes, nous présentons ci-dessous une perspective, sans prétendre nous limiter à celle-ci.

Q2 : Une étude des systèmes de réputation abstraits

Les modèles de bandits manchots ont de bonnes propriétés pour l'étude des systèmes de réputation, en particulier en permettant de représenter des politiques d'utilisation et de construction de la confiance. Une première étude de l'influence de ces politiques a été réalisée sur les systèmes que sont EigenTrust, FlowTrust et BetaReputation. Toutefois, chacun possède des propriétés statistiques qui lui sont propres. Par exemple, EigenTrust a une distribution de réputation avec un écart-type plus faible que les autres. Pour généraliser cette étude, nous proposons de définir une taxonomie des systèmes de réputation selon leurs propriétés en termes d'entrées (ex. : distribution des valeurs de confiance, procédures d'agrégation) et de sorties (ex. : distribution des valeurs de réputation, utilisation de la réputation) et d'en étudier les propriétés selon les politiques d'utilisation de la confiance.

Q4 : Une logique de la réputation de sincérité

Nous avons proposé une logique modale de la confiance en la sincérité d'un agent et avons étudié avec la *confiance commune* une des notions de confiance collective. Cependant, la confiance collective peut aussi prendre plusieurs aspects comme par exemple la *confiance réciproque* et la *confiance mutuelle*. Ici, la confiance réciproque pourrait être caractérisée par $T_{i,j}\phi \wedge T_{j,i}\phi$ et la confiance mutuelle par $T_{i,j}\phi \wedge T_{j,i}\phi \wedge T_{i,j}T_{j,i}\phi \wedge T_{j,i}T_{i,j}\phi$ et $T_{i,j}\phi \wedge T_{j,i}\phi$. Nous proposons donc de caractériser d'autres formes de confiance collective et d'en étudier les propriétés. Par exemple, il serait intéressant d'étendre cette logique de la confiance avec une notion de réputation (par exemple un agent est réputé être sincère pour un groupe d'agents si une majorité d'agents de ce groupe ont confiance en sa sincérité).

Q5 : Vers une généralisation des fonctions de filtrage

La notion de crédibilité dans les systèmes de réputation permet d'identifier des agents dont les témoignages semblent peu informatifs ou erronés. Toutefois, cette notion doit être utilisée conjointement avec une politique d'usage, représentée sous la forme d'une fonction de filtrage. Par exemple, nous avons montré que la divergence de Kullback-Leibler est une mesure de crédibilité efficace lorsque qu'elle est couplée avec un vote majoritaire entre un ensemble d'agents juges tirés au hasard. Nous proposons alors de définir une fonction de filtrage abstraite permettant d'exprimer, de par son paramétrage, une large famille de fonction. En effet, une généralisation de la fonction de filtrage permettrait une meilleure exploration expérimentale de l'espace des fonctions de filtrage et ainsi déterminer quels paramètres ont une influence positive et significative sur la robustesse du système de réputation. Intuitivement, ces paramètres pourraient être les types de témoignages considérés, la mesure de crédibilité, l'ensemble des agents impliqués dans le jugement et la règle qu'ils appliquent.

Q6 : Jeux hédoniques sous connaissances partielles

Dans notre étude de la robustesse des jeux hédoniques, nous avons montré que si les agents malhonnêtes n'avaient pas besoin de connaître le profil de préférence du jeu afin de mettre en œuvre une manipulation. Toutefois, nous avons aussi montré qu'il leur était nécessaire de connaître ce profil pour pouvoir décider si la manipulation est efficace. Nous pouvons nous interroger sur la robustesse de ces jeux lorsque cette hypothèse est remise en cause. Nous proposons alors d'introduire explicitement la connaissance d'un agent malhonnête par une représentation partielle du profil de préférence. Cette représentation pose alors la question de la redéfinition de l'efficacité d'une manipulation, par exemple autour de la probabilité qu'une manipulation soit efficace. Ceci permettrait d'étudier de manière plus fine la robustesse d'autres concepts de solution.

Q7 : Un modèle de jugement éthique des normes

Dans son acception courante, le jugement éthique permet de discriminer ce qui est acceptable ou inacceptable de faire au nom de valeurs et principes. Toutefois, l'éthique nous invite à porter un regard non seulement sur les actions que nous réalisons mais aussi sur les règles dont nous nous dotons. En effet, certaines règles ou lois peuvent être jugées iniques, contraires à la morale ou l'éthique. L'architecture de jugement éthique que nous avons proposé ne permet de juger que les actions des agents et non pas les règles qui contraignent ces actions. Nous proposons donc d'étendre ce modèle avec un jugement des normes, qu'elles soient individuelles ou collectives. Pour ce faire, une piste consisterait à formuler des décisions dans un espace contraint par des normes activées ou désactivée en fonction du jugement porté sur elles.

Q8 : Une instanciation de l'éthique de la confiance

L'architecture de jugement éthique que nous avons proposé s'intéresse principalement au jugement des actions des agents et nous avons proposé un modèle de confiance fondé sur les jugements portés sur les autres agents. Toutefois, dans ce travail, la décision de faire confiance n'est pas traitée explicitement comme une action et l'archicture doit être affinée afin de pouvoir exprimer une éthique de la confiance (qui caractériserait des contextes dans lesquels il est juste de faire confiance). Nous proposons donc d'étendre l'architecture afin de pouvoir formuler ces éthiques de la confiance, en s'appuyant par exemple sur les travaux en philosophie de [Horsburgh, 1960].

Q9 : Jeux de coalitions fondés sur un système de valeurs

Le modèle des *jeux de déviation hédoniques* permet à chaque agent d'exprime une valeur humaine cardinale, représentant sous forme de conjonctions des conditions qui lui sont propres, pour identifier les coalitions qui sont acceptables de son point de vue. Toutefois, l'éthique s'appuie rarement sur une unique valeur. Nous proposons alors de reformuler le modèle des jeux de déviation pour permettre aux agents d'exprimer une éthique des vertus non pas sur une unique valeur cardinale mais sur un ensemble de valeurs afin de modéliser la notion de *système de valeurs*. Par exemple, ces systèmes peuvent soit correspondre à des concepts de déviation satisfaisant plusieurs valeurs simultanément, soit à une relation de préférence entre plusieurs concepts de déviation.

3 Enrichissement de l'axe d'étude de la fiabilité

Comme indiqué précédemment, deux croisements entre axes de recherche et approches formelles n'ont pas pas été abordés dans ce mémoire : la question Q1 avec le croisement entre les questions de fiabilité et les modèles d'agents cognitifs, et la question Q3 avec le croisement entre les mêmes questions de fiabilité et les modèles de formation de coalitions. Pour chacun de ces croisements, nous proposons ci-dessous une piste de recherche qui nous semble pertinente pour une première approche de la question.

Q1 : Un modèle cognitif de l'usage de la confiance

La notion de confiance peut recouvrir de nombreux aspects : confiance interpersonnelle ou institutionnelle, confiance occurente ou dispositionnelle, confiance en la fiabilité, sincérité, honnêteté, ou coopération par exemple. Si de nombreux travaux se sont intéressés à modéliser les mécanismes cognitifs sous-jacents à ces aspects, très peu se sont penchés sur les mécanismes permettant à un agent de décider d'interagir avec un autre en vue de construire cette confiance. Par exemple, un agent peut avoir l'intention d'interagir avec un agent qu'il croit sincère mais qu'il ne croit pas encore fiable afin d'en éprouver la fiabilité et acquérir de l'information. Se posent alors les questions suivantes.

1. Comment modéliser l'intention d'interagir avec un autre agent ? En effet, selon la nature de la confiance entre les agents et sachant si un agent estime disposer de suffisament d'information, sa décision d'interagir avec un agent donné peut être différente.
2. Comment modéliser l'intention d'agir en vue que les autres agents puissent déduire que nous sommes fiables ? Cette question est liée à la précédente au sens où il s'agit de modéliser l'intention d'agir sachant un modèle de l'intention d'interagir des autres agents.

Q3 : Un système de réputation pour jeux de coalitions répétés

Dans le domaine de la formation de coalitions, la fonction caractéristique pour les jeux à utilité transférable ou le profil de préférences pour les jeux hédoniques est un paramètre exogène du jeu de coalitions, donné *a priori*. Toutefois, dans un contexte dynamique où les agents forment itérativement des coalitions au fur et à mesure que de nouveaux besoins se font sentir, la fonction caractéristique d'un jeu ou le profil de préférences est intuitivement lié à l'évaluation que les agents font de leurs précédentes interactions. Ainsi, il semble pertinent d'utiliser un système de réputation pour fonder un modèle dynamique de formation de coalitions. Cependant, les systèmes de réputation se fondent classiquement sur l'évaluation d'interaction deux-à-deux et non pas sur des interactions de groupe. Coupler systèmes de réputation et formation de coalitions dynamiques posent alors de nouvelles questions scientifiques.

1. Comment définir une relation de préférence sur les coalitions à partir des confiances et réputations individuelles des agents ? Si certaines approches, comme l'utilisation de préférences additives, permet d'agréger les préférences individuelles pour exprimer des préférences sur les coalitions, elles ne permettent pas de définir des politiques d'utilisation de la confiance, comme par exemple exprimer le fait un agent pourrait préférer être en coalition avec certains agents afin de pouvoir les évaluer par la suite.
2. Comment mettre à jour la confiance individuelle qu'un agent a envers un autre à partir de la seule observation de l'utilité obtenue par la coalition ? La difficulté principale de cette question tient au fait qu'il convient de prendre compte le contexte dans l'évaluation d'un agent. En effet, un agent peut être fiable au sein de certaines coalitions mais pas au sein d'autres.

Chapitre 7

Curriculum vitae

Sommaire

1 Informations personnelles

1.1 État civil

- Civilité : Monsieur
- Nom de famille : BONNET
- Prénom : Grégory
- Date de naissance : 14/08/1980
- Grade : MCF 27e section
- Établissement d'affectation : Université de Caen Normandie
- Unité de recherche : GREYC – CNRS UMR 6072

1.2 Formations et diplômes

- 1998 : Baccalauréat Littéraire option Expression Dramatique
- 2002 : DEUG Mathématiques Appliquées et Sciences Sociales (Limoges)
- 2003 : Licence d'informatique, mention Assez Bien (Limoges)
- 2004 : Maîtrise d'informatique, mention Bien (Limoges)
- 2005 : Master Recherche « Intelligence Artificielle », mention Assez Bien (Toulouse)
- 2008 : Doctorat d'informatique « Systèmes Embarqués » délivré par l'Université de Toulouse, intitulé « Coopération au sein d'une constellation de satellites » et soutenu le 17 novembre 2008 sous la direction de Catherine Tessier.

1.3 Parcours professionnel

- 2005 – 2008 : Études doctorales à l'Onera (Toulouse)
- 2009 – 2010 : Contrat post-doctoral à l'Institut Charles Delaunay (Troyes)
- depuis 2010 : Maître de conférences à l'Université de Caen Normandie (Caen)

2 Liste des publications

Les tables 7.1 et 7.2 résument l'ensemble de nos publications (respectivement internationales et nationales) depuis l'obtention de notre doctorat selon leur axe de recherche (en comptant nos travaux antérieurs mentionnés en introduction de ce mémoire sur les réseaux autonomes). Les publications internationales sont hiérarchisées selon leur classement Q1, Q2 et Q3 sur SJR [1] pour les journaux et A*, A, B et C sur CORE 2018 [2] pour

1. http://www.scimagojr.com/journalrank.php
2. http://portal.core.edu.au/conf-ranks/

les conférences. De plus, nous avons participé à la rédaction de trois rapports techniques et d'un livre blanc [3]. Nous avons aussi été co-éditeur de 3 numéros spéciaux de la Revue d'Intelligence Artificielle. Deux autres numéros spéciaux sont en cours de co-édition pour l'année 2018.

Thématique	**Journaux**			**Conférences**			
	Q1	**Q2**	**Q3**	**A***	**A**	**B**	**C**
Fiabilité					1	1	
Honnêteté				3			
Éthique				1		1	
Réseaux autonomes	1	1	1			5	4

TABLE 7.1 – Publications internationales par axe

Thématique	**Journaux**	**Conférences**
Fiabilité	1	2
Honnêteté	1	3
Éthique	3	4
Réseaux autonomes		2

TABLE 7.2 – Publications nationales par axe

2.1 Journaux internationaux

1. R. Makhloufi, G. Bonnet, G. Doyen et D. Gaiti (2014). A survey and performance evaluation of decentralized aggregation schemes for autonomic management. *International Journal of Network Management*, 24(6) :469–498.
2. I. Ullah, G. Doyen, G. Bonnet et D. Gaiti (2012). A bayesian approach for user aware peer-to-peer video streaming systems. *Signal Processing : Image Communication Special Issue on Advances in Video Streaming for P2P Network*, 27(5) :438–456.
3. I. Ullah, G. Doyen, G. Bonnet et D. Gaiti (2012). A survey and synthesis of user behavior measurements in video streaming systems. *IEEE Communications Surveys and Tutorials*, 14(3) :734–749.
4. G. Bonnet et C. Tessier (2009). Incremental adaptive organization for a satellite constellation. *Lecture Notes on Artificial Intelligence : Special Issue on Organized Adaptation in Multi-Agent Systems*, 5368 :108–125.

3. `https://ethicaa.greyc.fr/en/livrables.php`

2.2 Conférences internationales à comité de lecture

5. N. Cointe, G. Bonnet et O. Boissier (2018). Ethics-based cooperation in multi-agent systems. In *14th Annual Conference of the European Social Simulation Association*.
6. C. Leturc et G. Bonnet (2018). A normal modal logic for trust in the sincerity. In *17th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 175–183.
7. N. Cointe, G. Bonnet et O. Boissier (2016). Ethical judgment of agents' behaviors in multi-agent systems. In *15th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1106–1114.
8. A. Belloni, A. Berger, O. Boissier, G. Bonnet, G. Bourgne, P.-A. Chardel, J.-P. Cotton, N. Evreux, J.-G. Ganascia, P. Jaillon, B. Mermet, G. Picard, B. Rever, G. Simon, T. de Swarte, C. Tessier, F. Vexler, R. Voyer et A. Zimmermann (2015). Towards a framework to deal with ethical conflicts in autonomous agents and multi-agent systems. In *12th International Conference on Computer Ethics and Philosophical Enquiry*.
9. T. Vallée et G. Bonnet (2015). Using KL divergence for credibility assessment. In *14th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1797–1798.
10. T. Vallée, G. Bonnet et F. Bourdon (2014). Multi-armed bandit policies for reputation systems. In *13th International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 279–290.
11. T. Vallée, G. Bonnet, B. Zanuttini et F. Bourdon (2014). A study of Sybil manipulations in hedonic games. In *13th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 21–28.
12. G. Bonnet (2012). A protocol based on a game-theoretic dilemma to prevent malicious coalitions in reputation systems. In *20th European Conference on Artificial Intelligence*, pages 187–191.
13. R. Makhloufi, G. Doyen, G. Bonnet et D. Gaiti (2012). SAAM : A self-adaptive aggregation mechanism for autonomous management systems. In *13th IEEE/IFIP Network Operations and Management Symposium*, pages 667–670.
14. I. Ullah, G. Doyen, G. Bonnet et D. Gaiti (2012). An autonomous topology management framework for QoS enabled P2P video streaming systems. In *8th International Conference on Network and Service Management*, pages 126–134.
15. G. Bonnet, I. Ullah, G. Doyen, L. Fillatre, D. Gaiti et I. Nikiforov (2011). A semi-markovian individual model of users for P2P video streaming applications. In *4th International Conference on New Technologies, Mobility and Security*, pages 1–5.
16. R. Makhloufi, G. Doyen, G. Bonnet et D. Gaiti (2011). Situated vs. global aggregation shemes for autonomous management systems. In *12th IFIP/IEEE International Symposium on Integrated Network Management*, pages 1135–1139.

17. R. Makhloufi, G. Doyen, G. Bonnet et D. Gaiti (2011). Towards self-adaptive management frameworks : The case of aggregated information monitoring. In *7th International Conference on Network and Service Management*, pages 1–5.
18. R. Makhloufi, G. Doyen, G. Bonnet et D. Gaiti (2011). Impact of dynamics on situated and global aggregation schemes. In *5th IFIP International Conference on Autonomous Infrastructure, Management and Security*, pages 148–159.
19. I. Ullah, G. Doyen, G. Bonnet et D. Gaiti (2011). User behavior anticipation in P2P live video streaming systems through a bayesian network. In *12th IFIP/IEEE International Symposium on Integrated Network Management*, pages 337–344.
20. I. Ullah, G. Bonnet, G. Doyen et D. Gaiti (2010). Modeling user behavior in P2P live video streaming systems through a bayesian network. In *4th International Conference on Autonomous Infrastructure, Management and Security*, pages 2–13.
21. R. Makhloufi, G. Bonnet, G. Doyen et D. Gaiti (2009). Towards a P2P-based deployment of network management information. In *4th International Conference on Autonomous Infrastructure, Management and Security*, pages 26–37.
22. I. Ullah, G. Bonnet, G. Doyen et D. Gaiti (2009). Improving performance of ALM systems with bayesian estimation of peers dynamics. In *12th IFIP/IEEE International Conference on Management of Multimedia and Mobile Networks and Services*, pages 157–169.
23. G. Bonnet et C. Tessier (2007). Collaboration among a satellite swarm. In *6th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 287–294.

2.3 Ateliers internationaux à comité de lecture

24. N. Cointe, G. Bonnet et O. Boissier (2016). Multi-agent based ethical asset management. In *1st Workshop on Ethics in the Design of Intelligent Agents*, pages 52–57.
25. A. Belloni, A. Berger, O. Boissier, G. Bonnet, G. Bourgne, P.-A. Chardel, J.-P. Cotton, N. Evreux, J.-G. Ganascia, P. Jaillon, B. Mermet, G. Picard, B. Rever, G. Simon, T. de Swarte, C. Tessier, F. Vexler, R. Voyer et A. Zimmermann (2015). Dealing with ethical conflicts in autonomous agents and multi-agent systems. In *1st International Workshop Joint on Artificial Intelligence, Ethics and Society*.
26. I. Ullah, G. Doyen, G. Bonnet et D. Gaiti (2013). Toward user-classified P2P IPTV systems : A persona-based approach. In *5th International Workshop on Management of the Future Internet*, pages 1187–1190.
27. R. Makhloufi, G. Doyen, G. Bonnet et D. Gaiti (2011). Situated vs. global aggregation schemes for autonomous management systems. In *4th IFIP/IEEE Workshop on Distributed Autonomous Network Management Systems*, pages 1135–1139.

28. R. Makhloufi, G. Bonnet, G. Doyen et D. Gaiti (2009). Decentralized aggregation protocols in peer-to-peer networks : A survey. In *4th IEEE International Workshop on Modelling Autonomic Communications*, pages 111–116.
29. G. Bonnet et C. Tessier (2008). A trust model based on communication capabilities for physical agents. In *Workshop on Trust in Agent Societies at 7th AAMAS*, pages 2–6.
30. G. Bonnet et C. Tessier (2008). Multi-agent collaboration : A satellite constellation case. In *4th Starting Artificial Intelligence Researchers Symposium*, pages 24–35.
31. G. Bonnet et C. Tessier (2008). Incremental adaptive organization for a satellite constellation. In *Workshop on Organized Adaptation in Multi-Agent Systems at 7th AAMAS*, pages 63–78.
32. G. Bonnet et C. Tessier (2007). On-board cooperation for satellite swarms. In *Workshop on Artificial Intelligence in Space Application at 20th IJCAI*.

2.4 Journaux nationaux

33. G. Bonnet, B. Mermet et G. Simon (2017). Vérification formelle du respect de valeurs morales dans les SMA. *Revue d'Intelligence Artificielle*, 31(4) :449–470.
34. N. Cointe, G. Bonnet et O. Boissier (2017). Jugement éthique dans le processus de décision d'un agent BDI. *Revue d'Intelligence Artificielle*, 31(4) :471–499.
35. N. Cointe, G. Bonnet et O. Boissier (2017). Éthique collective dans les systèmes multi-agents. *Revue d'Intelligence Artificielle*, 31(1-2) :71–96.
36. T. Vallée, G. Bonnet et F. Bourdon (2015). Politiques de bandits manchots et crédibilité dans les systèmes de réputation. *Revue d'Intelligence Artificielle*, 29(3-4) :369–398.
37. G. Bonnet (2014). Un protocole fondé sur des dilemmes pour se prémunir des collusions dans les systèmes de réputation. *Revue d'Intelligence Artificielle*, 28(4) :411–431.
38. G. Bonnet et C. Tessier (2009). Évaluation d'un système multirobot : cas d'une constellation de satellites. *Revue d'Intelligence Artificielle*, 23(5-6) :565–592.

2.5 Conférences nationales à comité de lecture

39. C. Leturc et G. Bonnet (2018). Une logique modale pour la caractérisation des manipulations entre agents autonomes. In *12es Journées d'Intelligence Artificielle Fondamentale*.
40. N. Cointe, G. Bonnet et O. Boissier (2017). Coopération fondée sur l'éthique entre agents autonomes. In *25es Journées Francophones sur les Systèmes Multi-Agents*, pages 9–18.

41. C. Leturc et G. Bonnet (2017). Une logique modale normale de la confiance. In *11es Journées d'Intelligence Artificielle Fondamentale*, pages 189–198.

42. T. Vallée et G. Bonnet (2017). Jeux de coalitions hédoniques à concepts de solution multiples. In *25es Journées Francophones sur les Systèmes Multi-Agents*, pages 53–62.

43. G. Bonnet, B. Mermet et G. Simon (2016). Vérification formelle et éthique dans les SMA. In *24es Journées Francophones sur les Systèmes Multi-Agents*, pages 139–148.

44. N. Cointe, G. Bonnet et O. Boisser (2016). Jugement éthique dans les systèmes multi-agents. In *24es Journées Francophones sur les Systèmes Multi-Agents*, pages 149–158.

45. T. Vallée, G. Bonnet et F. Bourdon (2014). De l'utilisation des politiques de bandits manchots dans les systèmes de réputation. In *19e Congrès National sur la Reconnaissance de Formes et l'Intelligence Artificielle*.

46. G. Bonnet (2013). Un protocole fondé sur un dilemme pour se prémunir des collusions dans les systèmes de réputation. In *21es Journées Francophones sur les Systèmes Multi-Agents*, pages 9–18.

47. T. Vallée, G. Bonnet, B. Zanuttini et F. Bourdon (2013). Étude des attaques Sybil sur les jeux hédoniques. In *7es Journées Francophones sur les Modèles Formels de l'Interactions*.

48. R. Makhloufi, G. Doyen, G. Bonnet et D. Gaiti (2012). Une approche adaptative pour la surveillance d'informations agrégées dans les réseaux complexes. In *16e Colloque Francophone sur l'Ingénierie des Protocoles*.

49. I. Ullah, G. Bonnet, G. Doyen et D. Gaiti (2011). Un classifieur du comportement des utilisateurs dans les applications pair-à-pair de streaming vidéo. In *15e Colloque Francophone sur l'Ingénierie des Protocoles*.

50. G. Bonnet et G. Doyen (2010). Coopération entre systèmes multi-agents appliquée au contrôle de trafic sur les réseaux pair-à-pair. In *18es Journées Francophones sur les Systèmes Multi-Agents*, pages 181–190.

51. G. Bonnet et C. Tessier (2008). Évaluer un système multiagent physique : retour sur expérience. In *16es Journées Francophones sur les Systèmes Multi-Agents*, pages 13–22.

52. G. Bonnet et C. Tessier (2007). Coopération au sein d'une constellation de satellites. In *15es Journées Francophones sur les Systèmes Multi-Agents*, pages 171–180.

53. G. Bonnet, C. Tessier, M.-P. Charmeau et P. Dago (2006). Planification pour un essaim de satellite : rapport préliminaire. In *1er Journées Francophones Planification, Décision, Apprentissage pour la conduite de systèmes*, pages 47–48.

3 Animation et rayonnement scientifique

3.1 Encadrement doctoral

Au cours de notre post-doctorat, nous avons participé au co-encadrement (25%) de deux thèses à l'Université de Technologie de Troyes, toutes deux dirigées par Dominique Gaïti et Guillaume Doyen :

- **2008 – 2011.** Thèse de Ihsan Ullah. Gestion de performance des infrastructures de systèmes de multicast applicatif basé sur P2P. Bourse d'étude du Pakistan.
- **2008 – 2012.** Thèse de Rafik Maklhoufi. Vers une gestion adaptative des réseaux complexes : cas de la surveillance décentralisée des données agrégées. Allocation MRE.

Depuis notre prise de fonction à l'Université de Caen Normandie, nous avons co-encadré (ou co-encadront toujours l'une d'elles) trois thèses :

- **2012 – 2015.** Thèse de Thibaut Vallée. De la manipulation dans les systèmes multi-agents : une étude sur les jeux hédoniques et les systèmes de réputation. Allocation MRE. Co-encadrement à 90% avec le directeur de thèse François Bourdon, PR IUT de Caen.
- **2014 – 2017.** Thèse de Nicolas Cointe. Jugement éthique pour la décision et la coopération dans les systèmes multi-agents. Financement ANR. Co-direction à 50% avec Olivier Boissier, PR École Supérieure des Mines de Saint-Étienne.
- **Depuis 2016.** Thèse de Christopher Leturc. Modélisation des manipulations dans les systèmes multi-agents. Allocation MRE. Co-direction à 50% avec Bruno Zanuttini, PR Université de Caen Normandie.

À cela s'ajoutent l'encadrement de six stages de master en informatique :

- **2011.** Stage de Sami Hajlaoui. Modélisation d'attaques Sybil pour la simulation de comportements de free-riding sur les réseaux pair-à-pair.
- **2012.** Stage de Thibaut Vallée. Étude de la robustesse des jeux de coalitions hédoniques face aux attaques Sybil.
- **2015.** Stage de Florent Benavant. Formation de coalitions pour une gestion responsable du trading haute-fréquence.
- **2016.** Stage de Christopher Letruc. Un modèle de raisonnement pratique éthique.
- **2017.** Stage de Yohann Bacquey. Modélisation de vertus pour les agents autonomes.
- **2017.** Stage de Damien Lelerre. Système de réputation à témoignages confidentiels (co-encadré à 50% avec Laurent Vercouter, PR INSA Rouen).

La table 7.3 résume ces encadrements de stages et de doctorats selon nos axes de recherche.

Thématique	Encadrement	
	Master	Doctorat
Fiabilité	2	1
Honnêteté	1	1
Éthique	3	1
Réseaux autonomes P2P		2

TABLE 7.3 – Encadrements doctoral et de master par axe

3.2 Organisation d'événements

Dans le cadre de la thématique « Éthique et agents autonomes », nous avons été responsable de l'organisation de plusieurs ateliers ou tables rondes (nationaux et internationaux) :

- Atelier RDA2 (Rights and Duties of Autonomous Agents) à ECAI [4] 2012,
- Table ronde « Ethics and autonomous agents » à CEPE [5] 2014,
- Journée E&IA (Éthique et Intelligence Artificielle) lors de la PFIA [6] 2015,
- Atelier EDIA (Ethical Design of Intelligent Agents) à ECAI 2016,
- Table ronde « Ethics and autonomous agents » à ESOF [7] 2018,
- Journée E&IA (Éthique et Intelligence Artificielle) à la PFIA 2018.

Au-delà de cela, nous avons été membre du comité d'organisation des JFSMA 2012 et 2017, ainsi que de la PFIA 2017. En 2014, nous avons été président du comité de programme des RJCIA (Rencontres des Jeunes Chercheurs en Intelligence Artificielle).

3.3 Participation à des comités de programme

Depuis 2015, nous avons progressivement intégré les comités de programme de grandes conférences internationales liées à nos domaines de recherche (AAAI, AAMAS, IJCAI) ainsi que des conférences nationales dédiées à l'intelligence artificielle et aux systèmes multi-agents, ce qui est résumé sur la table 7.4. Nous avons aussi été relecteur à l'international pour JAAMAS (International Journal of Autonomous Agents and Multi-Agent Systems) et au national pour RIA (Revue d'Intelligence Artificielle).

4. European Conference on Artificial Intelligence
5. International Conference on Computer Ethics and Philosophical Enquiry
6. Plateforme Francophone Intelligence Artificielle
7. Euroscience Open Forums
8. International Conference on Autonomous Infrastructure, Management, and Security

Années	Conférences internationales	Conférences nationales
2012		RJCIA
2013	AIMS [8]	RJCIA
2014	AIMS	RJCIA
2015	IJCAI	JFSMA
2016	IJCAI	JFSMA
2017	AAMAS, IJCAI	JFSMA
2018	AAAI, AAMAS, IJCAI	CNIA, JFSMA

TABLE 7.4 – Participations à des comités de programme

3.4 Participation à des jurys de thèses

Nous avons été examinateur de 5 thèses extérieures à notre établissement.

- **2012.** Thèse de Frédéric Merle. Proposition d'une grille d'analyse pour la composition de systèmes P2P adaptés aux contextes applicatifs. Direction : Dominique Gaïti, Université de Technologie de Troyes.
- **2012.** Thèse de Yann Kupra. PrivaCIAS : Privacy as contextual integrity in decentralized multi-agent systems. Direction : Olivier Boissier et Laurent Vercouter, École Supérieure des Mines de Saint-Étienne.
- **2014.** Thèse de Omar Rihawi. Modelling and simulation of large scale situated multi-agent systems. Direction : Philippe Mathieu et Yann Secq, Université Lille 1.
- **2018.** Thèse de Lise-Marie Veillon. Apprentissage artificiel collectif. Direction : Henri Soldano et Gauvain Bourgne, Université Paris 13.
- **2018.** Thèse de Azzedine Benabbou. Génération dynamique de situations critiques en environnements virtuels : dilemme et ambiguïté. Direction : Dominique Lenne et Domitile Lourdeau, Université de Technologie de Compiègne.

3.5 Invitations et collaborations

Au niveau national, nous avons été invité en 2013 à donner un séminaire au LIP6 (équipe DESIR [9]) ayant pour titre « De l'utilisation des dilemmes pour limiter les manipulations dans les réseaux de confiance ». Nous avons aussi été invité au salon Documation 2014 pour participer à une table ronde autour du thème « La GED collaborative au service de la qualité et de l'efficacité » et à donner un conférence « Éthique et agents autonomes » à WACAI 2018.

Au niveau international, nous avons été invité à participer au séminaire Normative Human-Robot Interaction organisé à Innsbruck en 2015 par le groupe de recherche Moral

9. http://www-desir.lip6.fr/~sma-site/seminaires/seminaires2013.php

Competence in Computational Architectures for Robots [10]. Nous avons également été invité à donner un cours [11] intitulé « Architectures for ethical autonomous agents » à l'école d'été « Responsible Artificial Intelligence », organisée par Virginia Dignum (TU Delft) lors de ECAI 2016 (European Conference on Artificial Intelligence, notée A par le classement CORE 2018). Enfin, nous avons initié en 2018 une collaboration avec la juriste Maja Brkan (Maastricht University) et avons été invité à présenter ce travail lors du séminaire « Innov-AI-tion Law for Technology 4.0 ».

3.6 Activités de vulgarisation

Au-delà de participations régulières à la Fête de la Science, nous menons des actions de vulgarisation, en particulier associée à la thématique « Éthique et agents autonomes ». Dans ce cadre, nous avons été invité à donner plusieurs interventions (conférences de vulgarisation entre 2016 et 2018 pour :

- 2016 : Conférence pour Pint of Science Caen
- 2017 : Conférence pour Grand Témoin [12], Relais de Science Caen
- 2017 : Conférence pour Stella Incognita [13], Cherbourg
- 2018 : Conférence en Humanités Numériques (Université de Caen Normandie)
- 2018 : Conférence en Anthropologie sociale (Université de Caen Normandie)
- 2018 : Table ronde « Faut-il avoir peur de l'IA ? » pour l'association OPTIC, Caen

4 Responsabilités scientifiques et pédagogiques

4.1 Coordination de projets

Nous avons été coordinateur du projet ANR-CORD-13-0006 ETHICAA (Ethics and Autonomous Agents) qui a été financé de janvier 2014 à juillet 2018 (1000 keurs). Ce projet pluridisciplinaire (informatique et sciences humaines) a regroupé 6 partenaires (Ardans, Armines-Fayol, GREYC, LIP6, Institut Mines-Telecom, Onera).

4.2 Responsabilités scientifiques nationales

Nous sommes membre élus du Conseil National des Université section 27 pour la mandature 2016 – 2019.

10. https://hrilab.tufts.edu/muri13/
11. https://ethicaa.greyc.fr/media/files/ethicaa.tutorial.1.pdf
12. https://www.youtube.com/watch?v=G1ocyLvIqpM
13. http://stella-incognita.byethost18.com/

4.3 Responsabilités scientifiques locales

Au niveau de notre laboratoire, nous sommes responsable de la communication (site web) de notre équipe d'accueil depuis 2015 mais nous sommmes surtout responsable de l'organisation et de l'animation des séminaires I3 (Information, Interaction, Intelligence) du GREYC depuis 2011 avec 79 invités sur la période 2011 – 2018.

4.4 Responsabilités pédagogiques

Au niveau du département d'informatique de l'Université de Caen Normandie, nous avons été responsable de la communication sur la période 2012 – 2016 et membre élu du conseil du département sur la période 2013 – 2016. En 2016 et 2017, nous avons participé activement à la réflexion sur les maquettes de L3 et Master et à leurs rédactions. Enfin depuis 2017, nous sommes membre de la commission pédagogique du département et responsable de la formation en M1 Informatique (environ 60 étudiants avec la gestion des tâches de recrutement et de l'emploi du temps, ainsi que le suivi des projets et des stages).

D'un point de vue pédagogique, nous avons été responsable (et sommes toujours pour certaines) de 6 Unités d'Enseignement :

- Introduction à l'intelligence artificielle en L3 Informatique (2012 – 2016)
- Génie logiciel en L3 (2010 – 2016),
- Conception de logiciels en L2 (depuis 2014) et en L1 Informatique (depuis 2017)
- Intelligence artificielle distribuée en M1 Informatique (depuis 2011)
- Réseaux et système en Licence professionnelle Webmestre (2010 – 2016)

Nous avons bénéficié pour les années 2016 et 2017 d'une décharge d'enseignement de 45 heures au titre de la coordination du projet ANR ETHICAA. La table 7.5 de la page suivante résume l'ensemble de nos enseignements depuis notre prise de fonction à l'Université de Caen Normandie.

Intitulé	Responsabilité	Niveau	Année	CM	TD	TP
HTML & CSS		L1 Informatique	2010			36
Génie logiciel	×	L3 Informatique	2010	16,5	22,5	5
Réseaux & systèmes	×	LP Webmestre	2010	16,5		16,5
Intelligence artificielle distribuée		M1 Informatique	2010			18
Systèmes d'information		M2 AMI	2010	12		13
Agents mobiles	(partagée)	M2 LID	2010	2		
Projets & stages		L3, LP, M1, M2	2010			20
HTML & CSS		L1 Informatique	2011			36
Génie logiciel	×	L3 Informatique	2011	16,5	22,5	5
Réseaux & systèmes	×	LP Webmestre	2011	16,5		16,5
Intelligence artificielle distribuée	(partagée)	M1 Informatique	2011	4		12
Systèmes d'information		M2 AMI	2011	12		13
Agents mobiles	×	M2 LID	2011	12		
Projets & stages		L3, LP, M1, M2	2011			20
Génie logiciel	×	L3 Informatique	2012	16,5	22,5	5
Intro. à l'intelligence artificielle	×	L3 Informatique	2012	10	12	
Réseaux & systèmes	×	LP Webmestre	2012	16,5		16,5
Intelligence artificielle distribuée	×	M1 Informatique	2012	10	10	
Projets & stages		L3, LP, M1, M2	2012			50
Conception de logiciels	×	L2 Informatique	2013			26
Génie logiciel	(partagée)	L3 Informatique	2013	7,5	14	14
Intro. à l'intelligence artificielle	×	L3 Informatique	2013	10	12	
Réseaux & systèmes	×	LP Webmestre	2013	16,5		16,5
Intelligence artificielle distribuée	×	M1 Informatique	2013	10	10	
Agents et raisonnement		M2 DECIM	2013	2,5		
Projets & stages		L3, LP, M1, M2	2013			50
Conception de logiciels	×	L2 Informatique	2014			26
Génie logiciel	(partagée)	L3 Informatique	2014		13	13
Intro. à l'intelligence artificielle	×	L3 Informatique	2014	10	12	
Réseaux & systèmes	×	LP Webmestre	2014	16,5		16,5
Intelligence artificielle distribuée	×	M1 Informatique	2014	10	10	
Agents et raisonnement		M2 DECIM	2014	10		
Projets & stages		L3, LP, M1, M2	2014			50
Conception de logiciels	×	L2 Informatique	2015			26
Génie logiciel	(partagée)	L3 Informatique	2015		8,5	8,5
Intro. à l'intelligence artificielle	×	L3 Informatique	2015	10	12	
Réseaux & systèmes	×	LP Webmestre	2015	16,5		16,5
Intelligence artificielle distribuée	×	M1 Informatique	2015	10	10	
Agents et raisonnement		M2 DECIM	2015	12,5		
Projets & stages		L3, LP, M1, M2	2015			50
Conception de logiciels	×	L2 Informatique	2016			26
Génie logiciel		L3 Informatique	2016		8,5	8,5
Intro. à l'intelligence artificielle	×	L3 Informatique	2016	10	12	
Réseaux & systèmes	×	LP Webmestre	2016	16,5		16,5
Intelligence artificielle distribuée	×	M1 Informatique	2016	10	10	
Projets & stages		L3, LP, M1, M2	2016			50
Conception de logiciels	×	L1 Informatique	2017	10		40
Conception de logiciels	×	L2 Informatique	2017	8		30
Sécurité et IA	(partagée)	L2 Informatique	2017	2,5		32
Aide à la décision et IA	(partagée)	L3 Informatique	2017	5		7,5
Intelligence artificielle distribuée	×	M1 Informatique	2017	10		
Projets & stages		L3, LP, M1, M2	2017			30

TABLE 7.5 – Récapitulatif de nos enseignements

Bibliographie

A. Abdul-Rahman et S. Hailes (2000). Supporting trust in virtual communities. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, HICSS 2000*, page 10.

D. Abel, J. MacGlashan et M. Littman (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Workshops : AI, Ethics and Society*.

D. Abramson et L. Pike (2011). When formal systems kill : Computer ethics and formal methods. *APA Newsletter on Philosophy and Computers*, 11(1).

S. Aknine, S. Pinson et M. F. Shakun (2004). A multi-agent coalition formation method based on preference models. *Group Decision and Negotiation*, 13(6) :513–538.

H. Aldewereld, V. Dignum et Y. hua Tan (2015). *Handbook of Ethics, Values, and Technological Design*, chapter Design for values in software development. Springer-Verlag.

N. Alechina, B. Logan et M. Whitsey (2004). A complete and decidable logic for resource-bounded agents. In *Autonomous Agents and Multi-Agent Systems (AAMAS'04)*.

T. Alpcan et T. Başar (2010). *Network security : A decision and game-theoretic approach*. Cambridge University Press.

V. Anantharam, P. Varaiya et J. Walrand (1987). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. *IEEE Automatic Control*, 32(11) :968–976.

M. Anderson et S. Anderson (2014). Toward ensuring ethical behavior from autonomous systems : a case-supported principle-based paradigm. *Industrial Robot*, 42(4) :324–331.

M. Anderson, S. Anderson et C. Armen (2006). An approach to computing ethics. *IEEE Intelligent Systems*, 21(4) :56–63.

M. Anderson, S. L. Anderson et V. Berenz (2017). A value driven agent : Instantiation of a case-supported principle-based behavior paradigm. In *Workshops at the 31st AAAI Conference on Artificial Intelligence*.

P. Angin, B. Bhargava, R. Ranchal, N. Singh, M. Linderman, L. B. Othmane et L. Lilien (2010). An entity-centric approach for privacy and identity management in cloud computing. In *Proceedings of the 29th Symposium on Reliable Distributed Systems*, pages 177–183.

L. Antunes et H. Coelho (1999). Decisions based upon multiple values : the bvg agent architecture. In *Portuguese Conference on Artificial Intelligence*, pages 297–311.

R. Arkin (2009). *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall.

K. J. Arrow (1963). *Social Choice and Individual Values*. Yale University Press.

P. Aschwanden, V. Baskaran, S. Bernardini, C. Fry, M. Moreno, N. Muscettola, C. Plaunt, D. Rijsman et P. Tompkins (2006). Model-unified planning and execution for distributed autonomous system control. In *AAAI 2006 Fall Symposium*, pages 1–10.

K. Atkinson et T. Bench-Capon (2008). Addressing moral problems through practical reasoning. *Journal of Applied Logic*, 6(2) :135–151.

K. Atkinson et T. Bench-Capon (2016). Value based reasoning and the actions of others. In *22th European Conference on Artificial Intelligence*, pages 52–57.

P. Auer, N. Cesa-Bianchi et P. Fischer (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3) :235–256.

P. Auer, N. Cesa-Bianchi, Y. Freund et R. Schapire (1995). Gambling in a rigged casino : the adversarial multi-armed bandit problem. In *36th Annual Symposium on Foundations of Computer Science*, pages 322–331.

P. Auer et R. Ortner (2010). UCB revisited : Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2) :55–65.

R. J. Aumann (1985). On the non-transferable utility value : A comment on the roth-shafer examples. *Econometrica : Journal of the Econometric Society*, 53 :667–677.

R. J. Aumann et J. H. Dreze (1974). Cooperative games with coalition structures. *International Journal of game theory*, 3(4) :217–237.

B. Awerbuch et R. Kleinberg (2008). Competitive collaborative learning. *Journal of Computer and System Sciences*, 74(8) :1271–1288.

H. Aziz, F. Brandt et P. Harrenstein (2013a). Fractional hedonic games. In *12th International Joint Conference on Autonomous Agents and Multi-Agent Systems*.

H. Aziz, F. Brandt et P. Harrenstein (2013b). Pareto optimality in coalition formation. *Games and Economic Behavior*, 82 :562–581.

H. Aziz, F. Brandt et P. Harrenstein (2014). Fractional hedonic games. In *Proceedings of the 13th international conference on Autonomous agents and multi-agent systems, AAMAS 2014*, pages 5–12.

H. Aziz, F. Brandt et H. Seedig (2013c). Computing desirable partitions in additively separable hedonic games. *Artificial Intelligence*, 195 :316–334.

H. Aziz, F. Brandt et H. G. Seedig (2011). Stable partitions in additively separable hedonic games. In *The 10th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2011*, pages 183–190.

H. Aziz et M. Paterson (2009). False name manipulations in weighted voting games : splitting, merging and annexation. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2009*, pages 409–416.

t. M.-A. L. B. Williams (1990). *Éthique et les limites de la philosophie*. Gallimard.

Y. Bachrach et E. Elkind (2008). Divide and conquer : False-name manipulations in weighted voting games. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems, AAMAS 2008*, pages 975–982.

Y. Bachrach et J. Rosenschein (2008). Coalitional skill games. In *7th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1023–1030.

M. Baldoni, C. Baroglio, I. Gungui, A. Martelli, M. Martelli, V. M. nd V. Patti et C. Schifanella (2005). Reasoning About Agents' Interaction Protocols Inside DCaseLP. In *Declarative Agent Languages and Technologies II*, pages 112–131.

C. Ballester (2004). NP-completeness in hedonic games. *Games and Economic Behavior*, 49(1) :1–30.

J. Banzhaff III (1964). Weighted voting doesn't work : A mathematical analysis. *Rutgers University Law Review*, 19.

C. D. Batson (2014). *The altruism question : Toward a social-psychological answer*. Psychology Press.

C. Battaglino, R. Damiano et L. Lesmo (2013). Emotional range in value-sensitive deliberation. In *12th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 769–776.

G. Beavers et H. Hexmoor (2003). Types and limits of agent autonomy. In *International Workshop on Computational Autonomy*, pages 95–102.

G. Bekey (2005). *Autonomous robots : from biological inspiration to implementation and control*. MIT Press.

A. Belloni, A. Berger, O. Boissier, G. Bonnet, G. Bourgne, P.-A. Chardel, J.-P. Cotton, N. Evreux, J.-G. Ganascia, P. Jaillon, B. Mermet, G. Picard, B. Rever, G. Simon, T. de Swarte, C. Tessier, F. Vexler, R. Voyer et A. Zimmermann (2015a). Dealing with ethical conflicts in autonomous agents and multi-agent systems. In *1st International Workshop Joint on Artificial Intelligence and Ethics*.

A. Belloni, A. Berger, O. Boissier, G. Bonnet, G. Bourgne, P.-A. Chardel, J.-P. Cotton, N. Evreux, J.-G. Ganascia, P. Jaillon, B. Mermet, G. Picard, B. Rever, G. Simon, T. de Swarte, C. Tessier, F. Vexler, R. Voyer et A. Zimmermann (2015b). Towards a framework to deal with ethical conflicts in autonomous agents and multi-agent systems. In *12th International Conference on Computer Ethics and Philosophical Enquiry*.

T. Bench-Capon et K. Atkinson (2009). Abstract argumentation and values. In G. Simari et I. Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 45–64. Springer.

P. Bernoux (1985). *La sociologie des organisations*. Seuil.

F. Berreby, G. Bourgne et J.-G. Ganascia (2015). Modelling moral reasoning and ethical responsibility with logic programming. In *20th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning*, pages 532–548.

L. Bilge, T. Strufe, D. Balzarotti et E. Kirda (2009). All your contacts are belong to us : automated identity theft attacks on social networks. In *Proceedings of the 18th international conference on World wide web*, pages 551–560.

F. Bloch (1997). *New Directions in the Economic Theory of the Environment*, volume 25, chapter 10 Non-cooperative models of coalition formation in games with spillovers, page 311. Cambridge University Press (Cambridge, UK).

A. Bogomolnaia et M. Jackson (2002). The stability of hedonic coalition structures. *Games and Economic Behavior*, 38(2) :201–230.

O. Boissier, F. Balbo et F. Badeig (2010). Controlling multi-party interaction within normative multi-agent organizations. In *3rd Federated Workshops on Multi-Agent Logics, Languages, and Organisations*. CEUR Proceedings Vol. 627.

O. Boissier, G. Bonnet, N. Cointe, B. Mermet, G. Simon, C. Tessier et T. de Swarte (2017). Models for ethical autonomous agents. Technical report, ANR ETHICAA.

O. Boissier, J. Hübner et A. Ricci (2016). The JaCaMo framework. In *Social coordination frameworks for social technical systems*, pages 125–151. Springer.

G. Bonnet (2012). A protocol based on a game-theoretic dilemma to prevent malicious coalitions in reputation systems. In *20th European Conference on Artificial Intelligence*, pages 187–191.

G. Bonnet (2013). Un protocole fondé sur un dilemme pour se prémunir des collusions dans les systèmes de réputation. In *21es Journées Francophones sur les Systèmes Multi-Agents*, pages 9–18.

G. Bonnet (2014). Un protocole fondé sur des dilemmes pour se prémunir des collusions dans les systèmes de réputation. *Revue d'Intelligence Artificielle*, 28(4) :411–431.

G. Bonnet, B. Mermet et G. Simon (2016). Vérification formelle et éthique dans les sma. In *24es Journées Francophones sur les Systèmes Multi-Agents*, pages 139–148.

R. Bordini, M. Fisher, W. Visser et M. Wooldridge (2003). Verifiable multi-agent programs. In M. Dastani, J. Dix et A. Seghrouchni, editors, *ProMAS*.

R. Bordini, J. Hübner et M. Wooldridge (2007). *Programming Multi-Agent Systems in AgentSpeak Using **Jason***. John Wiley & Sons.

N. Borisov (2006). Computational puzzles as sybil defenses. In *Proceedings of the 6th IEEE International Conference on Peer-to-Peer Computing*, pages 171–176.

A. Bracciali, U. Endriss, N. Demetriou, T. Kakas, W. Lu et K. Stathis (2006). Crafting the mind of PROSOCS agents. *Applied Artificial Intelligence*, 20(2–4) :105–131.

R. Braithwaite (1955). *Theory of games as a tool for the moral philosopher*. Cambridge University Press.

M. Bramer, M. Bramer et M. Bramer (2007). *Principles of data mining*, volume 131. Springer.

F. Brandl, F. Brandt et M. Strobel (2015). Fractional hedonic games : Individual and group stability. In *14th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1219–1227.

M. Bratman (1987). *Intention, plans, and practical reason*. Harvard University Press.

M. Bratman (1990). What is intention ? In P. Cohen, J. Morgan et M. Pollack, editors, *Intentions in Communication*. MIT Press.

S. Bringsjord et J. Taylors (2012). Introducing divine-command robot ethics. In P. Lin, G. Bekey et K. Abney, editors, *Robot ethics : the ethical and social implication of robotics*, pages 85–108. MIT Press.

J. Broersen, M. Dastani, Z. Huang, J. Hulstijn et L. Van der Torre (2001). The BOID architecture : Conflicts between beliefs, obligations, intentions and desires. In *5th International Conference on Autonomous Agents*, pages 9–16.

C. Brooks et E. Durfee (2003). Congregation formation in multiagent systems. *Journal of Autonomous Agents and Multiagent Systems*, 7 :145–170.

R. Brooks (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1) :14–23.

R. Brooks (1991). Intelligence without representation. *Artificial Intelligence*, 47(1–3) :139–159.

P. Buzing, A. Eiben et M. Schut (2005). Emerging communication and cooperation in evolving agent societies. *Journal of Artificial Societies and Social Simulation*, 8(1) :27–52.

P. Caire (2009). Designing convivial digital cities : A social intelligence design approach. *AI and Society Journal*, 24(1) :97–114.

C. Carabelea, O. Boissier et A. Florea (2003). Autonomy in multi-agent systems : A classification attempt. *Lecture Notes in Computer Science*, 2969 :103–113.

C. Castelfranchi (1998). Modeling social action for AI agents. *Artificial Intelligence*, 103 :157–182.

C. Castelfranchi et R. Falcone (2003). From automaticity to autonomy : the frontier of artificial agents. In H. Hexmoor, C. Castelfranchi et R. Falcone, editors, *Agent autonomy*, pages 103–136. Kluwer Academic Publishers.

C. Castelfranchi et R. Falcone (2010). *Trust theory : A socio-cognitive and computational model*. John Wiley & Sons.

CERNA (2014). éthique de la recherche en robotique. Technical report, AllistÃ¨ne.

CERNA (2017). éthique de la recherche en apprentissage artificiel. Technical report, AllistÃ¨ne.

B. Chae, D. Paradice, J.-F. Courtney et C.-J. Cagler (2005). Incorporating an ethical perspective to problem formulation : Implications for decision support system design. *Decision Support Systems*, 40 :197–212.

G. Chalkiadakis, E. Elkind, E. Markakis, M. Polukarov et N. Jenning (2010). Cooperative games with overlapping coalitions. *Journal of Artificial Intelligence Research*, 39 :179–216.

G. Chalkiadakis, E. Markakis et C. Boutilier (2007). Coalition formation under uncertainty : Bargaining equilibria and the bayesian core stability concept. In *6th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 64–72.

N. Chamfort (1857). *Maximes, Pensées, Anecdotes, Caractères et Dialogues*. Durr.

R. K. Chang (2002). Defending against flooding-based distributed denial-of-service attacks : a tutorial. *Communications Magazine, IEEE*, 40(10) :42–51.

S. Chatterjee, S. Sarker et M. Fuller (2009). A deontological approach to designing ethical collaboration. *Journal of the Association for Information Systems*, 10 :138–169.

B. Chellas (1980). *Modal logic : an introduction*. Cambridge University Press.

A. Cheng et E. Friedman (2005a). Sybilproof reputation mechanisms. In *Proceedings of the 3rd Workshop on Economics of Peer-to-Peer Systems*, pages 128–132.

A. Cheng et E. Friedman (2005b). Sybilproof reputation mechanisms. In *3rd Workshop on Economics of Peer-to-Peer Systems*, pages 128–132.

A. Cheng et E. Friedman (2006). Manipulability of pagerank under sybil strategies.

R. Chisholm (1963). Contrary-to-duty imperatives and deontic logic. *Analysis*, 24 :33–36.

S. Chopra et L.-F. White (2011). A legal theory for autonomous artificial agents. Technical report, University of Michigan.

B. Christianson et W. Harbison (1997). Why isn't trust transitive ? In *Security protocols*, pages 171–176.

H. Coelho et A. da Rocha Costa (2009). On the intelligence of moral agency. In *Encontro Português de Inteligência Artificial*, pages 12–15.

H. Coelho, P. Trigo et A. da Rocha Costa (2010). On the operationality of moral-sense decision making. In *2nd Brazilian Workshop on Social Simulation*, pages 15–20.

P. Cohen et H. Levesque (1990). Intention is choice with commitment. *Artificial Intelligence*, 42(3) :213–261.

N. Cointe (2017). *Jugement éthique pour la décision et la coopération dans les systèmes multi-agents*. Thèse de doctorat, Université de Lyon.

N. Cointe, G. Bonnet et O. Boisser (2016a). Jugement éthique dans les systèmes multi-agents. In *24es Journées Francophones sur les Systèmes Multi-Agents*, pages 149–158.

N. Cointe, G. Bonnet et O. Boisser (2016b). Multi-agent based ethical asset management. In *1st Workshop on Ethics in the Design of Intelligent Agents*, pages 52–57.

N. Cointe, G. Bonnet et O. Boissier (2016c). Ethical judgment of agents' behaviors in multi-agent systems. In *15th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1106–1114.

N. Cointe, G. Bonnet et O. Boissier (2017a). Coopération fondée sur l'éthique entre agents autonomes. In *25es Journées Francophones sur les Systèmes Multi-Agents*, pages 9–18.

N. Cointe, G. Bonnet et O. Boissier (2017b). éthique collective dans les systèmes multi-agents. *Revue d'Intelligence Artificielle*, 31(1-2) :71–96.

N. Cointe, G. Bonnet et O. Boissier (2017c). Jugement éthique dans le processus de décision d'un agent BDI. *Revue d'Intelligence Artificielle*, 31(4) :471–499.

N. Cointe, G. Bonnet et O. Boissier (2018). Ethics-based cooperation in multi-agent systems. In *14th Annual Conference of the European Social Simulation Association*.

A. Comte (1852). *Catéchisme positiviste*. P. Arnaud, Paris.

A. Comte-Sponville (2012). *La philosophie*. PUF.

V. Conitzer, J. Lang et T. Sandholm (2003). How many candidates are needed to make elections hard to manipulate ? In *Proceedings of the 9th conference on Theoretical aspects of rationality and knowledge*, pages 201–214.

V. Conitzer et T. Sandholm (2006). Complexity of constructing solutions in the core based on synergies among coalitions. *Artificial Intelligence*, 170(6) :607–619.

B. Constant et E. Kant (2003). *Le droit de mentir*. Mille et une nuits.

G. Danezis et P. Mittal (2009). Sybilinfer : Detecting sybil nodes using social networks. In *Proceedings of the Network and Distributed System Security, NDSS2009*.

V. D. Dang et N. R. Jennings (2004). Generating coalition structures with finite bound from the optimal guarantees. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2004*, pages 564–571.

M. Dastani (2008). 2apl : a practical agent programming language. *Autonomous Agents and Multi-Agent Systems*, 16(3) :214–248.

M. Dastani, A. Herzig, J. Hulstijn et L. Van Der Torre (2004). Inferring trust. In *5th International Workshop on Computational Logic in Multi-Agent Systems*, pages 144–160.

DDHC (1789). Déclaration des Droits de l'Homme et du Citoyen de 1789 - Article 4.

G. De Clippel et R. Serrano (2005). Marginal contributions and externalities in the value. Technical report, Brown University.

C.-L. de Montesquieu (1867). *L'esprit des lois*. Libr. de F. Didot Frères.

Defense Science Board (2012). The role of autonomy in DoD systems. Technical report, Department of Defense.

G. Delannoi et O. Dowlen (2010). *Sortition, Theory and Practice*. Academic UK and USA.

V. Demiaux et Y. Si Abdallah (2017). Comment permettre Ã l'Homme de garder la main ? Technical report, CNIL.

R. Demolombe (2004). Reasoning about trust : A formal logical framework. In *2nd International Conference on Trust Management*, pages 291–303.

R. Demolombe et C. Liau (2001). A logic of graded trust and belief fusion. In *4th Workshop on Deception, Fraud and Trust in Agent Societies*, pages 13–25.

C. Dennett (1987). *The Intentional Stance*. The MIT Press.

D. Dennett (1971). Intentional systems. *The Journal of Philosophy*, 68(4) :87–106.

L. Dennis, M. Fisher et A. Winfield (2015). Towards verifiably ethical robot behaviour. In *1th International Workshop on AI and Ethics*.

B. Docherty (2012). Losing humanity - The case against killer robots. Technical report, Human Rights Watch.

G. Dorais, P. Bonasso, D. Kortenkamp, B. Pell et D. Schreckenghost (1999). Adjustable autonomy for human-centered autonomous systems. In *Workshop on Adjustable Autonomy Systems*.

J. R. Douceur (2002). The sybil attack. In *Peer-to-peer Systems*, pages 251–260. Springer.

J. Dreze et J. Greenberg (1980). Hedonic coalitions : Optimality and stability. *Econometrica*, 48(4) :987–1003.

J. H. Drèze et J. Greenberg (1980). Hedonic coalitions : Optimality and stability. *Econometrica*, 48(4) :987–1003.

T. Driessen (1991). A survey of consistency properties in cooperative game theory. *SIAM Review*, 33(1) :43–59.

A. Drogoul (1995). When ants play chess (or can strategies emerge form tactical behaviors ?). *Artificial Intelligence*, 957 :13–27.

B. Dundua et L. Uridia (2010). Trust and belief, interrelation. In *3rd Workshop on Agreement Technologies*.

E. Durfee (2001). Scaling up agent coordination strategies. *IEEE Computer*, 34(7) :39–46.

E. Durkheim (1893). *De la division du travail social : étude sur l'organisation des sociétés supérieures*. F. Alcan.

E. Durkheim (1897). *Le suicide : étude de sociologie*. F. Alcan.

E. Elkind et M. Wooldridge (2009). Hedonic coalition nets. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2009*, pages 417–424.

R. J. Ellison, D. A. Fisher, R. C. Linger, H. F. Lipson et T. Longstaff (1997). Survivable network systems : An emerging discipline. Technical report, DTIC Document.

R. Falcone, G. Pezzulo et C. Castelfranchi (2002). A fuzzy approach to a belief-based trust computation. In *5th Workshop on Deception, Fraud and Trust in Agent Societies*, pages 73–86.

M. Feldman, C. Papadimitriou, J. Chuang et I. Stoica (2006). Free-riding and whitewashing in peer-to-peer systems. *Selected Areas in Communications, IEEE Journal on*, 24(5) :1010–1019.

J. Ferber (1999). *Multi-agent systems - an introduction to distributed artificial intelligence*. Addison-Wesley-Longman.

I. Ferguson (1992). *Touring Machines : An architecture for dynamic, rational, mobile agents*. Thèse de doctorat, University of Cambridge.

K. Fischer, M. Schillo et J. Siekmann (2003). Holonic multiagent systems : A foundation for the organisation of multiagent systems. In *1st International Conference on Applications of Holonic and Multi-Agent Systems*, pages 71–80.

L. Floridi et J. Sanders (2004). On the morality of artificial agents. *Minds and Machines*, 14(3) :349–379.

S. Franklin et A. Graesser (1996). Is it an agent or just a program ? A taxonomy for autonomous agents. *Lecture Notes In Computer Science*, 1193 :21–35.

B. Friedman (1996). Value-sensitive design. *Interactions*, 3(6) :16–23.

B. Friedman, P. Kahn, A. Borning et A. Huldtgren (2013). Value sensitive design and information systems. In *Early engagement and new technologies : Opening up the laboratory*, pages 55–95. Springer Netherlands.

M. Frize, L. Yang, R. Walker et A. O'Connor (2005). Conceptual framework of knowledge management for ethical decision-making support in neonatal intensive care. *IEEE Transactions on Information Technology in Biomedicine*, 9(2) :205–215.

W. A. Gamson (1961). A theory of coalition formation. *American Sociological Review*, 26(3) :373–382.

J. Ganascia (2007). Modeling ethical rules of lying with answer set programming. *Ethics and Information Technology*, 9 :39–47.

J. Ganascia (2012). An agent-based formalization for resolving ethical conflicts. In *1st Workshop on Belief change, Nonmonotonic reasoning, and Conflict resolution*.

P. Gärdenfors (1976). Manipulation of social choice functions. *Journal of Economic Theory*, 13(2) :217–228.

L. Gasser (2001). Organizations in multi-agent systems. In *10th European Worshop on Modeling Autonomous Agents in a Multi-Agent World*.

T. Génin (2010). *Stratégies de formation de coalitions dans les systèmes multi-agents*. Thèse de doctorat, Paris 6.

T. Génin et S. Aknine (2011). Étude de protocoles et de stratégies de négociation pour l'obtention de structures de coalitions pareto optimales dans un problème de formation de coalitions. In *Sixièmes Journées Francophones Modèles formels de l'interaction, MFI 2011*, pages 187–196.

H. Gensler (1996). *Formal ethics*. Routledge.

G. D. Giacomo, Y. Lesperance et H. J. Levesque (2000). Congolog, a concurrent programming language based on the situation calculus. *Artificial Intelligence*, 121(1–2) :109–169.

A. Gibbard (1973). Manipulation of voting schemes : a general result. *Econometrica : journal of the Econometric Society*, 41(4) :587–601.

J. Golbeck (2006). Computing with trust : Definition, properties, and algorithms. In *Proceedings of the Securecomm and Workshops, 2006*, pages 1–7.

M. Grabisch et Y. Funaki (2012). A coalition formation value for games in partition function form. *European Journal of Operational Research*, 221(1) :175–185.

T. Grandison et M. Sloman (2000). A survey of trust in internet applications. *Communications Surveys & Tutorials, IEEE*, 3(4) :2–16.

J. Greenberg (1994). Coalition structures. *Handbook of game theory with economic applications*, 2 :1305–1337.

M. Guerini et O. Stock (2005). Toward ethical persuasive agents. In *IJCAI Workshop on Computational Models of Natural Arguments*.

J. Haidt (2001). The emotional dog and its rational tail : a social intuitionist approach to moral judgment. *Psychological Review*, 108(4) :814–834.

J. Hajduková *et al.* (2003). Computational complexity of stable partitions with b-preferences. *International Journal of Game Theory*, 31(3) :353–364.

J. Hajduková *et al.* (2004). Stable partitions with w-preferences. *Discrete Applied Mathematics*, 138(3) :333–347.

B. Hardin et M. Goodrich (2009). On using mixed-initiative control : a perspective for managing large-scale robotic teams. In *4th ACM/IEEE International Conference on Human-Robot Interaction*, pages 165–172.

J. C. Harsanyi (1963). A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2) :194–220.

S. Hart et M. Kurz (1983). Endogenous formation of coalitions. *Econometrica : Journal of the Econometric Society*, 51 :1047–1064.

L. Henkin (1949). The completeness of the first-order functional calculus. *Journal of Symbolic Logic*, 14 :159–166.

A. Herzig, E. Lorini, J. Hübner et L. Vercouter (2010). A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1) :214–244.

A. Herzig, E. Lorini, L. Perrussel et Z. Xiao (2016). Bdi logics for bdi architectures : old problems, new perspectives. *KI-Künstliche Intelligenz*, pages 1–11.

S. Hitlin et J. Piliavin (2004). Values : Reviving a dormant concept. *Annual Review of Sociology*, 30 :359–393.

J. Hoc (2000). From human-machine interaction to human-machine cooperation. *Ergonomics*, 43(7) :833–843.

M. Hoefer, D. Váz et L. Wagner (2014). Hedonic coalition formation in networks. In *Proceedings of the 28th Conference on Artificial intelligence, AAAI 2014*.

K. Hoffman, D. Zage et C. Nita-Rotaru (2009). A survey of attack and defense techniques for reputation systems. *ACM Computer Survey*, 42(1) :1–31.

A. Honarvar et N. Ghasem-Aghaee (2009). An artificial neural network approach for creating an ethical artificial agent. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 290–295.

B. Horling et V. Lesser (2004). A survey of multi-agent organizational paradigms. *American Society for Information Science and Technology*, 55(9) :783–793.

H. Horsburgh (1960). The ethics of trust. *The Philosophical Quartely*, 10(41) :343–354.

N. Howden, R. Rönnquist, A. Hodgson et A. Lucas (2001). Jack intelligent agents-summary of an agent infrastructure. In *5th International Conference on Autonomous Agents*.

J. Hubner, J. Sichman et O. Boissier (2002). Moise+ : Towards a structural, functional, and deontic model for the MAS organization. In *1st International Joint Conference on Autonomous Agents and Multi-Agent Systems*.

IEEE (2017). Ethically aligned design. Technical report, Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems.

S. Ieong et Y. Shoham (2008). Bayesian coalitional games. In *23rd AAAI Conference on Artificial Intelligence*, pages 95–100.

Y. Jin, Z. Gu et Z. Ban (2007). Restraining false feedbacks in peer-to-peer reputation systems. In *Proceedings of the Semantic International Conference on Computing, ICSC 2007*, pages 304–312.

J. Jones (2008). ALFUS - Autonomy Levels For Unmanned Systems Framework. Technical report, ALFUS Working Group.

A. Jøsang et R. Ismail (2002). The Beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, pages 41–55.

A. Josang, R. Ismail et C. Boyd (2007). A survey of trust and reputation systems for online service proposition. *Decision Support Systems*, 43(2) :618–644.

A. Jøsang, R. Ismail et C. Boyd (2007). A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2) :618–644.

M. Kacprzak, A. Lomuscio et W. Penczek (2004). Verification of multiagent systems via unbounded model checking. In *Autonomous Agents and Multi-Agent Systems (AAMAS'04)*.

L. Kaelbling, M. Littman et A. Cassandra (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2) :99–134.

S. D. Kamvar, M. T. Schlosser et H. Garcia-Molina (2003). The EigenTrust algorithm for reputation management in P2P networks. In *Proceedings of the 12th International World Wide Web Conference*, pages 640–651.

V. Kant et K. Bharadwaj (2013). Fuzzy computational models of trust and distrust for enhanced recommendations. *International Journal of Intelligent Systems*, 28(4) :332–365.

H. Keinänen (2010). An algorithm for generating nash stable coalition structures in hedonic games. In *Foundations of Information and Knowledge Systems*, pages 25–39. Springer.

A. S. Kelso Jr et V. P. Crawford (1982). Job matching, coalition formation, and gross substitutes. *Econometrica : Journal of the Econometric Society*, 50(05) :1483–1504.

K. Kim et H. Lipson (2009). Towards a 'theory of mind' in simulated robots. In *11th Annual Conference Companion on Genetic and Evolutionary Computation Conference*, pages 2071–2076.

B.-J. Koops et R. Leenes (2006). Identity theft, identity fraud and/or identity-related crime. *Datenschutz und Datensicherheit-DuD*, 30(9) :553–556.

D. E. Koulouriotis et A. Xanthopoulos (2008). Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation*, 196(2) :913–922.

E. Koutrouli et A. Tsalgatidou (2011). Credibility enhanced reputation mechanism for distributed e-communities. In *Proceedings of the 19th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, PDP 2011*, pages 627–634.

S. Kullback (1997). *Information theory and statistics*. Courier Dover Publications.

K. S. Larson et T. W. Sandholm (2000). Anytime coalition structure generation : an average case study. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1) :23–42.

M. Lemaître et G. Verfaillie (2007). Interaction between reactive and deliberative tasks for on-line decision-making. In *ICAPS'07 Workshop on Planning and Plan Execution for Real-World Systems*.

C. Lemaître et C. Excelente (1998). Multi-agent organization approach. In *2nd Iberoamerican Workshop on Distributed Artificial Intelligence and Multi-Agent Systems*.

C. Leturc et G. Bonnet (2017). Une logique modale normale de la confiance. In *Journées Intelligence Artificielle Fondamentale*.

C. Leturc et G. Bonnet (2018a). A normal modal logic for trust in the sincerity. In *17th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 175–183.

C. Leturc et G. Bonnet (2018b). Une logique modale pour la caractÃ©risation des manipulations entre agents autonomes. In *Journées Intelligence Artificielle Fondamentale*.

C.-J. Liau (2003). Belief, information acquisition, and trust in multi-agent systems : A modal logic formulation. *Artificial Intelligence*, 149(1) :31–60.

K. Liu et Q. Zhao (2010). Distributed learning in multi-armed bandit with multiple players. *Signal Processing, IEEE Transactions on*, 58(11) :5667–5681.

E. Lorini (2012). On the logical foundations of moral agency. In *11th International Conference on Deontic Logic in Computer Science*, pages 108–122.

Z. Malik et A. Bouguettaya (2009). Rateweb : Reputation assessment for trust establishment among web services. *International Journal on Very Large Data Bases*, 18(4) :885–911.

S. Marsh (1994). *Formalising Trust as a Computational Concept*. Thèse de doctorat, University of Stirling.

M. Martelli, V. Mascardi et F. Zini (1997). CaseLP : a Complex Application Specification Environment base on Logic Programming. In *Proc. of ICLP'97 workshop on Logc Programming and Mult-Agents*, pages 35–50.

S. Marti et H. Garcia-Molina (2006). Taxonomy of trust : Categorizing p2p reputation systems. *Computer Networks*, 50(4) :472–484.

K. Mathieson (2007). Dioptra : An ethics decision support system. In *13th Americas Conference on Information Systems*.

R. D. McKelvey, P. C. Ordeshook et M. D. Winer (1978). The competitive solution for n-person games without transferable utility, with an application to committee games. *American Political Science Review*, 72(02) :599–615.

B. McLaren (2003). Extensionally defining principles and cases in ethics : An AI model. *Artificial Intelligence*, 150(1–2) :145–181.

S. Mercier (2011). *Contrôle du partage de l'autorité dans un système d'agents hétérogènes*. Thèse de doctorat, Institut Supérieur de l'Aéronautique et de l'Espace, Toulouse.

B. Mermet et G. Simon (2009). GDT4MAS : an extension of the GDT model to specify and to verify multi-agent systems. In *8th International Conference on Autonomous Agents and Multiagent Systems*, pages 505–512.

J.-J. Meyer, J. Broersen et A. Herzig (2015). Bdi logics. Technical report, College Publications.

T. Michalak, T. Rahwan, J. Sroka, A. Dowell, M. Wooldridge, P. McBurney et N. Jennings (2009). On representing coalitional games with externalities. In *10th ACM Conference on Electronic Commerce*, pages 11–20.

T. Michalak, J. Sroka, T. Rahwan, M. Wooldridge, P. McBurney et N. R. Jennings (2010). A distributed algorithm for anytime coalition structure generation. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2010*, pages 1007–1014.

J. Mill (1869). *On liberty*. Longmans, Green, Reader, and Dyer.

J. Mill (1889). *L'utilitarisme*. Alcan.

O. Morgenstern et J. Von Neumann (1953). *Theory of games and economic behavior*. Princeton University Press.

L. Mui, M. Mohtashemi et A. Halberstadt (2002). A computational model of trust and reputation. In *Proceedings of the 35th Annual Hawaii International Conference on the System Sciences, HICSS 2002*, pages 2431–2439.

G. Muller et L. Vercouter (2004). Détection décentralisée d'agents menteurs (article court). In *Douxièmes journées francophones sur les systèmes multi-agents, JFSMA 04*, pages 243–248.

G. Muller et L. Vercouter (2005). Decentralized monitoring of agent communications with a reputation model. In *Trusting Agents for Trusting Electronic Societies*, pages 144–161. Springer.

J. Muller et M. Pischel (1993). The agent architecture InteRRap : Concept and application. Technical report, Deutsches Forschungszentrum fur Kunstliche Intelligenz.

J. F. Nash (1950). Equilibrium points in n-person games. *National Academy of Sciences of the United States of America*, 36(1) :48–49.

A. Nongaillard et P. Mathieu (2011). Reallocation problems in agent societies : a local mechanism to maximize social welfare. *Journal of Artificial Societies and Social Simulation*, 14(3) :5.

A. Nowak et T. Radzik (1994). A solidarity value for n-person transferable utility games. *International Journal of Game Theory*, 23 :43–48.

H. S. Nwana, L. C. Lee et N. R. Jennings (1996). Coordination in software agent systems. *The British Telecom Technical Journal*, 14(4) :79–88.

S. of Professional Journalists (2014). Code of ethics.

A. Okada (1996). A noncooperative coalitional bargaining game with random proposers. *Games and Economic Behavior*, 16(1) :97–108.

M. Okada, K. Yamamoto et K. Watanabe (2007). Conceptual model of health information ethics as a basis for computer-based instructions for electronic patient record systems. *Studies in Health Technology and Informatics*, 129 :1442–1446.

L. Orseau et S. Armstrong (2016). Safely interruptible agents. In *32nd Conference on Uncertainty in Artificial Intelligence*, pages 557–566.

L. Page, S. Brin, R. Motwani et T. Winograd (1999). The PageRank citation ranking : bringing order to the Web. Technical report, Stanford InfoLab.

D. Peters et E. Elkind (2015). Simple causes of complexity in hedonic games. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 617–623.

T. Powers (2005). Deontological machine ethics. Technical report, American Association for Artificial Intelligence.

T. Rahwan (2007). *Algorithms for coalition formation in multi-agent systems*. Thèse de doctorat, University of Southampton.

T. Rahwan et N. R. Jennings (2007). An algorithm for distributing coalitional value calculations among cooperating agents. *Artificial Intelligence*, 171(8) :535–567.

T. Rahwan et N. R. Jennings (2008). Coalition structure generation : Dynamic programming meets anytime optimization. In *Proceedings of the 23rd Conference on Artificial intelligence, AAAI 2008*, volume 8, pages 156–161.

T. Rahwan, T. Michalak, M. Wooldridge et N. Jennings (2015). Coalition structure generation : A survey. *Artificial Intelligence*, 229 :139–174.

T. Rahwan, S. D. Ramchurn, V. D. Dang, A. Giovannucci et N. R. Jennings (2007). Anytime optimal coalition structure generation. In *Proceedings of the 22nd Conference on Artificial intelligence AAAI 2007*, volume 7, pages 1184–1190.

F. Raimondi et A. Lomuscio (2004). Verification of multiagent systems via orderd binary decision diagrams : an algorithm and its implementation. In *Autonomous Agents and Multi-Agent Systems (AAMAS'04)*.

A. Rand (1964). *The virtue of selfishness*. Penguin.

A. Rand (2005). *The fountainhead*. Penguin.

A. Rao (1996). Agentspeak(l) : Bdi agents speak out in a logical computable language. In *7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, pages 42–55.

A. Rao et M. Georgeff (1991). Modeling rational agents within a BDI-architecture. In *2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484.

D. Ray et R. Vohra (1999). A theory of endogenous coalition structures. *Games and Economic Behavior*, 26(2) :286–336.

P. Resnick, K. Kuwabara, R. Zeckhauser et E. Friedman (2000). Reputation systems. *ACM Communications*, 43(12) :45–48.

P. Ricoeur (1995). *Oneself as another*. University of Chicago Press.

M. Ripeanu (2001). Peer-to-peer architecture case study : Gnutella network. In *1st International Conference on Peer-to-Peer Computing*, pages 99–100.

H. Robbins (1952). Some aspects of the sequential design of experiments. *Journal of the AMS*, 58(5) :527–535.

R.-W. Robbins et W.-A. Wallace (2007). Decision support for ethical problem solving : A multi-agent approach. *Decision Support Systems*, 43(4) :1571–1587.

M. S. Robinson (1985). Collusion and the choice of auction. *The RAND Journal of Economics*, pages 141–145.

M. Rodriguez-Moreno, G. Brat, N. Muscettola et D. Rijsman (2007). Validation of a multi-agent architecture for planning and execution. In *18th International Workshop on Principles of Diagnosis*, pages 368–371.

M. Rokeach (1973). *The nature of human values*. New York Free Press.

S. Russell et P. Norvig (1995). *Artificial Intelligence : a modern approach*. Prentice Hall.

S. Russell et P. Norvig (2003). *Artificial Intelligence - A Modern Approach*. Prentice Hall.

J. Sabater, M. Paolucci et R. Conte (2006). Repage : Reputation and image among limited autonomous partners. *Journal of artificial societies and social simulation*, 9(2).

J. Sabater et C. Sierra (2001). Regret : A reputation model for gregarious societies. In *Proceedings of the 4th workshop on deception fraud and trust in agent societies*, volume 70.

T. W. Sandhlom et V. R. Lesser (1997). Coalitions among computationally bounded agents. *Artificial intelligence*, 94(1) :99–137.

T. Sandholm, K. Larson, M. Andersson, O. Shehory et F. Tohmé (1999). Coalition structure generation with worst case guarantees. *Artificial Intelligence*, 111(1–2) :209–238.

T. W. Sandholm (1999). *Multiagent systems : a modern approach to distributed artificial intelligence*, chapter Distributed rational decision making, pages 201–258. MIT press.

A. Saptawijaya et L. Pereira (2014). Towards modeling morality computationally with logic programming. In *16th International Symposium on Practical Aspects of Declarative Languages*, pages 104–119.

J. B. Schafer, J. Konstan et J. Riedl (1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce*, pages 158–166.

D. Schmeidler (1969). The nucleolus of a characteristic function game. *SIAM Journal on Applied Mathematics*, 17(6) :1163–1170.

M. Schroeder (2016). Value theory. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

S. Schwartz (2012). An overview of schwartz theory of basic values. *Online Readings of Psychology and Culture*, 2(1).

W. Scott (1998). *Organizations : rational, natural and open systems*. Prentice Hall.

S. Shapiro, Y. LespÃ©rance et H. J. Levesque (2002). The Cognitive Agents Specification Language and Verification Environment for Multiagent Systems. In *AAMAS*, pages 19–26. ACM Press.

L. Shapley (1952). A value for n-person games. Technical report, RAND Corporation.

L. Shapley (1953). A value for n-person games. In H. Kuhn et A. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press.

O. Shehory et S. Kraus (1998). Methods for task allocation via agent coalition formation. *Artificial Intelligence*, 101(1) :165–200.

T. Sheridan et W. Verplank (1978). Human and computer control of undersea teleoperators. Technical report, MIT Man-Machine Systems Laboratory.

A. Shiloni, N. Agmon et G. Kaminka (2009). Of robot ants and elephants. In *8th International Conference on Autonomous Agents and Multiagent Systems*, pages 81–88.

Y. Shoham (1993). Agent-oriented programming. *Artificial Intelligence*, 60(1) :51–92.

J. Sichman, V. Dignum et C. Castelfranchi (2005). Agent organizations : a concise overview. *Special Issue in Agent Organizations in the Journal of the Brazilian Computer Society*, 11(1).

H.-A. Simon (1990). Invariants of human behavior. *Annual Review on Psychology*, 41 :1–19.

A. Singh *et al.* (2006). Eclipse attacks on overlay networks : Threats and defenses. In *Proceedings of the INFOCOM*.

M. Singh (2011). Trust as dependence : A logical approach. In *10th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 863–870.

C. Smith, A. Ambrossio, L. Mendoza et A. Rotolo (2011). Combinations of normal and non-normal modal logics for modeling collective trust in normative mas. In *4th International Conference on AI Approaches to the Complexity of Legal Systems*, pages 189–203.

S. M. Specht et R. B. Lee (2004). Distributed denial of service : Taxonomies of attacks, tools, and countermeasures. In *Proceedings of the International Conference on Parallel and Distributed Computing (and Communications) Systems*, pages 543–550.

M. Srivatsa, L. Xiong et L. Liu (2005). TrustGuard : countering vulnerabilities in reputation management for decentralized overlay networks. In *Proceedings of the 14th International World Wide Web Conference*, pages 422–431.

K. Stathis, A. Kakas, W. Lu, N. Demetriou, U. Endriss et A. Bracciali (2004). PROSOCS : a platform for programming software agents in computational logic. In J. Müller et P. Petta, editors, *Proceedings of the Fourth International Symposium "From Agent Theory to Agent Implementation" (AT2AI-4)*, pages pages 523–528, Vienna, Austria.

M. Stephen (1994). *Formalising trust as a computational concept.* Thèse de doctorat, University of Stirling, scotland.

P. Stone et M. Veloso (2000). Multiagent systems : A survey from a machine learning perspective. *Autonomous Robots*, 8(3) :345–383.

S. Sung et D. Dimitrov (2007). On myopic stability concepts for hedonic games. *Theory and Decision*, 62(1) :31–45.

T. Suzuki *et al.* (2015). Solutions for cooperative games with and without transferable utility. Technical report, School of Economics and Management.

S. Swchartz (1992). Universals in the content and structure of values : Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 25 :1–65.

G. Theodorakopoulos et J. S. Baras (2006). On trust models and trust evaluation metrics for ad hoc networks. *Selected Areas in Communications, IEEE Journal on*, 24(2) :318–328.

M. Timmons (2012). *Moral theory : An introduction.* Rowman and Littlefiled.

W. Truszkowski, L. Hallock, C. Rouff, J. Karlin, J. Rash, M. Hinchey et R. Sterritt (2009). *Autonomous and Autonomic Systems with Applications to NASA Intelligent Spacecraft Operations and Exploration Systems.* Springer-Verlag.

M. Tufis et J.-G. Ganascia (2012). Normative rational agents : A BDI approach. In *1st Workshop on Rights and Duties of Autonomous Agents*, pages 37–43. CEUR Proceedings Vol. 885.

S. Turkle et A. Shapiro (2011). Social robots raise moral, ethical questions. Morning Edition.

T. Vallée (2015). *De la manipulation dans les systèmes multi-agents : une étude sur les jeux hédoniques et les systèmes de réputation.* Thèse de doctorat, Normadie Université.

T. Vallée et G. Bonnet (2015). Using kl divergence for credibility assessment. In *14th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1797–1798.

T. Vallée et G. Bonnet (2017). Jeux de coalitions hédoniques à concepts de solution multiples. In *25es Journées Francophones sur les Systèmes Multi-Agents*, pages 53–62.

T. Vallée et G. Bonnet (2017). Jeux de coalitions hédoniques à concepts de solution multiples. In *25es Journées Francophones sur les Systèmes Multi-Agents*, pages 53–62.

T. Vallée, G. Bonnet et F. Bourdon (2014a). De l'utilisation des politiques de bandits manchots dans les systèmes de réputation. In *19e Congrès National sur la Reconnaissance de Formes et l'Intelligence Artificielle.*

T. Vallée, G. Bonnet et F. Bourdon (2014b). Multi-armed bandit policies for reputation systems. In *13th International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 279–290.

T. Vallée, G. Bonnet et F. Bourdon (2015). Politiques de bandits manchots et crédibilité dans les systèmes de réputation. *Revue d'Intelligence Artificielle*, 29(3-4) :369–398.

T. Vallée, G. Bonnet, B. Zanuttini et F. Bourdon (2013). étude des attaques sybil sur les jeux hédoniques. In *21es Journées Francophones sur les Systèmes Multi-Agents*.

T. Vallée, G. Bonnet, B. Zanuttini et F. Bourdon (2014c). A study of sybil manipulations in hedonic games. In *13th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 21–28.

L. Vanhée (2015). *Using Culture and Values to Support Flexible Coordination*. Thèse de doctorat, Université de Montpellier 2 and Utrecht Universiteit.

G. Vauvert et A. El Fallah-Seghrouchni (2001). Coalition formation among strong autonomous and weak rational agents. In *Proceedings. 10th European Workshop Modelling Autonomous Agents in a Multi-agent World*.

J. Vermorel et M. Mohri (2005). Multi-armed bandit algorithms and empirical evaluation. In *Proceedings of the 16th European Conference on Machine Learning, Porto, ECML 2005*, pages 437–448.

B. Waggoner, L. Xia et V. Conitzer (2012). Evaluating resistance to false-name manipulations in elections. In *Proceedings of the 26th Conference on Artificial intelligence, AAAI 2015*.

J. L. Wang et S.-P. Huang (2007). Fuzzy logic based reputation system for mobile ad hoc networks. In *11th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, pages 1315–1322.

B. Werger (1999). Cooperation without deliberation : A minimal behavior-based approach to multi-robot teams. *Artificial Intelligence*, 110 :293–320.

D. Weyns, A. Omicini et J. Odell (2007). Environment as a first class abstraction in multiagent systems. *Autonomous Agents and Multi-agent Systems*, 14 :5–30.

A. Whitby, A. Jøsang et J. Indulska (2004). Filtering out unfair ratings in bayesian reputation systems. In *Proceedings of the 7th Int. Workshop on Trust in Agent Societies*, volume 6, pages 106–117.

B. Whitworth (2006). Socio-technical systems. *Encyclopedia of human computer interaction*, pages 553–541.

T. Wieder (2008). The number of certain k-combinations of an n-set. *Applied Mathematics E-Notes*, 8 :45–52.

V. Wiegel (2006). Building blocks for artificial moral agents. *Artificial Life X*.

A. Winfield, C. Blum et W. Liu (2014). Towards and Ethical Robot : Internal Models, Consequences and Ethical Action Selection. In *Advances in Autonomous Robotics Systems*, volume 8717, pages 85–96.

E. Winter (1989). A value for cooperative games with levels structure of cooperation. *International Journal of Game Theory*, 18(2) :227–240.

E. Winter (1991). On non-transferable utility games with coalition structure. *International Journal of Game Theory*, 20(1) :53–63.

M. Wooldridge et N. Jennings (1995). Agent theories, architectures and languages : a survey. In M. Wooldridge et N. Jennings, editors, *Intelligent Agents*, pages 1–22. Springer-Verlag.

D. Xiu et Z. Liu (2005). A formal definition for trust in distributed systems. In *Information Security*, pages 482–489. Springer.

H. Yanco et J. Drury (2004). Classifying human-robot interaction : an updated taxonomy. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 2841–2846.

C. Yang (1997). A family of values for n-person cooperative transferable utility games : An extension to the shapley value. Technical report, University of New-York Buffalo.

X. Yang et J. Gao (2014). Uncertain core for coalitional game with uncertain payoffs. *Journal of Uncertain Systems*, 8 :13–21.

H. Yu, M. Kaminsky, P. B. Gibbons et A. Flaxman (2006). SybilGuard : defending against Sybil attacks via social networks. *SIGCOMM Computer Communication Review*, 36(4) :267–278.

H. Zhao et X. Li (2009). H-trust : A group trust management system for peer-to-peer desktop grid. *Journal of Computer Science and Technology*, 24(5) :833–843.

R. Zhou et K. Hwang (2007). Powertrust : A robust and scalable reputation system for trusted peer-to-peer computing. *Parallel and Distributed Systems, IEEE Transactions on*, 18(4) :460–473.