Une logique modale pour la caractérisation des manipulations entre agents autonomes

Christopher Leturc, Grégory Bonnet

Une logique modale pour la caractérisation des manipulations entre agents autonomes

Christopher Leturc[1] **Grégory Bonnet**[1]

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

prenom.nom@unicaen.fr

Résumé

La *manipulation* est l'intention délibérée d'instrumentaliser une victime tout en veillant à lui dissimuler cette intention. Dans cet article, nous définissons un cadre logique nous permettant d'exprimer et raisonner sur cette définition de la manipulation. En remarquant qu'il est difficile d'associer des liens logiques entre une stratégie de manipulation et la manipulation visée par la stratégie, nous définissons une *hypothèse de non trivialité de la manipulation* qui nous permet de réduire les stratégies de manipulation à des formes à part entière de manipulations. Nous nommons ces formes des *manipulations approchées*. Enfin, nous montrons que notre cadre permet d'exprimer beaucoup de stratégies de manipulation comme des stratégies de dissimulation de sources, de mensonge crédible, ou encore de baratinage intentionnel.

Abstract

Manipulation is the deliberate intention to instrumentalise a victim while making sure to conceal that intent. In this article we define a logical framework allowing us to express and reason about this definition of manipulation. Noting that it is difficult to associate logical links between a manipulation strategy and the manipulation targeted by the strategy. We define a *hypothesis of non-triviality of manipulation* that allows us to reduce manipulation strategies to forms of manipulation. We call these forms *approximate manipulations*. Finally, we show that our framework allows to express many strategies of manipulation like source concealment strategies, believable lies, or intentional bullshit.

Intuitivement, la manipulation peut être définie comme l'intention de modifier l'état d'un objet, d'une personne et plus généralement d'un agent sans que ce dernier ne s'en rende compte. Dans des domaines comme les systèmes de réputation [23], la théorie des choix sociaux [18] ou la théorie des jeux algorithmique [10, 34], les agents peuvent trouver des intérêts à manipuler. Par exemple, les systèmes de réputation sont des systèmes permettant d'évaluer la confiance qu'un agent peut accorder à un autre en fonction des confiances que disent lui accorder les autres agents. Un agent peut avoir alors intérêt à mentir sur ses relations de confiance afin d'induire en erreur les autres et les pousser à interagir avec certains agents spécifiques. Il s'agit d'une manipulation au sens où, pour que cette stratégie soit efficace, il est nécessaire que les agents trompés ne soient pas au courant des actions que l'agent trompeur est en train d'entreprendre.

Dans le domaine de l'intelligence artificielle, très peu de travaux se sont intéressés à modéliser la manipulation. Ceux qui s'en rapprochent sont des travaux de logique formelle qui se sont intéressés à la modélisation de l'influence sociale [20] et de la tromperie [36, 25]. Or, telle que nous l'avons intuitivement définie précédemment, la manipulation peut être vue comme la combinaison de ces deux approches, c'est-à-dire qu'un agent va avoir l'intention d'influencer un autre, tout en veillant à le tromper sur ses véritables intentions. Ainsi, l'objectif de cet article est de proposer un cadre formel nous permettant de modéliser et de raisonner sur cette notion de manipulation.

En nous fondant sur des travaux de sciences sociales, nous donnons une définitions générale de la manipulation en section 1. Nous passons ensuite en revue plusieurs travaux de la littérature en intelligence artificielle au regard de cette définition. En section 2, nous proposons un cadre logique permettant de raisonner sur la notion de manipulation et étudions certaines propriétés logiques de ce cadre. Dans la section 3, nous utilisons notre cadre pour définir différentes formes de manipulations.

1 État de l'art

Dans un premier temps, nous allons nous intéresser aux travaux sur la manipulation du point de vue des sciences

sociales. Nous présenterons les travaux qui se sont penchés sur la modélisation de cette notion, en particulier au travers de l'influence et la malhonnêteté.

1.1 La manipulation dans les sciences humaines

Dans le domaine de la politique ou du marketing, la manipulation est définie comme l'action consistant à altérer le jugement des individus, les privant d'une partie de leur jugement et de leurs choix délibérés [29, 31]. Au-delà d'une influence sur les états mentaux d'une personne, les sciences humaines [13], en particulier la psychiatrie [9, 12, 16], définissent aussi la manipulation comme l'instrumentalisation de la personne : la personne manipulée est vue comme un outil que le manipulateur va chercher à contrôler à son avantage. Toutefois, cette définition fait entrer la persuasion rationnelle, la tromperie ou même la coercition dans le champ de la manipulation. Par exemple, un vol à main armée entrerait dans ce champ, ce qui peut sembler contre-intuitif. En effet, au vu d'un consensus dans la littérature en sciences humaines, nous considérons dans cette article une synthèse des définitions de [1, 7, 24].

Définition 1.1 *On appelle* manipulation *l'intention, délibérée, d'instrumentaliser une victime tout en veillant à lui dissimuler cette intention.*

De manière intéressante, cela rejoint la définition du Grand Larousse : « l'action d'orienter la conduite de quelqu'un, d'un groupe dans le sens qu'on désire et sans qu'ils s'en rendent compte » [19]. Ainsi, la manipulation est bien à distinguer de la persuasion rationnelle, de la coercition et elle n'est pas non plus réduite à la tromperie.

En revanche comme le rappelle le philosophe Patrick Todd [30], manipuler quelqu'un est influencer son comportement sans qu'il ne s'en rende compte, c'est-à-dire par moyen de la tromperie. La tromperie est donc nécessaire à la mise en œuvre de la manipulation. Une forme de tromperie peut être par exemple le mensonge qui consiste à avoir l'intention d'influencer les croyances d'un autre agent pour le détourner de la vérité que l'on croit détenir [25]. Par ailleurs, une stratégie visant à influencer des croyances, des désirs ou de la confiance est nécessairement une manipulation visant à instrumentaliser la victime. En effet, mentir délibérément pour amener simplement autrui à croire une information erronée n'a guère de sens s'il n'y a pas une autre intention derrière. Comme l'affirmait Rousseau : « Mentir sans profit ni préjudice de soi ni d'autrui n'est pas mentir : ce n'est pas mensonge, c'est fiction ».

Ainsi, puisqu'un mensonge précède systématiquement une manipulation, il doit exister un lien logique liant le mensonge à la manipulation.

1.2 Modèles formels de l'influence et de la tromperie

Peu de travaux en intelligence artificielle s'intéressent à décrire la manipulation en soi. Les travaux qui s'y intéressent cherchent plutôt à construire des manipulations ou cherchent à proposer des systèmes robustes pour des cas applicatifs spécifiques comme c'est le cas dans les systèmes de réputation [33], en théorie du choix social [11] ou encore en théorie des jeux [37]. Dans cet article, nous nous intéressons plutôt à une formalisation générale décrivant la manipulation à l'aide des logiques modales. En effet, les logiques modales nous permettent de représenter explicitement les notions d'intention, de croyance et de connaissance qui sont fondamentales pour les manipulations. De manière intéressante, plusieurs approches de ce type se sont déjà intéressées à des notions connexes comme l'influence [6, 20], au mensonge et la tromperie [25, 36].

Lorini [20] définit via les logiques STIT l'influence sociale comme le fait qu'un agent i veille à ce que l'agent j réalise dans le futur l'action espérée par l'agent i. Bottazzi et al. [6] définissent quant à eux l'influence via les logiques BIAT comme l'intention d'un agent i d'amener un autre agent j à amener un certain état de monde. Si ces deux définitions de l'influence ont du sens, nous optons dans cet article pour la seconde car elle ne fait pas intervenir d'autres modalités que l'intention.

Le mensonge est l'intention d'un agent i de faire croire quelque chose à un agent j alors que l'agent i croit en son contraire et informe l'agent j du contraire [36, 25]. Van Ditmarsch *et al.* [36] utilisent des logiques épistémiques dynamiques avec une modalité pour décrire l'action d'énoncer publiquement ou de manière privée une information. Sakama *et al.* [25] utilisent une logique modale et introduisent une modalité de communication entre deux agents, une modalité de croyance ainsi qu'une modalité d'intention. Dans cet article, nous n'utilisons pas de modalité explicite de communication. En effet, le mensonge peut se résumer à une intention délibérée d'un agent i de faire croire quelque chose à un agent j alors que l'agent i croit en son contraire. Ici, la modalité de communication est réduite à l'intention délibérée de faire croire quelque chose à un agent.

Concernant la manipulation elle-même, un seul travail a cherché à la représenter avec un langage de logique modale à notre connaissance. C'est au psychologue Joel Rudinow que nous devons cette première approche [24]. Celui-ci définit la manipulation comme l'intention d'un agent manipulateur i de jouer sur les motivations complexes d'un agent j par le moyen de la tromperie ou d'une faiblesse supposée de j. Cependant, aucune sémantique n'est donnée et il est difficile d'associer une sémantique explicite aux termes *motivations complexes* ou encore *faiblesse supposée*. Joel Rudinow nous donne l'intuition que les logiques épistémiques peuvent être un moyen pour exprimer ces termes.

De manière intéressante, la définition de Rudinow n'entre pas en contradiction avec la notre qui est *l'intention, délibérée, d'instrumentaliser une victime tout en veillant à lui dissimuler cette intention*. En effet, lorsqu'un agent joue sur une faiblesse supposée, il semble naturel qu'il va aussi veiller à ce que l'autre agent ne s'en rende pas compte.

2 Une logique de la manipulation

Dans cet article, nous proposons un cadre qui va considérer plusieurs modalités pour exprimer l'intention, la croyance et la connaissance.

2.1 Le besoin d'une logique d'action

Étant donné qu'une manipulation est une intention délibérée, nous avons donc besoin d'une logique d'action pour représenter les actes délibérés et les conséquences des actions. Pour formaliser la notion d'actions en logique, il existe de nombreux formalismes [26]. Nous pouvons citer les logiques dynamiques [14, 3] et les logiques temporelles [2, 17] qui considèrent des actions comme des programmes associés à une sortie, ou les logiques épistémiques dynamiques qui expriment les conséquences logiques générées par des annonces dans les croyances des agents [35]. Toutefois, deux approches nous paraissent pertinentes : le formalisme STIT [4, 20] et le formalisme BIAT [22, 32] qui considèrent de manière abstraite le fait de veiller à ce que quelque chose se réalise.

En effet, les approches STIT et BIAT considèrent les actions comme le fruit de leurs conséquences. Or, pour modéliser la manipulation, nous n'avons pas besoin de considérer l'action explicite effectuée par un agent mais plutôt son résultat car une manipulation ne peut être réduite à une action unique mais bien à des stratégies diverses. Il s'agit d'un niveau d'abstraction bien adapté à ces formalismes. En effet, les approches BIAT considèrent une modalité E_i qui signifie que l'agent *i brings it about* tandis que les approches STIT représentent une modalité $[STIT]_i$ qui décrit que le fait que l'agent *i sees to it that*. Bien que ces deux approches sont souvent confondues, la différence entre ces deux formalismes réside dans le niveau d'abstraction des modalités. Les approches STIT expriment le fait qu'un agent veille à ce que quelque chose se réalise et introduisent une notion de temporalité. Les approches BIAT s'affranchissent de cette notion de temporalité et considèrent sous sa forme la plus abstraite le fait d'amener quelque chose à être vrai.

Nous ne considérons pas ici la notion de temporalité mais uniquement celle d'intention, de connaissance et de croyance. En particulier, nous faisons la distinction entre une notion d'actions effectuées de façon délibérées (E_i^d) et une notion pour capturer les conséquences d'une ou plusieurs actions effectuées de façon délibérée ou non (E_i). Ces deux modalités décrivent donc le fait qu'un agent *i* effectue une ou plusieurs actions qui amènent à rendre vrai un nouvel état du monde (E_i) et le fait qu'un agent *i* effectue *de façon délibérée* une ou plusieurs actions qui amènent à rendre vrai un nouvel état du monde (E_i^d). En effet, la modalité E_i décrit toutes les conséquences de l'action ou la série d'actions qu'un agent *i* peut effectuer à un instant donné, tandis que la modalité E_i^d décrit les conséquences (désirées) que l'agent *i* a souhaité obtenir et ce de manière délibérée. Cette modalité d'intention délibérée E_i^d peut sembler, par son nom, très proche de la modalité de STIT délibéré introduite par Lorini [20]. En réalité les deux modalités sont différentes : Lorini considère que quelque chose est réalisé de manière délibérée s'il est possible du contraire dans le monde courant. Pour notre part, nous considérons qu'un acte est délibéré si l'agent a eu l'intention de rendre vrai cet état du monde et ce, peu importe l'état actuel du monde courant. Imaginons la situation dans laquelle un individu *i* veille à poignarder un autre individu *j* allongé dans son lit, pour qu'il meurt. Or, cet individu *j* était déjà mort avant que *i* le poignarde et le tue. Ainsi, pour le STIT délibéré de Lorini [20], puisque l'agent *j* était déjà mort, il n'était pas le cas que l'agent *i* puisse avoir l'intention délibérée de tuer l'agent *j*. Or nous pensons que justement, l'agent *i* avait pour intention délibérée de tuer l'agent *j* même s'il était déjà mort. Nous allons voir dans la suite de l'article que notre modalité E_i est en réalité un STIT alors que notre modalité E_i^d est un BIAT classique.

2.2 Langage

Soit un ensemble de lettres propositionnelles $\mathcal{P} = \{a, b, c, ...\}$, un ensemble d'agents $\mathcal{N}$ avec $i, j \in \mathcal{N}$ deux agents, et $p \in \mathcal{P}$ une variable propositionnelle. Nous définissons le langage $\mathcal{L}_{KBE}$ avec la règle de grammaire BNF suivante :

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

Nous considérons des modalités E_i, E_i^d, K_i et B_i pour chaque agent *i*. Ainsi les formules $E_i\phi$, $E_i^d\phi$, $K_i\phi$ et $B_i\phi$ signifient respectivement qu'un agent *i* effectue [1] une ou plusieurs actions menant à une conséquence ϕ, qu'un agent *i* effectue de manière délibérée une ou plusieurs actions menant à une conséquence ϕ, que l'agent sait que ϕ est vrai et que l'agent croit que ϕ est vrai.

2.3 Sémantique associée

Pour interpréter notre langage, nous considérons la sémantique $C = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^d\}_{i\in\mathcal{N}})$ associée à $\mathcal{L}_{KBE}$ où :

1. Par soucis de lisibilité, nous pourrons utiliser dans la suite de cet article les expressions "veiller à ce que" et ou "faire en sorte que". Dans tous les cas, l'interprétation sémantique est "effectuer une ou plusieurs actions qui mènent à".

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

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

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

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

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

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

— $\{\mathcal{E}_i^d\}_{i\in\mathcal{N}}$ un ensemble de fonctions effectives :

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

Nous définissons un modèle comme $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^d\}_{i\in\mathcal{N}}, i)$ avec $i : \mathcal{P} \to 2^{\mathcal{W}}$ une fonction d'interprétation. Pour tout monde $w \in \mathcal{W}$, pour toute formule $\phi, \psi \in \mathcal{L}_{KBE}$ et pour tout $p \in \mathcal{P}$:

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

Rappelons que ϕ est valide dans un modèle $\mathcal{M}$ (noté $\mathcal{M} \models \phi$) si, et seulement si, pour tout monde $w \in \mathcal{W}$, ϕ est satisfiable dans w i.e $\mathcal{M}, w \models \phi$ est vrai. Une formule ϕ est valide dans un cadre C (noté $\models_C \phi$ ou $C \models \phi$) si, et seulement si, pour tout modèle $\mathcal{M}$ fondé sur C, $\mathcal{M} \models \phi$. Dans ce cas, ϕ est un théorème du cadre, noté $\vdash \phi$.

Pour les modalités de connaissance et de croyance, nous contraignons de manière classique notre cadre C de telle sorte que, pour tout agent $i \in \mathcal{N}$, $\mathcal{K}_i$ est une relation d'équivalence (transitive, réflexive et symétrique) et $\mathcal{B}_i$ est sérielle, transitive et euclidienne. Les contraintes et relations entre ces deux modalités ont déjà été bien étudiées [28]. Ainsi, nous considérons tout d'abord qu'un agent i croit ce qu'il sait, c'est-à-dire :

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

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

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

De la même façon, un agent sait ce qu'il ne croit pas :

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

La contrainte E1 exprime le fait qu'une fois les actions menant à ϕ effectuées par l'agent i, alors ϕ est vraie (axiome T).

$$\forall w \in \mathcal{W} : w\mathcal{E}_i w \quad (E1)$$

En effet, lorsqu'un agent met en œuvre une ou plusieurs actions, s'il les a effectué, c'est que c'est bel et bien le cas et que l'action a bien eu la conséquence ϕ attendue. De cette contrainte, nous déduisons immédiatement que la relation $\mathcal{E}_i$ est aussi *sérielle* et donc nous déduisons dans un tel système la propriété D.

Ensuite $\mathcal{E}_i$ est *transitive* puisqu'un agent i lorsqu'il veille à rendre vraie ϕ, ce dernier veille aussi à ce que son action soit bien effectuée, et donc veille à veiller à ce que ϕ soit vraie.

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

Définissons maintenant les contraintes pour la modalité d'intention délibérée. La différence principale avec la modalité d'intention est qu'un agent i ne peut pas effectuer une ou plusieurs actions de manière délibérée qui mènent à rendre vrai une tautologie (appelée contrainte nNEd).

$$\forall w \in \mathcal{W} : W \notin \mathcal{E}_i^d(w) \quad (nNEd)$$

De plus, lorsqu'un agent i effectue de manière délibérée une ou plusieurs actions menant à rendre vrai un nouvel état du monde alors l'agent i effectue une ou plusieurs actions menant à rendre vrai ce nouvel état du monde. Il y a donc un lien entre modalité d'action délibérée et action non délibérée, représenté par la contrainte EdE.

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

Enfin, la modalité d'action délibérée dispose d'introspection positive (EdKP) et négative (EdKN) en lien avec la modalité de connaissance.

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

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

Respectivement ces contraintes signifient qu'un agent qui effectue une ou plusieurs actions de manière délibérée menant à une certaine conséquence sait ce qu'il est en train de faire de manière délibérée, et que s'il n'est pas le cas qu'un agent i effectue une ou plusieurs actions de manière délibérée menant à une certaine conséquence alors l'agent i sait qu'il ne les a pas effectué de manière délibérée.

2.4 Système axiomatique correspondant

Nous prouvons dans cette section que compte tenu des contraintes données précédentes sur notre cadre, nous pouvons prouver que le système axiomatique correspondant est celui de la figure 1. Rappelons que $\vdash \phi$ signifie que ϕ est une tautologie.

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

FIGURE 1 – Axiomatique minimale du cadre KBE

Proposition 2.1 *Le système KBE est correct.*

Preuve 2.1 *Pour des raisons d'espace, nous allons donner un schéma de preuve.*

Montrer que la substitution, le modus ponens et la nécessitation préservent la validé pour toute modalité normale est standard à démontrer [?]. De plus, il est bien connu qu'une sémantique d'une modalité normale d'un système S5 qui préserve la validité est une relation d'équivalence [?]. En effet, comme les relations $\mathcal{K}_i$ et $\mathcal{E}_i$ sont des relations d'équivalence, les règles de S5 préservent la validité.

Concernant les bridges rules entre la modalité K_i et B_i, Stalnaker [?] démontre que les propriétés de la sémantique de notre système que nous proposons, préserve la validité. De plus, il est bien connu qu'une relation sérielle, transitive et euclidienne préserve la validité dans un système $KD45$ et il est bien le cas pour la modalité B_i.

Enfin, les propriétés sur la sémantique que nous proposons concernant la modalité E_i^d et les liens avec la modalité E_i préservent aussi la validité [?]. Nous procédons par un raisonnement par contraposition. Pour les sens ($\Rightarrow$), nous construisons un modèle tel qu'il satisfasse la condition du cadre donnée par la contraposée et montrons alors qu'il existe un modèle et un monde tel que la formule correspondante ne soit pas vérifiée. Pour les sens ($\Leftarrow$), nous procédons aussi par contraposition et considérons qu'il existe un modèle et un monde tel que la formule correspondante ne soit pas vérifiée et nous déduisons alors que les propriétés du cadre ne sont pas vérifiées.

2.5 Complétude du système

Théorème 2.1 *Le système KBE est complet.*

Preuve 2.2 *Pour des raisons liées à l'espace disponible. Nous allons présenter un schéma de la preuve de complétude. La preuve repose sur la méthode de Henkin et les ensembles de formules maximaux S-consistants. L'astuce de la preuve repose sur la définition du modèle canonique suivante. On appelle* modèle canonique*, un modèle $\mathcal{M}^c = (\mathcal{W}^c, \{\mathcal{B}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{K}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^{d^c}\}_{i\in\mathcal{N}}, V^c)$ tel que : $\forall i \in \mathcal{N}, \forall w, v \in \mathcal{W} : w\mathcal{B}_i^c v$ ssi $B_i\phi \in w \Rightarrow \phi \in v$, $\forall i \in \mathcal{N}, \forall w, v \in \mathcal{W} : w\mathcal{K}_i^c v$ ssi $K_i\phi \in w \Rightarrow \phi \in v$, $\forall i \in \mathcal{N}, \forall w, v \in \mathcal{W} : w\mathcal{E}_i^c v$ ssi $E_i\phi \in w \Rightarrow \phi \in v$, $\forall i \in \mathcal{N}, \forall w \in \mathcal{W} : \mathcal{E}_i^{d^c}(w) := \{\|\phi\| : E_i\phi \in w\}$ avec $\|\phi\| := \{w | w \in \mathcal{W}^c \wedge \phi \in w\}$ et $\forall p \in \mathcal{P}, V^c(p) = \|p\|, \|p\| := \{w | w \in \mathcal{W}^c \wedge p \in w\}$. Nous remarquons que la partie du modèle canonique qui concerne la sémantique de voisinage repose sur la notion de* modèle canonique minimal *[21]. Nous prouvons que ce modèle canonique satisfait le lemme de vérité [21] et prouvons qu'il satisfait toutes les propriétés du cadre. Les propriétés du modèle canonique à démontrer concernant les relations $\mathcal{K}_i$, $\mathcal{B}_i$ et $\mathcal{E}_i$ sont classiques [5]. Pour les autres propriétés concernant la sémantique associée à la modalité E_i^d, il existe une petite difficulté concernant la propriété associée à l'axiome $(5_{K_iE_i})$. Il faut remarquer que pour un $X \notin \mathcal{E}_i^{d^c}(w)$ avec $w, v \in \mathcal{W}^c$ et $w\mathcal{K}_i^c v$, il existe deux cas. Lorsque X peut s'écrire sous la forme $X = \|\phi\|$ et donc $\neg E_i\phi \in w$. Puis en appliquant la méthode de Henkin, on déduit que $K_i\neg E_i\phi \in w$, puis nous déduisons par application de la définition du modèle canonique que $w\mathcal{K}_i v$, on a $X \notin \mathcal{E}_i^{d^c}(v)$. Mais lorsque X ne peut pas s'écrire sous cette forme, par application de la définition du modèle canonique on en déduit alors que quelque soit $u \in \mathcal{W}^c$, $X \notin \mathcal{E}_i^{d^c}(u)$ et donc pour $w\mathcal{K}_i v$, on a $X \notin \mathcal{E}_i^{d^c}(v)$.*

Par conséquent, pour toute formule ϕ valide dans C, on a donc que $\mathcal{M}^c \models \phi$ et donc $\forall w \in \mathcal{W}^c : \mathcal{M}^c, w \models \phi$, et par le lemme de vérité [21], $\forall w \in \mathcal{W}^c, \phi \in w$. Enfin, par une propriété sur les ensembles maximaux S-consistants $\forall w \in \mathcal{W}^c, \phi \in w$ ssi $\vdash \phi$. Nous venons donc de prouver que le système KBE est complet.

Nous prouvons enfin la forte complétude de notre système.

Théorème 2.2 *Le système KBE est fortement complet, c'est-à-dire, pour toute formule $\phi \in \mathcal{L}_{KBE}$ et tout ensemble de formules $\Gamma \subset \mathcal{L}_{KBE}$, si $\Gamma \models \phi$ alors $\Gamma \vdash \phi$.*

Preuve 2.3 *Démontrons la complétude forte par contraposition. Soit $\Gamma \subset \mathcal{L}_{KBE}$ un ensemble de formules tel que*

$\Gamma \nvdash \phi$. *On a donc que* $\Gamma \cup \{\neg\phi\}$ *est un ensemble de formules S-consistant*[2]. *On prouve facilement que dans notre cadre [21], puisque* $\Gamma \cup \{\neg\phi\}$ *est S-consistant, il existe un modèle canonique* $\mathcal{M}$ *tel que* $\mathcal{M} \models \Gamma \cup \{\neg\phi\}$, *c'est-à-dire,* $\mathcal{M} \models \Gamma$ *et* $\mathcal{M} \models \neg\phi$. *Par conséquent, on vient donc de prouver qu'il existe un modèle* $\mathcal{M}$ *tel que* $\mathcal{M}, \Gamma \not\models \phi$.

2.6 Quelques propriétés du cadre

Remarquons que l'axiome (D) est valide pour E_i^d.

Théorème 2.3 $\vdash \neg E_i^d \bot \quad (D_{E_i^d})$

Preuve 2.4
$\vdash E_i^d \bot \Rightarrow E_i \bot$
$\vdash (E_i^d \bot \Rightarrow E_i \bot) \Rightarrow (\neg E_i \bot \Rightarrow \neg E_i^d \bot)$
$\vdash \neg E_i \bot \Rightarrow \neg E_i^d \bot$
$\vdash \neg E_i \bot$
$\vdash \neg E_i^d \bot$

D'autres théorèmes intéressants peuvent être déduits. En particulier, lorsqu'un agent veille à ce qu'un autre agent croit quelque chose, il veille alors à ce que l'autre agent ne croit pas que cet agent puisse savoir le contraire.

Théorème 2.4 *Dissimulation :*

1. *des croyances contraires :* $\vdash E_i B_j \phi \Rightarrow E_i \neg B_j K_i \neg\phi$
2. *des croyances :* $\vdash E_i \neg B_j \phi \Rightarrow E_i \neg B_j K_i \phi$

Plutôt que de donner une longue preuve fondée sur les systèmes de Hilbert, nous présentons une intuition de la preuve du premier théorème par un exemple.

Preuve 2.5
(1) $\vdash K_i B_j \phi \Rightarrow B_j \phi$
$\vdash B_j \phi \wedge B_j K_i \neg\phi \Rightarrow B_j \phi$
$\vdash B_j(K_i \neg\phi \Rightarrow \neg\phi)$
$\vdash B_j(K_i \neg\phi \Rightarrow \neg\phi) \Rightarrow (B_j K_i \neg\phi \Rightarrow B_j \neg\phi)$
$\vdash B_j K_i \neg\phi \Rightarrow B_j \neg\phi$
$\vdash B_j \neg\phi \Rightarrow \neg B_j \phi$
$\vdash B_j K_i \neg\phi \Rightarrow B_j \neg\phi \Rightarrow \neg B_j \phi$
$\vdash (B_j K_i \neg\phi \Rightarrow B_j \neg\phi \Rightarrow \neg B_j \phi) \Rightarrow ((B_j K_i \neg\phi \Rightarrow B_j \neg\phi) \Rightarrow (B_j K_i \neg\phi \Rightarrow \neg B_j \phi))$
$\vdash B_j K_i \neg\phi \Rightarrow \neg B_j \phi$
$\vdash ((B_j \phi \wedge B_j K_i \neg\phi \Rightarrow B_j \phi)) \Rightarrow ((B_j \phi \wedge B_j K_i \neg\phi \Rightarrow \neg B_j \phi) \Rightarrow \neg(B_j \phi \wedge B_j K_i \neg\phi))$
$\vdash \neg(B_j \phi \wedge B_j K_i \neg\phi)$
$\vdash \neg(B_j \phi \wedge B_j K_i \neg\phi) \equiv (B_j \phi \Rightarrow \neg B_j K_i \neg\phi)$
$\vdash B_j \phi \Rightarrow \neg B_j K_i \neg\phi$
$\vdash E_i(B_j \phi \Rightarrow \neg B_j K_i \neg\phi)$
$\vdash E_i(B_j \phi \Rightarrow \neg B_j K_i \neg\phi) \Rightarrow (E_i B_j \phi \Rightarrow E_i \neg B_j K_i \neg\phi)$
$\vdash E_i B_j \phi \Rightarrow E_i \neg B_j K_i \neg\phi$

(2) Soit $\Box_j \in \{B_j, K_j\}$,
$\vdash \Box_j(K_i \phi \Rightarrow \phi) \Rightarrow \Box_j K_i \phi \Rightarrow \Box_j \phi$
$\vdash K_i \phi \Rightarrow \phi$
$\vdash \Box_j(K_i \phi \Rightarrow \phi)$
$\vdash \Box_j K_i \phi \Rightarrow \Box_j \phi$
$\vdash (\Box_j K_i \phi \Rightarrow \Box_j \phi) \Rightarrow (\neg\Box_j \phi \Rightarrow \neg\Box_j K_i \phi)$
$\vdash \neg\Box_j \phi \Rightarrow \neg\Box_j K_i \phi$
$\vdash E_i(\neg\Box_j \phi \Rightarrow \neg\Box_j K_i \phi)$
$\vdash E_i(\neg\Box_j \phi \Rightarrow \neg\Box_j K_i \phi) \Rightarrow E_i \neg\Box_j \phi \Rightarrow E_i \neg\Box_j K_i \phi$
$\vdash E_i \neg\Box_j \phi \Rightarrow E_i \neg\Box_j K_i \phi$
$\vdash K_j \phi \Rightarrow B_j \phi$
$\vdash (K_j \phi \Rightarrow B_j \phi) \Rightarrow (\neg B_j \phi \Rightarrow \neg K_j \phi)$
$\vdash \neg B_j \phi \Rightarrow \neg K_j \phi$
$\vdash E_i(\neg B_j \phi \Rightarrow \neg K_j \phi)$
$\vdash E_i(\neg B_j \phi \Rightarrow \neg K_j \phi) \Rightarrow E_i \neg B_j \phi \Rightarrow E_i \neg K_j \phi$
$\vdash E_i \neg B_j \phi \Rightarrow E_i \neg K_j \phi$
Pour $\Box_j = K_j$, *on a donc :*
$\vdash E_i \neg B_j \phi \Rightarrow (E_i \neg K_j \phi \Rightarrow E_i \neg K_j K_i \phi)$
$\vdash (E_i \neg B_j \phi \Rightarrow (E_i \neg K_j \phi \Rightarrow E_i \neg K_j K_i \phi)) \Rightarrow ((E_i \neg B_j \phi \Rightarrow E_i \neg K_j \phi) \Rightarrow (E_i \neg B_j \phi \Rightarrow E_i \neg K_j K_i \phi))$
$\vdash E_i \neg B_j \phi \Rightarrow E_i \neg K_j K_i \phi$

Un raisonnement analogue permet de déduire le second théorème. De plus, en remarquant que la contraposition de $K_i \phi \Rightarrow B_i \phi$ est $\neg B_i \phi \Rightarrow \neg K_i \phi$, nous déduisons deux corollaires immédiats à ces théorèmes :

1. $\vdash E_i B_j \phi \Rightarrow E_i \neg K_j K_i \neg\phi$
2. $\vdash E_i \neg B_j \phi \Rightarrow E_i \neg K_j K_i \phi$

Ces deux corollaires nous permettent de mieux caractériser *ce qui n'est pas* de la manipulation. Par exemple, le fait qu'un agent i veille à ce qu'un agent j croit quelque chose tout en veillant à ce que ce dernier ne sache pas que l'agent i sait le contraire n'est pas de la manipulation puisque le premier corollaire nous permet de déduire cette implication systématiquement. En revanche, lorsque l'agent i veille à la même chose mais de manière délibérée (avec la modalité E_i^d) alors nous ne pouvons plus déduire ces théorèmes. C'est alors cette modalité d'action délibérée qui va nous permettre de définir les manipulations.

3 Modéliser les manipulations

Nous avons vu en Section 1 que les manipulations sont caractérisées entre autre par une forme de dissimulation, qu'elle porte soit sur les intentions, les croyances ou les connaissances des agents. À partir du système KBE présenté précédemment, nous modélisons trois catégories de manipulations en fonction de l'objet sur lequel porte la dissimulation. Chacune de ces catégories permet d'exprimer des *manipulations constructives*, c'est-à-dire qui visent à

2. En effet, par l'absurde, si nous avions $\Gamma \cup \{\neg\phi\}$ est S-inconsistant, on aurait qu'il existe $\psi_1, \ldots, \psi_n \in \Gamma$ telles que : $\vdash \neg(\psi_1 \wedge \ldots \wedge \psi_n \wedge \neg\phi)$ et donc par le théorème de déduction que nous pouvons prouver aisément dans notre système, on aurait aussi $\vdash (\psi_1 \wedge \ldots \wedge \psi_n) \Rightarrow \phi$ et donc on déduirait aussi $\Gamma \vdash \phi$, ce qui contredit l'hypothèse que $\Gamma \nvdash \phi$.

amener un agent à faire quelque chose, croire ou connaître quelque chose et des *manipulations destructives* qui visent à empêcher un agent de faire quelque chose, de croire ou connaître quelque chose.

3.1 Manipulations pures constructives et destructives

En termes de manipulations, un agent qui veille à dissimuler ses intentions peut le faire selon qu'il veille à ce qu'un autre fasse quelque chose, croit quelque chose ou apprenne quelque chose. Une *manipulation pure constructive* consiste à veiller qu'un agent fasse quelque chose sans que ce dernier sache qu'un manipulateur a veillé à ce qu'il fasse cette chose. Ceci est exprimé par le prédicat suivant :

$$M^{+}_{i,j}\phi = E^{d}_{i}(E_j\phi \wedge \neg K_j E^{d}_{i} E_j\phi) \quad M^{d+}_{i,j}\phi = E^{d}_{i}(E^{d}_{j}\phi \wedge \neg K_j E^{d}_{i} E_j\phi)$$

Illustrons ce prédicat avec un exemple lié à la publicité. De manière générale, nous savons tous que les intentions d'un publicitaire est de conduire les acheteurs potentiels à acheter le produit. Cette intention n'est pas dissimulée et il ne s'agit donc pas d'une manipulation. En revanche, cela le devient lorsque le publicitaire utilise pour cela une technique de vente (composée d'une ou plusieurs actions) qu'il cherche à dissimuler aux futurs acheteurs, comme par exemple l'utilisation d'images subliminales. Ainsi, le publicitaire ne cherche pas à dissimuler son intention de faire que le client achète le produit mais à dissimuler de façon délibérée la technique qu'il utilise pour inciter le client à acheter.

Nous appelons *manipulation pure destructive* :

$$M^{-}_{i,j}\phi = E^{d}_{i}(\neg E_j\phi \wedge \neg K_j E^{d}_{i} \neg E_j\phi) \quad M^{d-}_{i,j}\phi = E^{d}_{i}(\neg E^{d}_{j}\phi \wedge \neg K_j E^{d}_{i} \neg E_j\phi)$$

Cet aspect de la manipulation vise donc à empêcher un agent d'agir sans qu'il s'aperçoive de l'intention délibérée du manipulateur. Par exemple, le cas des attaques DoS [38, 27] peut être capturé par cette formule car le pirate veille, au moment où il agit, à ce que l'opérateur du réseau n'identifie pas les paquets qui circulent comme malveillant (l'intention délibérée du pirate) avant qu'il ne soit trop tard.

Remarquons qu'en pratique il n'existe qu'une unique forme de manipulation. En effet la manipulation pure destructive se ramène à sa forme constructive. En effet, par l'axiome d'introspection positive des E_i, nous pouvons dire que, lorsqu'un agent i veille de façon délibérée à ce qu'un agent j n'effectue pas une certaine action ou série d'action, cela revient intuitivement à dire que l'agent i a l'intention délibérée que l'agent j effectue une ou plusieurs actions qui l'amènent à ne pas effectuer l'action initiale.

3.2 Hypothèse de non trivialité de la manipulation

En pratique, prouver qu'un agent manipule est extrêmement difficile car il est nécessaire d'en identifier le caractère délibéré de l'intention de l'agent d'instrumentaliser l'autre ainsi que son intention de le dissimuler. Par ailleurs, nous remarquons qu'il est souvent plus simple d'identifier directement la stratégie de manipulation comme par exemple, reconnaître un mensonge plutôt que de prouver l'intention finale du manipulateur. Ensuite, nous pensons qu'il existe un lien logique (i.e une implication) non trivial (i.e qui n'est pas une tautologie) liant la stratégie de manipulation adoptée par un agent et notre définition de la manipulation. Cependant, comme il est difficile d'établir ce lien logique entre la stratégie de manipulation et la manipulation elle-même, nous proposons de définir et de faire une *hypothèse de non trivialité de la manipulation* nous permettant ainsi de réduire la manipulation à sa stratégie.

Notons $\mathcal{S} \subsetneq \mathcal{L}_{KBE}$ l'ensemble des formules qui décrivent les stratégies de manipulation des sections suivantes. Nous disons que $s_{i,j}(\phi) \in \mathcal{L}_{KBE}$ est une *stratégie de manipulation* d'un agent i envers un agent j si, et seulement si, il existe un prédicat p d'arité 3 tel que $p(i, j, \phi) = s_{i,j}(\phi)$ et $s_{i,j}(\phi) \in \mathcal{S}$. Nous formalisons l'hypothèse de non trivialité de la manipulation de la manière suivante :

Définition 3.1 *Un modèle $\mathcal{M}$ du cadre C* confirme *une stratégie de manipulation $s_{i,j}(\phi) \in \mathcal{S}$ d'un agent i envers un autre agent j si, et seulement si, il existe une formule $\psi \in \mathcal{L}_{KBE}$ telle qu'un des deux cas est vérifié :*

— $\forall w \in \mathcal{W} : \mathcal{M}, w \models s_{i,j}(\phi) \Rightarrow M^{+}_{i,j}\psi$

— $\forall w \in \mathcal{W} : \mathcal{M}, w \models s_{i,j}(\phi) \Rightarrow M^{-}_{i,j}\psi$

Hypothèse 3.1 *(Hypothèse de non trivialité de la manipulation) En pratique, un modèle $\mathcal{M}$ ne confirme que rarement une stratégie de manipulation $s_{i,j}(\phi) \in \mathcal{S}$. Par conséquent, nous affirmons que nous faisons* l'hypothèse de non trivialité de la manipulation *lorsque nous considérons $s_{i,j}(\phi)$ comme une* manipulation*. On dit alors que $s_{i,j}(\phi)$ est une* manipulation approchée*.*

Imaginons un modèle $\mathcal{M}$ dans lequel nous déduisons qu'un agent i est en train de mentir à un agent j (voir définition du mensonge dans Section 1.1), et supposons que l'agent i a la volonté que son mensonge soit crédible, c'est-à-dire :

$$\forall w \in \mathcal{W} : \mathcal{M}, w \models K_i\neg\phi \wedge E^{d}_{i} B_j\phi \wedge E^{d}_{i} \neg K_j K_i \neg\phi$$

Par application de l'axiome $(C_{E^{d}_{i}})$, nous déduisons que :

$$\forall w \in \mathcal{W} : \mathcal{M}, w \models K_i\neg\phi \wedge E^{d}_{i}(B_j\phi \wedge \neg K_j K_i \neg\phi)$$

Cependant, nous ne pouvons pas déduire dans ce modèle les intentions à termes de l'agent i d'influencer le comportement de l'agent j afin de déduire la manipulation pure.

En effet, nous pouvons par exemple imaginer qu'à cet instant, la manipulation pure n'a pas encore eu lieu et nous ne déduisons qu'une stratégie et donc une tentative de manipulation. C'est donc pour cette raison que nous ne pouvons pas la confirmer. Ainsi, dans la suite de l'article, nous faisons *l'hypothèse de non trivialité de la manipulation* et réduisons ainsi la manipulation à sa stratégie.

3.3 Stratégies de manipulation

Dans la littérature, un grand nombre de stratégies de manipulation existent, qu'elles reposent sur la rationalité des agents [10], sur les connaissances ou croyances des agents [25], sur les buts ou désirs des agents [18, 34], sur les instincts et émotions [15], sur les normes [8] ou encore sur les confiances que les agents entretiennent entre eux [23]. Dans cet article, nous n'allons nous intéresser qu'aux principales stratégies de manipulation qui portent sur les intentions, les croyances et les connaissances des agents dans le système. La figure 2 représente les différentes catégories, i.e. stratégies de manipulations scindées en deux ensembles de formules représentant les stratégies constructives ou destructives. Cette catégorisation est construite sur l'idée qu'une stratégie de manipulation repose sur deux éléments principaux : les intentions du manipulateur et la nature de la dissimulation. Tous ces ensembles de formules forment une partie de l'ensemble $\mathcal{S}$[3]. Par exemple, lorsqu'un agent i veille de manière délibérée à dissimuler sa véritable croyance pour qu'un agent j veille à réaliser quelque chose, cette stratégie appartient à l'ensemble $\mathcal{S}^{E_j}_{B_i}$:

$$B_i\phi \wedge E^d_i(E_j\phi \wedge \neg K_j B_i\phi) \in \mathcal{S}^{E_j}_{B_i}$$

De la même façon sa manipulation destructive appartient à la même catégorie mais pas au même sous-ensemble car :

$$B_i\phi \wedge E^d_i(\neg E_j\phi \wedge \neg K_j B_i\phi) \in \mathcal{S}^{\bar{E}_j}_{B_i}$$

Remarquons par exemple que les stratégies de manipulation consistant à réaliser une manipulation pure est décrite par la catégorie :

$$\mathcal{S}^{E_j}_{E^d_i} \cup \mathcal{S}^{\bar{E}_j}_{E^d_i}$$

Ainsi, $M^+_{i,j}\phi \in \mathcal{S}^{E_j}_{E^d_i}$ et $M^-_{i,j}\phi \in \mathcal{S}^{\bar{E}_j}_{E^d_i}$.

Dans la suite de cette section et pour des raisons d'espace, nous avons fait le choix de ne présenter que les stratégies de manipulation portant sur la dissimulation de l'instrumentalisation, des croyances et de la méconnaissance.

3. Nous n'affirmons pas que l'ensemble $\mathcal{S}$ est décrit comme l'union de tous ces ensembles de stratégies de manipulation car il existe des stratégies de manipulations $\mathcal{S}$ qui ne peuvent pas être décrites par notre cadre. Nous pouvons mentionner l'exemple des stratégies de manipulation qui reposent sur la confiance entre agents ou encore de celles qui reposent sur les désirs des agents.

Intention / Dissimulation	Faire faire	Faire savoir	Faire croire
Instrumentalisation	$\mathcal{S}^{E_j}_{E^d_i} \cup \mathcal{S}^{\bar{E}_j}_{E^d_i}$	$\mathcal{S}^{K_j}_{E^d_i} \cup \mathcal{S}^{\bar{K}_j}_{E^d_i}$	$\mathcal{S}^{B_j}_{E^d_i} \cup \mathcal{S}^{\bar{B}_j}_{E^d_i}$
Connaissance	$\mathcal{S}^{E_j}_{K_i} \cup \mathcal{S}^{\bar{E}_j}_{K_i}$	$\mathcal{S}^{K_j}_{K_i} \cup \mathcal{S}^{\bar{K}_j}_{K_i}$	$\mathcal{S}^{B_j}_{K_i} \cup \mathcal{S}^{\bar{B}_j}_{K_i}$
Croyance	$\mathcal{S}^{E_j}_{B_i} \cup \mathcal{S}^{\bar{E}_j}_{B_i}$	$\mathcal{S}^{K_j}_{B_i} \cup \mathcal{S}^{\bar{K}_j}_{B_i}$	$\mathcal{S}^{B_j}_{B_i} \cup \mathcal{S}^{\bar{B}_j}_{B_i}$
Non croyance	$\mathcal{S}^{E_j}_{\bar{B}_i} \cup \mathcal{S}^{\bar{E}_j}_{\bar{B}_i}$	$\mathcal{S}^{K_j}_{\bar{B}_i} \cup \mathcal{S}^{\bar{K}_j}_{\bar{B}_i}$	$\mathcal{S}^{B_j}_{\bar{B}_i} \cup \mathcal{S}^{\bar{B}_j}_{\bar{B}_i}$
Méconnaissance	$\mathcal{S}^{E_j}_{\bar{K}_i} \cup \mathcal{S}^{\bar{E}_j}_{\bar{K}_i}$	$\mathcal{S}^{K_j}_{\bar{K}_i} \cup \mathcal{S}^{\bar{K}_j}_{\bar{K}_i}$	$\mathcal{S}^{B_j}_{\bar{K}_i} \cup \mathcal{S}^{\bar{B}_j}_{\bar{K}_i}$

FIGURE 2 – Taxonomie des stratégies de manipulation

Certaines de ces stratégies de manipulation ont été décrites dans l'article de Sakama *et al.* [25] comme le mensonge, la tromperie ou les bullshits. Nous allons donc reprendre et adapter ces stratégies de manipulation afin de les insérer dans notre cadre.

3.3.1 Dissimulation des intentions

Une stratégie de manipulation qui cherche à influencer les croyances d'un agent sans que ce dernier ne s'aperçoive de cette influence est une manipulation de la catégorie $\mathcal{S}^{B_j}_{E^d_i}$ que nous appelons *stratégie de manipulation doxastique* constructive. Ainsi lorsqu'un agent veille à dissimuler son intention de faire croire quelque chose à un autre et se caractérise par le prédicat :

$$M^{B_j}_{E^d_i}\phi = E^d_i(B_j\phi \wedge \neg K_j E^d_i B_j\phi)$$

L'agent manipulateur i effectue une ou plusieurs actions de manière délibérée et qui amènent à ce que j croit quelque chose tout en veillant à lui dissimuler son intention de lui faire croire ϕ. Cette stratégie de manipulation s'illustre par exemple dans la propagation des rumeurs. En effet, lorsqu'un agent i veille de manière délibérée à ce qu'un agent j croit que, par exemple, ϕ :="une certaine personne est mauvaise", l'agent i veille aussi de manière délibérée à ce que l'agent j ne sache pas que l'agent i veille de manière délibérée à ce que l'agent j croit en cette proposition.

La version destructive de cette stratégie de manipulation consiste à veiller de manière délibérée à empêcher quelqu'un de croire quelque chose sans pour autant que cette personne soit au courant de notre intention délibérée pour l'empêcher de croire.

$$M^{\bar{B}_j}_{E^d_i}\phi = E^d_i(\neg B_j\phi \wedge \neg K_j E^d_i \neg B_j\phi)$$

Ceci peut s'illustrer par la dissimulation de menaces non crédibles : un agent i peut veiller de façon délibérée à ce qu'un agent j ne croit pas que "l'agent i n'a aucun moyen de pression sur l'agent j".

Un dernier aspect sur les stratégies de manipulation qui ont pour principe de dissimuler l'instrumentalisation d'un agent sont les *stratégies de manipulation épistémique*. La forme constructive de cette stratégie consiste à veiller à ce que quelqu'un sache une vérité sans pour autant que cette personne soit au courante de l'intention délibérée du manipulateur de lui donner accès à cette vérité.

$$M^{K_j}_{E^d_i}\phi = E^d_i(K_j\phi \wedge \neg K_j E^d_i K_j\phi)$$

Dans une élection, un candidat i veille de manière délibérée à ce que pour tout [4] électeur j, j sache que "le candidat k est un voleur ainsi qu'un tricheur". Ce candidat veille de manière délibérée à ce que tout électeur j ne sache pas son intention délibérée pour ne pas être accusé de délation.

De la même façon sa forme destructive consiste à empêcher un agent d'avoir accès à une certaine vérité ou une fausse connaissance.

$$M^{\bar{K}_j}_{E^d_i}\phi = E^d_i(\neg K_j\phi \wedge \neg K_j E^d_i \neg K_j\phi)$$

Intéressons-nous désormais aux stratégies portant sur la dissimulation des croyances de la figure 2.

3.3.2 Dissimulation des croyances

Le mensonge [25] est par exemple une stratégie de manipulation qui vise à dissimuler ses croyances. En effet, lorsqu'un agent i ment à un autre agent j, l'agent i veille de façon délibérée à ce que l'agent j croit une proposition ϕ alors que l'agent i croit en son contraire, i.e. $B_i\neg\phi$. Cependant, pour que cette stratégie puisse fonctionner, il est aussi nécessaire que l'agent i veille à ce que son mensonge soit crédible, c'est-à-dire que l'agent i veille de façon délibérée à ce que l'agent j ne sache pas que l'agent i croit $\neg\phi$. Ainsi, nous appelons *stratégie du mensonge crédible* une manipulation approchée de $\mathcal{S}^{B_j}_{B_i}$ telle que :

$$lie_{i,j}\phi = B_i\neg\phi \wedge E^d_i(B_j\phi \wedge \neg K_j B_i\neg\phi) \in \mathcal{S}^{B_j}_{B_i}$$

Nous remarquons qu'il est possible de construire beaucoup d'autres formes de stratégies de manipulation de $\mathcal{S}^{B_j}_{B_i}$. Par exemple, nous nommons la manipulation approchée suivant comme la forme *canonique constructive* de $\mathcal{S}^{B_j}_{B_i}$:

$$M^{B_j}_{B_i}\phi = E^d_i(B_j\phi \wedge \neg K_j B_i\phi) \in \mathcal{S}^{B_j}_{B_i}$$

Par exemple, cette stratégie peut correspondre au fait qu'un agent i veille de façon délibérée à ce qu'un autre agent j croit que "l'agent j fait du mauvais travail", tout en veillant de façon délibérée à ce que l'agent j ne sache pas que l'agent i croit la même chose. Une autre version antagoniste à cette stratégie, et que nous appelons forme *non canonique constructive* de $\mathcal{S}^{\bar{B}_j}_{B_i}$, consiste en ce qu'un agent i veille de façon délibérée à ce qu'un agent j croit quelque chose tout en veillant à ce qu'il ne sache pas que l'agent i croit en son contraire. Remarquons que cette forme est en réalité une conséquence logique de la stratégie du mensonge crédible.

En dernier exemple, considérons le cas de la manipulation *canonique destructive* de $\mathcal{S}^{\bar{B}_j}_{B_i}$. Cette stratégie est décrite par le fait qu'un agent i veille de façon délibérée à ce qu'un agent j ne croit pas quelque chose que l'agent i croit, c'est-à-dire :

$$M^{B_j}_{B_i}\phi = E^d_i(\neg B_j\phi \wedge \neg K_j B_i\phi) \in \mathcal{S}^{\bar{B}_j}_{B_i}$$

Cette manipulation approchée vise à empêcher un agent de croire, par exemple, que "la fin du monde est proche" tout en veillant de façon délibérée à ce qu'il ne sache pas que l'agent manipulateur croit justement que la fin est proche. Remarquons que cette forme décrit l'intention délibérée d'un agent à tromper un autre agent.

Pour plus de détails sur les manipulations sur la tromperie, nous pouvons renvoyer le lecteur intéressé à la discussion de Sakama *et al.* dans [25]. Illustrons maintenant la ligne de notre tableau 2 qui porte sur les stratégies de manipulation sur la dissimulation de la méconnaissance.

3.3.3 Dissimulation de la méconnaissance

Une stratégie de manipulation peut considérer en le fait d'annoncer des bêtises, c'est-à-dire des formules dont on ne sait pas la valeur de vérité, afin de tromper un autre agent. Sakama *et al.* [25] définissent ainsi le *baratinage intentionnel* comme le fait qu'un agent i veille de manière délibérée à ce qu'un agent j acquiert une information ϕ alors qu'il n'est pas le cas que l'agent i croit ϕ ni ne croit $\neg\phi$. Dans cet exemple, il s'agit bien de dissimuler de la méconnaissance car, par contraposition sur le lien entre connaissance et croyance,nous déduisons qu'il n'est pas le cas qu'un agent i sait ϕ, ni ne sait $\neg\phi$. Pour que ce baratinage intentionnel soit crédible, il est nécessaire, à l'instar du mensonge crédible, que l'agent i veille de façon délibérée à ce que l'agent j ne sache pas la méconnaissance de l'agent i.

Cependant, contrairement à Sakama *et al.*, nous n'avons pas de modalité d'acquisition d'information décrivant le fait qu'un agent i transmet une information ϕ à un agent j. Nous pourrions être tenté de réduire cette acquisition d'information à une simple intention délibérée de faire croire quelque chose à un agent. Cependant, cela pourrait correspondre à "l'agent i veille de façon délibérée que l'agent j croit cette information ϕ", ou bien encore que "l'agent i

4. Remarquons que nous n'étendons pas notre cadre à la manipulation d'un groupe d'agents mais il semble simple de définir qu'un agent ou groupe d'agents I manipulent un ou plusieurs agents J si, et seulement quelque soit les agents $i \in I$ et quelque soit les agent $j \in J$, l'agent i manipule j.

veille de façon délibérée à ce que l'agent j croit que l'agent i croit ϕ", ou bien encore que "l'agent i veille de façon délibérée à ce que l'agent j croit que i croit que j croit ϕ", et ainsi de suite. Cependant, nous pouvons définir une manipulation approchée de $\mathcal{S}^{B_j}_{\bar{K}_i}$ qui se rapproche du baratinage intentionnel par la formule suivante :

$$\neg K_i\phi \wedge \neg K_i\neg\phi \wedge E^d_i((B_jB_i \dots B_j\phi) \wedge \neg K_j(\neg K_i\phi \wedge \neg K_i\neg\phi)) \in \mathcal{S}^{B_j}_{\bar{K}_i}$$

Il est aisé de remarquer le nombre conséquent de manipulations de $\mathcal{S}^{B_j}_{\bar{K}_i}$ pouvant être exprimées. Toutefois, lorsqu'une telle stratégie de manipulation est réalisée qu'il semble assez naturel d'affirmer que l'agent i veille de façon délibérée à ce que l'agent j ne sache pas les intentions de l'agent i. Ainsi donc, nous ne sommes jamais très loin de la manipulation pure.

4 Conclusion et perspectives

Dans cet article, nous avons proposé un cadre permettant d'exprimer la manipulation comme l'intention délibérée d'un agent d'instrumentaliser un autre agent tout en veillant à ce que ce dernier ne s'aperçoive pas des intentions du premier. Comme le lien logique entre stratégie de manipulation et manipulations n'était pas trivial, nous avons alors émis une hypothèse de non trivialité de la manipulation qui nous a permis de réduire des stratégies de manipulation à des formes à part entière de la manipulation. Cela nous permet d'exprimer beaucoup de stratégies de manipulation comme le mensonge crédible, le baratinage intentionel ou plusieurs formes de dissimulation des intentions. Au-delà d'étendre note cadre pour cartographier plus de manipulations, une première piste de recherche consiste à étendre le cadre aux manipulations s'appuyant sur la confiance des agents [23] ou sur les préférences exprimées des agents [34]. Une seconde piste consistant à caractériser les manipulations de groupes (manipulations communes ou simplement distribuées) nous semble intéressant.

Références

[1] Akopova, Asya Savvichna: *Linguistic Manipulation : Definition and Types*.

[2] Alur, Rajeev, Thomas A Henzinger et Orna Kupferman: *Alternating-time temporal logic*. Dans *Compositionality : The Significant Difference*, pages 23–60. Springer, 1998.

[3] Balbiani, Philippe: *Propositional dynamic logic*. Stanford encyclopedia of philosophy, 2008.

[4] Belnap, Nuel et Michael Perloff: *Seeing to it that : a canonical form for agentives*. Theoria, 54(3) :175–199, 1988.

[5] Blackburn, Patrick, Maarten De Rijke et Yde Venema: *Modal Logic : Graph. Darst*, tome 53. Cambridge University Press, 2002.

[6] Bottazzi, Emanuele et Nicolas Troquard: *On Help and Interpersonal Control*. Dans *The Cognitive Foundations of Group Attitudes and Social Interaction*, pages 1–23. Springer, 2015.

[7] Bursten, Ben: *The manipulative personality*. Archives of general psychiatry, 26(4) :318–321, 1972.

[8] Cialdini, Robert B: *Harnessing the science of persuasion*. Harvard Business Review, 79(9) :72–81, 2001.

[9] Clair, Harvey R St: *Manipulation*. Comprehensive psychiatry, 7(4) :248–258, 1966.

[10] Ettinger, David et Philippe Jehiel: *A theory of deception*.

[11] Gibbard, Allan: *Manipulation of voting schemes : a general result*.

[12] Hamilton, J DeVance, Norman Decker et Ruben D Rumbaut: *The manipulative patient.* American journal of psychotherapy, 1986.

[13] Handelman, Sapir: *Thought manipulation : the use and abuse of psychological trickery*. ABC-CLIO, 2009.

[14] Harel, David, Dexter Kozen et Jerzy Tiuryn: *Dynamic logic*. Dans *Handbook of philosophical logic*, pages 99–217. Springer, 2001.

[15] Joule, Robert Vincent, Jean Léon Beauvois et Jean Claude Deschamps: *Petit traité de manipulation à l'usage des honnêtes gens*. Presses universitaires de Grenoble Grenoble, 1987.

[16] Kligman, Michael et Charles M Culver: *An analysis of interpersonal manipulation*. The Journal of medicine and philosophy, 17(2) :173–197, 1992.

[17] Lamport, Leslie: *Hybrid systems in TLA+*.

[18] Lang, Juan, Matt Spear et Shyhtsun Felix Wu: *Social Manipulation of Online Recommender Systems.* Dans *SocInfo*, pages 125–139. Springer, 2010.

[19] Larousse, Pierre: *Grand dictionnaire universel du XIXe siècle*. Larousse, 1867.

[20] Lorini, Emiliano et Giovanni Sartor: *A STIT Logic for Reasoning About Social Influence*. Studia Logica, 104(4) :773–812, 2016.

[21] Pacuit, Eric: *NEIGHBORHOOD SEMANTICS FOR MODAL LOGIC*. Springer, 2017.

[22] Pörn, Ingmar: *Action theory and social science : Some formal models*, tome 120. Springer Science & Business Media, 2012.

[23] Ruan, Yefeng et Arjan Durresi: *A survey of trust management systems for online social communities – Trust modeling, trust inference and attacks*.

[24] Rudinow, Joel: *Manipulation*. Ethics, 88(4) :338–347, 1978.

[25] Sakama, Chiaki, Martin Caminada et Andreas Herzig: *A formal account of dishonesty*.

[26] Segerberg, Krister, John Jules Meyer et Marcus Kracht: *The logic of action*. 2009.

[27] Specht, Stephen M et Ruby B Lee: *Distributed Denial of Service : Taxonomies of Attacks, Tools, and Countermeasures.* Dans *ISCA PDCS*, pages 543–550, 2004.

[28] Stalnaker, Robert: *On logics of knowledge and belief*. Philosophical studies, 128(1) :169–199, 2006.

[29] Sunstein, Cass R: *Fifty shades of manipulation*. 2015.

[30] Todd, Patrick: *Manipulation*. The international encyclopedia of ethics, 2013.

[31] Torka, Nicole et Stefan Zagelmeyer: *Malignant manipulation at work : a qualitative exploration of strategies and tactics*.

[32] Troquard, Nicolas: *Reasoning about coalitional agency and ability in the logics of "bringing-it-about"*.

[33] Vallée, Thibaut, Grégory Bonnet et François Bourdon: *Multi-armed bandit policies for reputation systems*.

[34] Vallée, Thibaut, Grégory Bonnet, Bruno Zanuttini et François Bourdon: *A study of sybil manipulations in hedonic games*.

[35] Van Ditmarsch, Hans, Wiebe van Der Hoek et Barteld Kooi: *Dynamic epistemic logic*, tome 337. Springer Science & Business Media, 2007.

[36] Van Ditmarsch, Hans, Jan Van Eijck, Floor Sietsma et Yanjing Wang: *On the logic of lying*. Dans *Games, actions and social software*, pages 41–72. Springer, 2012.

[37] Wagner, Alan R et Ronald C Arkin: *Robot deception : recognizing when a robot should deceive*.

[38] Wood, Anthony D et John A Stankovic: *Denial of service in sensor networks*. computer, 35(10) :54–62, 2002.