Formalisation de systèmes d'agent cognitif, de la confiance, et des émotions

Jonathan Ben-Naim, Dominique Longin, Emiliano Lorini

Chapitre 1

Formalisation de systèmes d'agent cognitif, de la confiance, et des émotions

Résumé

Un agent cognitif est un agent dont la conception se fonde sur des propriétés que l'on attribue habituellement aux êtres humains. La cognition est ici vue comme un mécanisme général de gestion intelligente (par opposition à une gestion de type stimuli-réponse) de l'information : acquisition de nouvelles informations en provenance de l'environnement, raisonnement, prise de décision, *etc.* Cet article présente les différents concept couramment utilisés pour formaliser des agents cognitifs, puis aborde la formalisation de deux concepts particuliers en relation avec l'humain : la confiance, et l'émotion. Le formalisme présenté pour les agents cognitifs est exclusivement de type logique car c'est un outil particulièrement adapté pour représenter la connaissance et formaliser le raisonnement. En revanche, si confiance et émotion sont sans contexte des concepts d'ordre cognitif pouvant aisément être formalisés par la logique, nous montrons qu'ils peuvent (et sont) également représentés à l'aide de formalismes numériques.

1.1 Introduction

Définir un agent n'est jamais une tâche aisée tant il existe de multiples manières de le faire à l'aide de notions très différentes, voire parfois antagonistes. Dans ce qui suit, les agents sont définis comme des systèmes informatiques ayant certaines propriétés telles l'autonomie (capacité à agir sans intervention humaine et à contrôler ses actions et ses états internes), la réactivité (capacité à interagir avec d'autres agents — éventuellement humains — *via* un langage de communication), la pro-activation (capacité à adopter

0. Coordinateur : D. Longin ; auteurs : J. Ben-Naïm, D. Longin et E. Lorini (CNRS-IRIT).

un comportement dirigé par un but en prenant l'initiative), *etc.* Comme le résume Wooldridge, cela revient à considérer les agents comme des systèmes informatiques « capables de décider pour eux-mêmes quoi faire dans une situation donnée » (« *deciding for themselves* what to do in any given situation » (Wooldridge, 2000)).

Plus spécifiquement dans le domaine de l'intelligence artificielle (IA), il est très fréquent que ces propriétés soient décrites à l'aide de concepts habituellement associés aux humains : les attitudes mentales (croyance, connaissance, but, désir, intention, *etc.*), les attitudes sociales (engagement, croyance ou intention collective, acceptance, *etc.*), le temps et l'action. Les propriétés elles-mêmes peuvent également être plus spécifiques aux humains. On peut citer notamment la rationalité (dans un sens très large, cela signifie que l'agent n'agit pas de manière contradictoire : il ne croit pas simultanément une chose et son contraire, il agit conformément aux buts qu'il s'est fixé, *etc.*), la sincérité (l'agent ne cherche pas à induire les autres en erreur), *etc.* Ces propriétés dépendent de l'univers dans lequel l'agent va évoluer. Par exemple, est-il opportun de supposer qu'un agent soit sincère s'il doit jouer au poker, ou qu'il ne le soit pas s'il est destiné à communiquer ses observations météo ? La réponse est vraisemblablement « non » dans les deux cas. Ainsi, toutes ces propriétés sont utilisées par les concepteurs de systèmes d'agent pour développer des systèmes particuliers. On parle alors d'agent cognitif pour spécifier le fait que ces derniers soient construits à partir de concepts sensés représenter la cognition humaine.

Dans ce qui suit, nous appelons « système d'agent cognitif » (ou « système cognitif » pour faire court) les systèmes dont le comportement est prévisible selon les attitudes mentales qu'on leur attribue. Le problème est alors de choisir les « bonnes » attitudes mentales pour un (ensemble d')agent(s) donné, c'est-à-dire celles dont on a besoin pour formaliser les propriétés que l'on souhaite attribuer à cet (ensemble d')agent(s). Un avantage certain des systèmes cognitifs est qu'ils permettent de décrire à peu près n'importe quoi, même des objets fonctionnels. Ils sont très utilisés en IA car ils jouissent d'un certain nombre de propriétés intéressantes : ils s'assoient sur une littérature philosophique très importante ; les formalismes associés sont mathématiquement eux-aussi bien connus ; ils se situent à un niveau d'abstraction par nature très élevé (permettant de dissocier la formalisation d'un tel système, d'une part du fonctionnement réel du modèle qui a inspiré cette formalisation et d'autre part de l'implantation de celle-ci en machine) ; enfin, ils ont un fort pouvoir explicatif (l'action accomplie par un agent découlant mathématiquement des propriétés qu'on lui a attribuées et de ses connaissances).

Dans ce qui suit, nous nous attacherons ainsi en premier lieu à la formalisation de systèmes d'agent cognitif (Section 1.2). Un tel agent est supposé être capable de : se représenter l'environnement dans lequel il évolue ; se représenter ce qu'il souhaiterait qu'il fusse ; raisonner à partir de ces représentations afin d'agir pour le transformer. [1] La logique est un outil particulièrement adapté pour ces tâches si on souhaite les réaliser finement et nous nous intéresserons donc dans cette section uniquement aux formalisations logiques d'agents cognitifs, plus précisément aux logiques modales. En relation avec ce qui précède, ces logiques doivent inclure nécessairement trois types d'opérateur : la croyance ou la connaissance (représentation de l'environnement), les

1. Notons que le terme « agent » vient d'ailleurs du verbe latin *agere* qui signifie agir, faire.

désirs, les buts, les préférences, *etc.* (représentation des évolutions souhaitées de cet environnement), l'action et le temps (explicites ou non). [2]

En second lieu, nous présenterons deux concepts particuliers fortement reliés à la cognition (et donc aisément manipulables au sein de systèmes d'agent cognitif) : la confiance (Section 1.3) et l'émotion (Section 1.4). Par nature, ces deux concepts se prêtent bien à être formalisés par la logique pour ce qui est de capturer leur structure cognitive sous-jacente, c'est-à-dire l'état mental dans lequel un agent se trouve nécessairement lorsqu'il a confiance ou éprouve une émotion donnée. En revanche la logique se prête moins bien à la représentation de l'intensité de ces concepts, terrain où les formalismes numériques sont intuitivement plus adaptés. Cela explique pourquoi on trouve à la fois des systèmes logiques et des systèmes numériques traitant de la confiance et de l'émotion, et nous donnerons un aperçu de ces deux types d'approche.

1.2 Systèmes formels d'agent cognitif

1.2.1 Bref historique des systèmes BDI

On peut dire que l'histoire des systèmes formels tels qu'ils sont aujourd'hui est aussi longue que celle de la philosophie qui, depuis Aristote, s'est toujours interrogée à propos d'un certain nombre de concepts : logiques aléthiques (logiques du nécessaire et du possible), logiques épistémiques ou doxastiques (connaissance/savoir, ou croyance), déontiques (obligation, interdiction, permissions), temporelles, conditionnels, dynamique (logiques de l'action, explicite ou non), *etc.*

Il s'agit ici essentiellement de *logiques modales*, c'est-à-dire de logiques comprenant des opérateurs non véri-fonctionnels : ainsi, si $\Box$ est un opérateur modal, alors la formule de la logique modale $\Box\varphi$ (où φ est elle-aussi une formule de la logique) est vraie indépendamment du fait que φ soit vrai ou non. Cette opérateur $\Box$ peut représenter les croyances, les buts, les intentions, *etc.*. Par exemple, si Bel_i *beau* représente le fait que l'agent i croit qu'il fait beau, l'agent i peut avoir cette croyance indépendamment du fait que *beau* (représentant le fait qu'il fasse beau) soit vrai ou non.

Tous ces travaux formels, ainsi qu'un certain nombre d'autres, notamment en philosophie (voir notamment (Searle, 1983), mais surtout (Bratman, 1987)) ont contribué à l'élaboration, fin des années 80 début des années 90, de la logique BDI de Cohen et Levesque où l'intention est définie de manière non primitive à partir des croyances et des buts (Cohen and Levesque, 1990) et où le cadre formel utilisé sert également à caractériser les capacités communicatives de l'agent (Cohen et al., 1990). On peut dire que ces travaux ont été la pierre angulaire des systèmes formels d'agent cognitif [3] en voyant les théories d'agent (notamment celle du langage de ces agents) comme des théories de l'action. [4]

2. C'est pourquoi ces logiques sont habituellement appelées « logiques BDI » (pour *belief*, *desire*, *intention*) dans la littérature. Par analogie, on parle aussi (de système) d'agent BDI.

3. Leur article dans *Artificial Intelligence* a reçu le *AAMAS most influential paper award* en 2008.

4. Cela explique d'ailleurs le succès de la théorie des actes de langage (Austin, 1962; Searle, 1969) au sein de la communauté agent : dans ces théories, le langage est vu comme l'accomplissement d'action, facilitant *de facto* l'union formelle entre actions physiques et linguistiques.

Ces travaux ont été suivis par ceux de Rao et Georgeff qui, reprenant les principes logiques adoptés par Cohen et Levesque, ont cherché un cadre formel plus rigoureux au sein d'une logique temporelle ramifiée dotée d'une axiomatique et d'une sémantique (Rao and Georgeff, 1991). À noter que dans ces travaux l'intention est définie de manière non primitive. Dans la lignée directe de ces travaux, il faut citer ceux de Wooldridge qui introduit la logique LORA (*LOgic of Rational Agent*) dans (Wooldridge, 2000) où le but est non seulement de formaliser une architecture d'agent de type BDI, mais également son évolution dans le temps.

Au niveau national, on peut également signaler les travaux de Sadek (voir sa thèse ou KR'92) qui, au sein d'un cadre formel de la même famille, s'est attaché à définir des règles de rationalité afin de guider le comportement d'un agent rationnel au sein d'un système d'interaction rationnelle. Sa théorie a d'ailleurs directement influencé un langage de communication d'agent (*agent communication language* ou ACL) devenu une référence internationale et ayant [été utilisé dans / donné lieu à] d'innombrables travaux dans la communauté agent : le langage FIPA [5].

Milieu des années 90, des langages plus opérationnels apparaissent, dans le sens où le but n'est plus seulement de disposer d'un formalisme logique permettant de capturer finement les concepts servant à construire le système d'agent visé, mais également de l'implémenter. Ainsi, des systèmes BDI formalisés en *calcul des situations* (*situation calculus* en anglais) font leur apparition (voir par exemple les travaux de Shapiro, Lespérance & Levesque à Toronto). De véritables langages de programmation basés sur des primitives de type BDI apparaissent également : on peut citer par exemple GOLOG ou ConGolog. Cette communauté a donné lieu à ce qu'il convient d'appeler aujourd'hui la robotique cognitive, dont le laboratoire de même nom à Toronto est le représentant le plus illustre.

Des formalismes se sont également attachés à décrire des systèmes normatifs, c'est-à-dire des systèmes où l'agent doit considérer non seulement ce qu'il croit (ou sait) et ce qu'il a pour but de réaliser, mais également ce qu'il est obligé de réaliser. Cet aspect utilise (et hérite des questions théoriques de) la logique déontique. À titre d'exemple, on peut citer l'architecture BOID (où O représente la composante obligation du système BDI) de van der Torre *et al.* (voir par exemple l'article paru à AGENTS'01).

De là on observe un glissement : les systèmes BDI ne manipulent plus seulement des attitudes mentales (en plus du temps et/ou de l'action), mais également des concepts sociaux ou des contraintes extérieures. L'obligation peut être vue soit comme une norme internalisée (elle est alors formalisée à l'aide d'un opérateur indicé par un (groupe d')agent(s)), soit comme un loi externe à laquelle tous les agents du système modélisé doivent se plier. (Elle est alors représentée par un opérateur non indicé par un agent.)

Vers la fin des années 90, les systèmes BDI tels qu'ils sont formalisés alors sont beaucoup critiqués par la communauté agent car ils induisent des hypothèses fortes sur les états mentaux, notamment des hypothèses de sincérité. Ainsi, dans FIPA par exemple, un agent croit tout ce que lui dit un autre agent car il suppose toujours que ce dernier a dit la vérité. Pour éviter ce problème, des travaux décrivent l'effet d'une action linguistique en séparant ce que le locuteur veut signifier de ce que l'auditeur croit sur la base d'hypothèse faites par ce dernier à propos de la sincérité et de la com-

5. http://www.fipa.org/repository/aclspecs.html

pétence du premier. D'autres travaux ont cherché des concepts alternatifs permettant de se dégager de ces hypothèses sur les états internes des agents. Il y a eu par exemple une quantité importante de travaux sur l'engagement social (*commitment*) destiné à capturer ce sur quoi une personne s'engage publiquement lorsqu'elle dit quelque chose. Par exemple, quand quelqu'un affirme quelque chose, il s'engage sur la valeur de vérité de cette proposition : il ne pourra pas dire qu'il ne l'a pas dit, et ne peut agir ou dire quelque chose qui le mettrait en porte-à-faux par rapport à ce qu'il a dit. (Voir à ce sujet les travaux de Singh (Singh, 1999) et de Colombetti en Suisse.) Néanmoins, ces approches ont elles-mêmes leurs défauts : d'autres hypothèses sont bien souvent prises (par exemple, le caractère public des actions linguistiques et le fait qu'elles soient correctement interprétées par leurs destinataires) et il n'est pas évident que ce concept soit sans lien avec les états mentaux de l'agent contractant un tel engagement. Enfin, cette notion n'a pas été analysée de manière satisfaisante en tant que concept non primitif[6], bien qu'elle contienne visiblement une composante normative et conventionnelle ainsi que des conditions de violation. Dans ces circonstances, ces approches sont à la limite des systèmes d'agent BDI puisqu'ils ne font pas intervenir d'état mentaux : le lien existe intuitivement mais il reste à l'établir formellement.

D'autres concepts traditionnels se sont également retrouvés confrontés à ce problème, telle la croyance commune. Celle-ci est définie habituellement comme la conjonction infinie est croyances alternatives entre agents. Par exemple, s'il y a croyance commune entre les agents i et j à propos de φ, alors c'est que : i croit φ, j croit φ, i croit que j croit φ, j croit que i croit φ, i croit que j croit que i croit φ, *etc.* Le problème est donc, dans un système implémenté, d'arriver à établir s'il y a croyance commune sans accéder à « ce qu'il y a dans la tête des agents ». Au mieux, peut-on construire une croyance commune subjective, c'est-à-dire la croyance d'un agent selon lequel il y aurait croyance commune (mais peut-être n'est-ce pas le cas). De nombreux travaux en philosophie ont tourné autour de cette question (voir par exemple (Gilbert, 1989)) pour aboutir à des notions comme l'acceptance (voir (Lorini et al., 2009) par exemple).

Parallèlement à tout cela, des problèmes que l'IA se posait déjà depuis de nombreuses années ont été transférés dans le cadre BDI et ont donné lieu à une abondante littérature : le *frame problem* (comment décrire l'environnement de manière économique et exhaustive), le problème de la caractérisation des actions (quelles sont toutes les conditions nécessaires et suffisantes pour accomplir une action donnée), les problèmes de révision (comme faire évoluer les connaissances d'un agent à travers le temps) et de ramification des actions (comment décrire l'impact d'une action sur le domaine, y compris les états mentaux des agents). Par exemple, avec l'arrivée des systèmes BDI s'est ensuite posé le problème de la révision de ces attitudes mentale (voir (van der Hoek et al., 2007) par exemple).

Plus récemment, ce problème est devenu le cœur d'une branche du domaine : celle des logiques épistémiques dynamiques (voir plus loin). Simplement, leur but est d'intégrer dans la sémantique même de la logique le fait que les croyances (ou les connaissances) d'un agent peuvent évoluer : celui-ci peut changer d'avis, apprendre que certaines propositions sont vraies, que certaines autres sont fausses, *etc.* Au prix de certaines contraintes techniques, les logiques des annonces publiques donnent à

6. C'est-à-dire un concept décrit à l'aide de concepts de plus bas niveau.

l'épineux problème du changement des états mentaux une réponse adéquate. À titre de synthèse sur ce sujet, voir (van Ditmarsch et al., 2007).

Enfin, des plates-formes d'agent se sont développées comme par exemple AgentSpeak par Rao, Jason par Hübner & Bordini ou 2APL par Dastani. Ces plates-formes permettent l'implantation d'agents et de systèmes multi-agents mais n'utilisent pas jusqu'à présent tout le pouvoir expressif des logiques BDI. En particulier, elles ne disposent pas d'un ensemble complet d'opérateurs booléens et ne font pas appel à des démonstrateurs de théorèmes (qui, par ailleurs, existent déjà pour des (familles) de logiques bien connues).

Les concepts proposés dans le domaine des systèmes BDI ont été aussi utilisés dans d'autres domaines de l'IA comme par exemple celui de l'argumentation. À titre d'exemple, on peut citer les travaux de Amgoud sur l'utilisation des méthodes de l'argumentation pour la génération de désirs et de plans dans un agent autonome (Amgoud and Rahwan, 2006).

1.2.2 Concepts de base

Dans ce qui suit, nous présentons les concepts de base généralement utilisés pour la formalisation d'un (système d') agent cognitif. Tous les systèmes n'utilisent bien sûr pas tous les concepts simultanément car, comme nous l'avons dit en introduction, la manière de caractériser un (système d')agent dépend du domaine auquel cet agent ou ce système est destiné.

Dès lors que l'on a besoin d'imbriquer les opérateurs, les logiques modales sont particulièrement adaptées, car une formule de la logique modale dans la portée d'un opérateur modal forme une nouvelle formule modale. On peut ainsi imbriquer naturellement une infinité d'opérateurs modaux les uns dans les autres tout en conservant une formule du langage objet. C'est une propriété particulièrement importante dans le domaine de la cognition car on peut avoir des croyances, par exemple, sur à peu près n'importe quoi, y compris d'autres croyances : l'agent i peut croire que l'agent j croit que l'agent k croit que l'agent i croit p, *etc.*

Les opérateurs de croyance

La notion de croyance a été étudiée en profondeur dans le domaine des logiques doxastiques et épistémiques, et ce depuis le début des années 60 (voir (Gochet and Gribomont, 2006) pour un article complet sur ce sujet). C'est très certainement une des notions les plus étudiées en logique, sous toutes ses formes (logique classique, logique modale, avec degrés ou sans représentant généralement le degré de force des croyances ou connaissances de l'agent [7]).

Une logique communément utilisée est la logique modale propositionnelle sans degré où « l'agent i croit que φ est vrai » est noté $Bel_i\,\varphi$, où Bel_i (pour tout i appartenant à l'ensemble des agents) est un opérateur appelé « opérateur modal de croyance de l'agent i » et où φ est une formule quelconque. Traditionnellement, la vérité de $Bel_i\,\varphi$

7. Dans ce travail nous ne considérons que les approches qualitatives de la croyance. Nous ne discutons pas des approches quantitatives qui formalisent la notion de degré de croyance (voir par example (Laverny and Lang, 2005)).

dans un certain monde w_0 est interprétée comme le fait que φ soit vrai dans tous les mondes que l'agent i envisage comme possibles depuis ce monde w_0, et ce sans que cet agent ne puisse distinguer lequel de ces mondes est le monde réel, et sans même avoir la certitude que le monde réel appartienne à cet ensemble de mondes épistémiques. (Il peut se tromper.) Le substrat sémantique de cela est une famille de relations indexées par chacun des agents appartenant à l'ensemble de tous les agents. Ainsi, dire que l'agent i croit que φ est vrai dans le monde réel w_0 s'écrit : $w_0 \Vdash \mathit{Bel}_i\,\varphi$. Sémantiquement, cela signifie que φ est vrai dans tous les mondes accessibles depuis w_0 *via* la relation représentant la croyance de l'agent i et notée $\mathscr{B}_i$.

Il y a un large consensus dans la littérature qui s'accorde à dire que la logique de la croyance est le système modal normal de type KD45 (Chellas, 1980), même si cette logique constitue une *idéalisation* de certains principes. Par exemple, cette logique entraîne qu'un agent connaît toutes les croyances qui découlent de ses croyances et ce, instantanément (omniscience), et qu'il est conscient de tout ce qu'il croit (introspection positive) et de tout ce qu'il ne croit pas (introspection négative). Ce sont toutefois des critiques qu'il faut tempérer par le fait qu'elles constituent des idéalisations (et non des propriétés aberrantes) qui, au niveau d'un agent artificiel, ne sont pas nécessairement contre-intuitives.

FIGURE 1 illustre la sémantique de l'opérateur de croyance de l'agent i. L'ensemble des mondes accessibles depuis le monde w_0 est noté $\mathscr{B}_i(w_0)$ où $\mathscr{B}_i$ est la relation d'accessibilité aux mondes épistémiques pour l'agent i et est représentée graphiquement par les flèches.

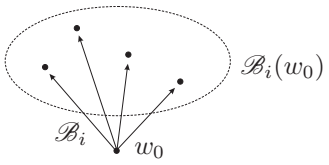

Figure 1 – *Sémantique de Kripke de l'opérateur* Bel_i

Les opérateurs temporels

Il y a de nombreuses logiques temporelles, selon que l'on souhaite une représentation du temps ramifié ou linéaire, avec des indices temporels explicites ou non, *etc.* Les logiques temporelles sont relativement bien étudiées en logique modale et en informatique théorique (van Benthem, 1991). Leur sémantique est basée sur des relations de transition entre des états possibles et sont donc assimilables à des automates (potentiellement infinis).

Nous choisissons ici une des notions les plus simples : le temps linéaire. Comme cette notion est à considérer de concert avec les croyances de l'agent, cela signifie que ces dernières portent en fait non pas sur des mondes épistémiques mais sur des ensembles de mondes ordonnés linéairement dans le temps appelés des « histoires ». Cela permet ainsi de simuler un temps arborescent puisque, épistémiquement, chaque histoire possible correspond à un déroulement (que l'agent croit possible) des événements dans le futur.

Par exemple, FIGURE 2 représente les quatre histoires crues par l'agent i. Les points situés sur les histoires indiquent le présent, et les tirets les instants passés ou futurs. Ainsi : l'agent croit actuellement que p est vrai ($Bel_i\, p$) ; il envisage le fait que r soit actuellement vrai et devienne faux par la suite ($\neg Bel_i\, \neg(r \wedge F\neg r)$) ; *etc.*

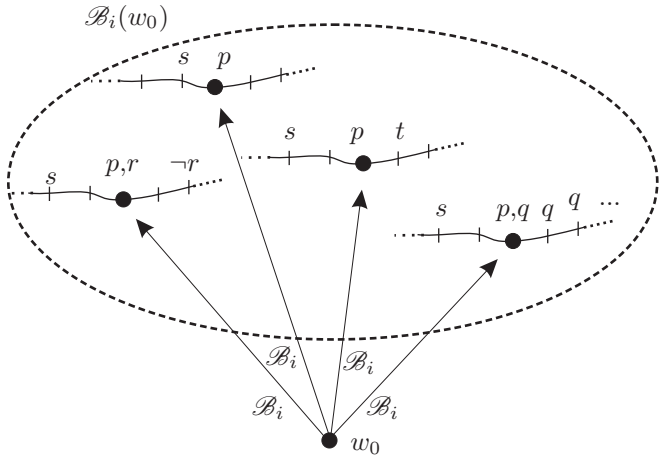

Figure 2 – *Temps linéaire et mondes épistémiques*

De façon symétrique par rapport aux opérateurs G et F, nous pouvons définir les opérateurs H et P parlant du passé.

Techniquement, le temps est défini dans une logique modale du temps linéaire de type S4.3$_t$ (*cf.* (Burgess, 2002) pour plus de détails). Néanmoins, ces opérateurs peuvent être définis sémantiquement selon une structure arborescente. (C'est d'ailleurs de choix qui est fait par dans (Rao and Georgeff, 1991).)

Enfin, nous utilisons parfois également deux opérateurs X et X^{-1} tels que $X\varphi$ signifie « φ sera vrai l'instant juste après maintenant dans l'histoire considérée » et $X^{-1}\varphi$ signifie « φ était vrai l'instant juste avant maintenant dans l'histoire considérée ». Il existe évidemment des liens formels entre ces opérateurs et les opérateurs temporels précédents.

Les opérateurs de but

La notion de but a été étudiée très largement dans la littérature, et a été employée dans des sens très différents (voir par exemple les notions de *goal* chez Cohen et Levesque (Cohen and Levesque, 1990) ou Rao & Georgeff (voir (Rao and Georgeff, 1991) par exemple), ou de choix chez Sadek (voir sa thèse ou KR'92). Ici, nous donnons à la notion de but la définition de *but choisi* (ou *préféré*), c.-à-d. le sous-ensemble cohérent des propositions dont on souhaite actuellement qu'elles soient vraies. Les opérateurs primitifs de but associés sont notés $Choice_i$ (où i appartient à l'ensemble des agents) tel que $Choice_i\, \varphi$ signifie « l'agent i choisit présentement que le but φ soit actuellement vrai ». Ce choix porte sur n'importe quel φ, donc en particulier sur des propositions pouvant représenter l'état de chose présent. C'est la distinction que nous introduisons

ici par rapport aux opérateurs de but à accomplir (abandonnés quand l'état de chose visé est réalisé) ou à maintenir (où on cherche à ce que l'état de chose visé soit atteint en permanence). De façon similaire à la croyance, nous interprétons la vérité de $Choice_i\,\varphi$ dans un monde w_0 comme le fait que φ soit vrai dans tous les mondes que l'agent préfère dans le monde actuel w_0. En toute généralité, les buts sont des pré-ordres partiels, mais nous ne considérons pas ici cet aspect afin de rendre les choses plus simples : nous focalisons donc sur des buts binaires non ordonnés et non contradictoires.

Une question difficile et peu étudiée est : comment naissent ces buts ? Du point de vue cognitif, il semble que cela soit au travers d'un processus délibératif portant sur des attitudes motivationnelles plus primitives appelées « désirs » ou des attitudes morales telles que les idéaux et les impératifs (voir (Rao and Georgeff, 1991; Conte and Castelfranchi, 1995; Castelfranchi and Paglieri, 2007) à ce sujet). L'ensemble des buts que nous caractérisons est celui issus d'un processus de filtrage des idéaux et des désirs destiné à résoudre les conflits entre ces deux concepts et à éliminer les cas impossibles. Les buts choisis d'un agent satisfont alors deux principes de rationalité fondamentaux : ils doivent être consistants (*i.e.*, un agent rationnel ne peut choisir simultanément deux buts contradictoires) ; les buts choisis doivent être nécessairement en relation avec les croyances de l'agent qui les choisit. Dans (Cohen and Levesque, 1990), la relation entre croyances et buts est une relation d'inclusion : si un agent croit actuellement que φ est vrai alors nécessairement il a actuellement pour but que φ. (Cette relation est nommée *réalisme fort*.) On peut également imposer une relation dite de *réalisme faible* où il est seulement requis que les mondes épistémiquement possibles et ceux préférés aient une intersection non vide.

Les idéaux

Il existe de très nombreux systèmes normatifs en logique avec des caractéristiques très différentes, plus ou moins complexes, adaptés à une classe de problèmes plutôt qu'à une autre. Ces normes peuvent avoir des origines différentes : la loi proprement dite, le règlement de telle ou telle structure au sein de laquelle l'agent évolue (qui s'ajoute à la loi), la morale (religieuse ou non), *etc.*

Certaines normes particulières propres à un agent donné sont appelées des idéaux. Nous introduisons un nouvel ensemble d'opérateurs tels que $Idl_i\,\varphi$ signifie : « φ est un état idéal pour l'agent i ». Cela signifie que i s'adresse un ordre à lui-même : une sorte d'impératif à réaliser φ (quand φ est actuellement faux) ou à le maintenir vrai (quand il l'est déjà) (Castaneda, 1975).

Il y a différentes manières d'expliquer comment un état φ devient un état idéal pour un certain agent. Une explication plausible est basée sur l'hypothèse que les idéaux sont juste des normes sociales internalisées (ou adoptées) par cet agent (Conte and Castelfranchi, 1995). Supposons qu'un agent croie que dans un certain groupe (ou institution) il existe une certaine norme (par exemple une obligation) prescrivant qu'un état φ doive être vrai, alors même qu'il s'identifie comme membre de ce groupe. Dans ce cas, l'agent adopte cette norme externe (qui ne provient pas de lui et qui n'a pas encore été approuvée ni reconnue en tant que telle par l'agent) qui devient alors un idéal de l'agent. Par exemple, si l'agent i croit qu'en France il est obligatoire de payer ses impôts et qu'il s'identifie comme citoyen français, il adopte cette obligation

en s'imposant de payer ses impôts.

Il n'y a pas de relation particulière avec les autres opérateurs, sauf avec la croyance, si l'on suppose qu'un agent est conscient de ses idéaux.

L'action explicite

Quand on cherche à définir ce que signifie « l'agent i est capable de faire l'action α », nous nous devons de regarder du côté des logiques de l'action. D'une manière générale, ces logiques modélisent les actions en terme de systèmes de transitions entre des états. Il y a principalement deux courants, l'un où l'action est explicite et l'autre où elle est implicite (*cf.* section suivante).

La principale logique de l'action explicite est la logique dynamique propositionnelle (PDL) qui étudie l'interaction entre une action et ses effets (Harel et al., 2000). Il a été démontré (par exemple dans (van Linder et al., 1998)) que la logique dynamique est particulièrement adaptée à la caractérisation de concepts de capacité et de pouvoir. Il y a une littérature très abondante sur l'intégration de la logique dynamique avec les logiques de la croyance et du but (voir par exemple la logique épistémique dynamique (Baltag and Moss, 2004) ou la logique dynamique doxastique (Segerberg, 1992, 1995)).

PDL distingue des actions telles que α de formules telles que φ et ψ, et son ensemble de constantes non logiques est construit à partir de ces deux catégories distinctes. La formule $\mathit{After}_\alpha\,\varphi$ exprime que φ sera vrai après toute exécution possible de l'action α. Ainsi, $\mathit{After}_\alpha\,\bot$ exprime le fait que α est inexécutable. [8]

Plusieurs extensions ont été proposées dans lesquelles un agent est ajouté comme argument aux opérateurs de PDL. Dans de telles extension, la formule $\mathit{After}_{i:\alpha}\,\varphi$ exprime le fait que φ est vrai après toute exécution possible de l'action α par l'agent i. Pour toute action α et tout agent i, $\mathit{After}_{i:\alpha}$ est un opérateur modal d'action.

D'un point de vu sémantique, l'action est traitée ici comme la transition d'un monde réel vers un ensemble d'autres mondes réels. (Des contraintes sémantiques supplémentaires peuvent faire en sorte que cet ensemble se réduise à un singleton.) La FIGURE 3 représente cette transition.

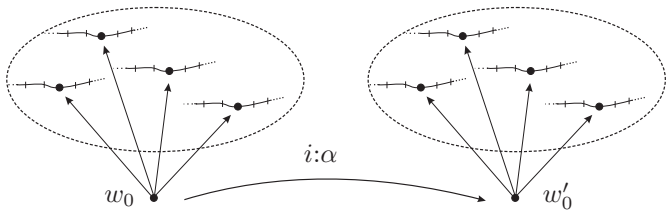

Figure 3 – *Transition du monde w_0 vers le monde w'_0* via *l'execution de l'action $i{:}\alpha$*

À DEON'2008, Lorini & Demolombe ont enrichi le langage de PDL avec des opérateurs $\mathit{Does}_{i:\alpha}$ où la formule $\mathit{Does}_{i:\alpha}\,\varphi$ se lit « L'agent i va faire l'action α après quoi φ sera vrai ». Cela permet de parler à la fois de ce qu'un agent peut faire ($\neg\mathit{After}_{i:\alpha}\,\bot$) et de ce qu'un agent fait ($\mathit{Does}_{i:\alpha}\,\top$).

8. En dehors des logiques BDI, l'opérateur After_α se note souvent $[\alpha]$.

L'action implicite

L'action demeure implicite dans les « logiques de l'agentitude » (*logics of agency*) qui étudient l'interaction entre un agent et les effets que cet agent fait en sorte de provoquer. La particularité de ces logiques est qu'elles s'abstraient des actions qui provoquent ces effets (seul leurs résultats comptent).

Par exemple, dans la logique *STIT* (Belnap et al., 2001), les actions sont identifiées par des formules induisant un agent et parlant des effets que cet agent fait en sorte de provoquer. Ainsi, l'action décrite dans « i achète le produit p » est identifiée avec les formules de l'agentitude « i fait en sorte que le produit p soit acheté par l'agent i ».

Les formules de l'agentitude sont de la forme $STIT_i\,\varphi$, ce qui se lit : « l'action présentement choisie par l'agent i assure le fait que φ est vrai, et ce quoi que fassent les autres agents ». En raccourci, « i fait en sorte que φ ». L'opérateur modal $STIT_i$ est appelé *operateur d'agentitude*.

Dynamique des états mentaux

Ces dernières années un certain nombre de chercheurs travaillant dans le domaine des logiques pour la modélisation des agents autonomes et des systèmes multi-agents ont proposé des logiques pour la représentation de la dynamique des états mentaux. Ces logiques font partie de la grande famille des logiques épistémiques dynamiques (ou *Dynamic Epistemic Logic* (DEL) ; voir par exemple (Ditmarsch et al., 2007)). DEL est un terme chapeau utilisé pour indiquer les extensions dynamiques de la logique de la croyance et de la connaissance, mais aussi de la logique des préférences et des normes (logique déontique) (Baltag and Moss, 2004; Kooi, 2007; van Benthem and Liu, 2007). Dans ces logiques, des opérateurs modaux sont introduits afin de décrire les effets sur les états mentaux de différents types d'événement informatif (transmission de messages publics ou privés, commandes, *etc.*).

Ici nous considérons la logique épistémique dynamique la plus connue : la logique des annonces publiques ou *Public Announcement Logic* (PAL) (Ditmarsch et al., 2007). De façon informelle, nous pouvons dire qu'un fait p est annoncé publiquement si et seulement si : chaque agent apprend que p est vrai ; chaque agent apprend que chaque agent apprend que p est vrai ; chaque agent apprend que chaque agent apprend que chaque agent apprend que p est vrai, *etc.* jusqu'à l'infini. Dans la logique PAL, les annonces publiques sont des événements qui mettent à jour les croyances et les connaissances des agents : la fonction d'une annonce publique est de restreindre l'ensemble des mondes possibles aux mondes dans lesquels le fait publiquement annoncé est vrai et de restreindre les relations d'accessibilité épistémiques à ces mondes. PAL utilise la notation $p!$ pour l'annonce publique de p et introduit des opérateurs modaux de la forme $[p!]$ pour décrire les effets d'une annonce publique sur les croyances des agents : la formule $[p!]q$ signifie que q sera vrai après l'annonce publique de p. Nous présentons ici un exemple pour illustrer le fonctionnement de ces opérateurs dynamiques.

Marie, Paul et Alice sont assis autour d'une table sur laquelle sont posées trois cartes. Sur la face non visible de chaque carte est écrit un numéro compris entre 1 et 3 différent de celui marqué sur les deux autres. Ainsi, chaque carte est identifiée par un numéro entre 1 et 3 (carte 1, carte 2, carte 3). Marie, Paul et Alice prennent une carte

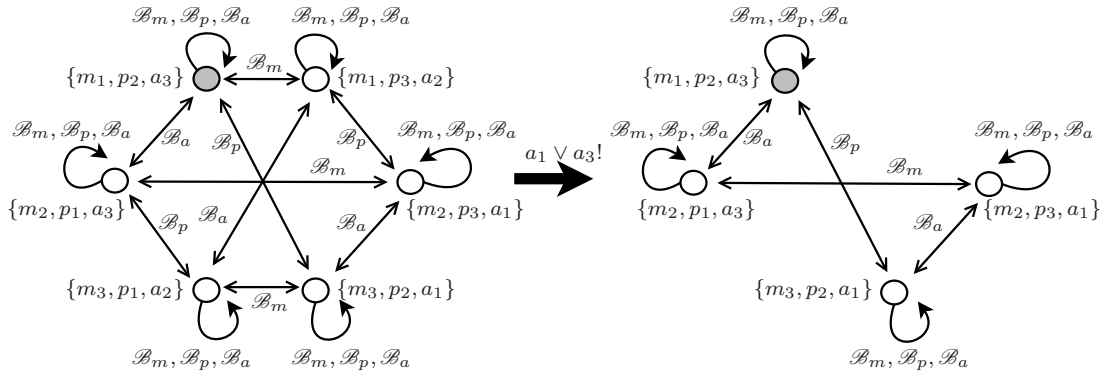

Figure 4 – *Exemple des cartes*

chacun et on suppose que Marie prend la carte 1 (que l'on notera m_1), Paul prend la carte 2 (noté p_2) et Alice prend la carte 3 (a_3). Chaque joueur regarde sa carte sans que les autres joueurs ne la voient et la repose face cachée sur la table. Chaque joueur ne connaît donc que sa carte.

Dans FIGURE 4 le modèle à gauche de la flèche représente les croyances de Marie, de Paul et d'Alice dans la situation initiale venant d'être décrite. Il y a six mondes possibles et celui en gris est le monde réel. Les flèches qui partent de chaque monde représentent les relations d'accessibilité $\mathscr{B}$ aux mondes épistémiques pour chaque joueur. Par exemple, dans le monde réel Marie envisage comme possible le monde dans lequel Marie a la carte 1, Paul a la carte 2 et Alice a la carte 3 et le monde dans lequel Marie a la carte 1, Paul a la carte 3 et Alice a la carte 2. Donc, dans le monde réel Marie est incertaine de la distribution des cartes.

Supposons qu'il soit annoncé publiquement que Alice a une carte avec un numéro impair. Cette annonce est représentée par l'événement $a_1 \vee a_3!$ (Alice a la carte 1 ou bien Alice a la carte 3). FIGURE 4, le modèle à droite de la flèche représente les croyances de Marie, Paul et Alice après cette annonce. Grâce à cette dernière, Marie apprend que Paul a la carte 2 et que Alice a la carte 3. En fait, l'effet de l'annonce publique est de restreindre l'ensemble des mondes possibles aux mondes dans lesquels Alice a une carte avec un numéro impair et de restreindre les relations d'accessibilité à ces mondes. Donc, dans le monde réel après l'annonce, Marie connaît la distribution des cartes : Marie a la carte 1, Paul a la carte 2 et Alice a la carte 3, et Marie le croit. Ce dernier fait est représenté par la formule $m_1 \wedge p_2 \wedge a_3 \wedge Bel_m\,(m_1 \wedge p_2 \wedge a_3)$ qui est vraie dans le mondé réel du modèle de droite. Au contraire, l'annonce publique ne fait rien apprendre à Paul et Alice : après l'annonce publique, Paul et Alice restent incertain sur la distribution des cartes.

Dans ce qui précède, nous venons de donner un aperçu des différents concepts relatifs aux systèmes d'agent cognitif et une façon de les formaliser. Dans ce qui suit, nous présentons deux concepts complexes particuliers pouvant être décrits en termes d'états mentaux, de temps, et d'action. Les systèmes d'agent cognitif sont donc parti-

culièrement adaptés pour capturer ces concepts. Néanmoins, ils font également l'objet d'une mise en œuvre au sein de formalismes numériques dont nous donnons également un aperçu.

1.3 Formalisation de la confiance

Les systèmes de confiance sont des outils utilisés dans certains systèmes multi-agents du domaine de l'intelligence artificielle pour aider les utilisateurs à prendre des décisions alors même que les agents constituant ces systèmes peuvent être incompétents ou mal intentionnés. Par conséquent, un utilisateur a besoin d'une évaluation de ces agents par rapport à certains aspects avant d'interagir avec eux. Un système de confiance est un système qui construit, pour un utilisateur u donné, une telle évaluation. Typiquement, cette évaluation est obtenue sur la base de deux types d'information :

- les résultats des interactions antérieures entre u et les agents ;
- les recommandations que certains agents fournissent à propos d'autres agents.

La position d'un agent a dans une telle évaluation peut être naturellement vue comme la confiance que u peut rationnellement accorder à a par rapport à certains aspects.

Concernant les applications, elles sont de plus en plus nombreuses et déployées pour de très grands nombres d'utilisateurs : e-commerce (Ebay, Amazon, *etc.*), grands wikis (Wikipedia, Planetmath, *etc.*), réseaux sociaux (Facebook, Myspace, *etc.*), pages web et liens hypertext, articles et citations, *etc.*

Typiquement, l'introduction de la confiance dans des systèmes d'agent a permit le développement de nouveaux modèles, l'exhibition de propriétés intuitivement désirables que ces modèles satisfont (ce qui leur fournit des justifications théoriques), l'obtention d'un bon comportement de ces modèles dans des plateformes de test (ce qui leur fournit une validation expérimentale).

À notre connaissance, il existe principalement deux types de modèles de systèmes de confiance : les modèles logiques (essentiellement à base de logiques modales) et les modèles numériques (essentiellement à base de probabilités). Un résumé des différents modèles de la confiance qui ont été développés à ce jour peut être trouvé dans (Sabater and Sierra, 2005) par exemple.

1.3.1 Modèles logiques des systèmes de confiance

Dans l'approche logique, le but n'est pas tant d'obtenir un système d'agent basé sur la confiance que de tenter de caractériser ce qu'est la confiance, ce que cela signifie de faire confiance à quelqu'un, et quel est l'état mental d'un agent lorsqu'il a confiance en quelqu'un.

L'un des modèles théoriques les plus importants de la confiance est le modèle cognitif de confiance de Castelfranchi et Falcone (que nous désignons par la suite comme étant le modèle C&F) (Castelfranchi and Tan, 2001). Contrairement à des approches plus computationnelles (voir paragraphe précédent), le modèle C&F ne se réduit pas plus ou moins à des probabilité subjectives mises à jour à la lumière d'interactions directes avec le *trustee* (la personne en qui on a confiance) et d'information à propos de sa réputation. De manière informelle, le modèle C&F définie la confiance comme une

croyance individuelle du *truster* (ce qui fait ou cherche à faire confiance au *trustee*) à propos de propriétés (telles : la capacité, l'intention, la disposition, *etc.*) que le premier attribue au second.

En accord avec la théorie C&F, et selon l'analyse menée dans (Herzig et al., 2010), la confiance est construite autour de quatre composants : un *truster* i, un *trustee* j, une action α de j, et un but φ de i. Selon leur définition, « i fait confiance à j pour exécuter α afin de réaliser le but φ » si et seulement : i a le but φ ; i croit que j est capable de faire α ; i croit que j, en faisant α, va rendre φ vrai ; i croit que j a l'intention de faire α. Par exemple, quand i fait confiance à j pour lui envoyer un certain produit p dans le but de posséder p, alors : i a le but de posséder p, et il croit que j est capable de lui envoyer p, que le fait que j lui envoie p réalisera son but de posséder p, et que j a bel et bien l'intention de le lui envoyer.

En d'autres termes, la confiance peut être définie formellement de la manière suivante :

$$Trust(i, j, \alpha, \varphi) \stackrel{def}{=} Goal_i\, \varphi \wedge Bel_i\, (Capable_j(\alpha) \wedge After_{j:\alpha}\, \varphi \wedge Intend_j\, (\alpha))$$

où chacun des opérateurs utilisés ci-dessus doit être un opérateur de base ou correspondre à un opérateur complexe définissable à partir des opérateurs de base (*cf.* Section 1.2.2) :

- $Goal_i\, \varphi \stackrel{def}{=} Choice_i\, F\varphi$ signifie que « l'agent i choisit présentement que le but $F\varphi$ soit actuellement vrai » ;
- $Capable_j(\alpha) \stackrel{def}{=} \neg After_{j:\alpha}\, \bot$ signifie que l'agent j est capable d'accomplir l'action α si et seulement si cette action est d'ores et déjà exécutable. [9]
- $Intend_j\, (\alpha) \stackrel{def}{=} Choice_j\, Does_{j:\alpha}\, \top$ signifie l'agent i a l'intention de faire l'action α si et seulement si il a pour but choisi que α soit exécutée (ici et maintenant).

Il existe tout un domaine où la confiance n'est pas décrite en termes de propriétés logiques mais numériques.

1.3.2 Modèles numériques des systèmes de confiance

Dans ce qui précède, la confiance était vue comme une croyance particulière portant sur un certain nombre de propriétés. En fonction du fait qu'il faisait confiance ou non à un agent j à propos d'une certaine proposition φ, un agent i était alors en mesure de juger s'il devait ou non croire ce que j lui disait concernant φ. Il en est de même ici, mais il s'agit maintenant de se demander comment représenter de manière numérique la confiance qu'un agent i peut placer en un agent j. Différents formats de représentation ont été étudiés. Par exemple, la confiance peut être représentée par un nombre, un intervalle, ou même un intervalle flou.

Dans de nombreuses approches, la confiance est simplement représentée par un nombre. Une des premières approches de ce type est (Marsh, 1994). Un des plus importants systèmes de ce type est Pagerank (Page et al., 1998), le système de confiance

9. On pourrait être tenté d'objecter que cela devrait être une condition suffisante, mais non nécessaire. En effet, il suffit que l'agent i croie que l'agent j sera en mesure d'exécuter α « dans les délais qu'on s'est donné pour obtenir φ ». Il faut néanmoins se rappeler que nous formalisons ici une notion de confiance ici et maintenant, pas une confiance potentielle.

à la base du moteur de recherche Google. Une page web peut être vue comme un agent et un lien hypertext d'une page x vers une page y peut être vu comme une recommandation de x en faveur de y. Pagerank associe un nombre réel compris entre 0 et 1 à chaque page web x. Ce nombre peut être vu comme la confiance qu'un utilisateur peut placer dans les informations contenues dans x.

Plus précisément, Pagerank est basé sur une certaine procédure dans laquelle chaque page web possède une certaine quantité de crédit. À chaque étape, chaque page transfère une fraction de ses crédits à chaque autre page. Le point crucial est que si une page x recommande une page y, alors la quantité de crédit transférée de x vers y est plus grande. Cette procédure est réglée de telle manière qu'elle converge. La confiance qu'un utilisateur peut placer en une page x est définie comme la limite de la quantité de crédit que x possède quand le nombre d'étapes tend vers l'infini. Une étude complète des questions liées à Pagerank, ainsi que plusieurs versions alternatives de ce système, peuvent être trouvés dans e.g. (Langville and Meyer, 2005). Une version importante de Pagerank adaptée aux systèmes pairs-à-pairs a été développée dans (Kamvar et al., 2003).

Un même format de représentation peut avoir différentes significations. Par exemple, dans certaines approches, un agent est soit fiable (c.-à-d. que c'est quelqu'un de confiance) soit non-fiable, et les auteurs manipulent un nombre indiquant la probabilité que l'agent soit fiable. Dans d'autres approches, un agent est fiable à un certain degré, et les auteurs travaillent avec un nombre indiquant ce degré. Or, la probabilité d'être (totalement) fiable n'est pas la même chose qu'un degré de fiabilité. Par exemple, supposons que le nombre associé à un agent et représentant sa fiabilité soit $0,5$. Dans le premier cas, cela signifie que l'agent réussi parfaitement un but sur deux, alors que dans le second cas cela signifie que tous les buts effectuées par l'agent soit à moitié réussis.

Enfin, une autre facette du problème de l'évaluation de la confiance est qu'il peut faire référence à différents types de données. Typiquement, ces données prennent la forme d'opinions que les agents expriment à propos de leurs pairs. C'est par exemple le cas dans Pagerank. Mais, les informations dont un agent i dispose à propos d'un agent j peuvent aussi venir d'interactions directes entre i et j. En fait, interactions directes et opinions d'un tiers sont les principales données d'entrée des systèmes de confiance développés à ce jour.

1.3.3 Applications liées au concept de confiance

Il existe plusieurs systèmes multi-agents dans lesquels il est utile d'incorporer un système de confiance. Voici quelques exemples :

E-commerce (Ebay, Amazon, *etc.*). Les agents sont les vendeurs et les acheteurs. Les utilisateurs sont les agents eux-mêmes. Il existe des mauvais acheteurs ainsi que des mauvais vendeurs, donc les utilisateurs ont besoin d'une évaluation des agents avant de procéder à des transactions. Après chaque transaction, l'acheteur peut rendre publique son avis sur le vendeur et vice versa. Pour aider un utilisateur u à prendre des décisions, un système de confiance peut donc exploiter ces opinions, ainsi que les résultats des transactions antérieures de u afin de lui fournir une évaluation des agents.

Grands wikis (Wikipedia, Planetmath, *etc.*). Les agents sont les contributeurs du wikis, c'est à dire ceux qui créent, effacent, ou modifient des articles. Un utilisateur est un lecteur ou un patrouilleur. Le rôle d'un patrouilleur est d'empêcher ou de réparer les actes de vandalisme. Il serait facile de modifier un wiki pour que les contributeurs puissent fournir des opinions sur leurs pairs, en particulier à la suite des longues discussions qu'ils ont sur les sujets à controverse. Un système de confiance pourrait exploiter ces opinions pour aider les utilisateurs à choisir les articles à lire ou à valider.

Réseaux sociaux (Facebook, Myspace, *etc.*). Les agents sont des humains, des applications, des sociétés, des produits, etc. Les utilisateurs sont les agents. Certains agents sont mal intentionnés, donc une évaluation des agents serait utile aux utilisateurs pour décider avec qui partager des informations personnelles.

Pages web et liens hypertext. Les agents sont les pages web. Les utilisateurs sont les entités (humains, ordinateurs, etc.) qui recherchent des informations ou des services sur le web. Un lien d'une page x vers une page y peut être vu comme un avis favorable de x envers y. Ces opinions peuvent être exploitées pour aider les utilisateurs à déterminer la confiance qu'ils peuvent avoir dans le fait qu'une page web contienne des informations ou des services importants.

Articles et citations. Les agents sont les articles. Les utilisateurs sont des lecteurs. Une citation d'un article x vers un article y peut (comme pour les pages web) être vue comme un avis favorable de x envers y. Ces avis peuvent être la base d'une évaluation aidant les utilisateurs à faire des choix de lecture.

1.4 Formalisation des émotions

Il y a une littérature très abondante sur les émotions, tant en philosophie[10] (Gordon, 1987) qu'en psychologie (Lazarus, 1991; Ortony et al., 1988), en économie (Loewenstein, 2000) ou en sciences cognitives (Lane and Nadel, 2000). En informatique, les émotions sont désormais une thématique importante des systèmes d'agent dans lesquels elles interviennent à différents niveaux. De nombreux travaux se concentrent sur le rendu visuel des émotions au travers de la modélisation faciale et gestuelle d'agents conversationnels animés (ACA) (voir par exemple (Gratch and Marsella, 2005; Pelachaud, 2009)). Les ACA utilisent des modèles de l'émotion pour représenter également les émotions des utilisateurs ou pour montrer leur état affectif ou une personnalité particulière.

Le but est que ces agents paraissent le plus réaliste possible au sens où l'utilisateur de tels systèmes aurait l'impression d'interagir avec un autre être humain. Cela suppose d'une part un réalisme des aspects expressifs de l'agent (mouvements faciaux et corporels, intonation, expression verbale, *etc.*), et d'autre part la capacité à reconnaître et à prendre en compte dans leur système de raisonnement les émotions des utilisateurs (aussi bien que les leurs) afin de parler ou d'agir de la manière la plus adéquate possible.

10. Platon établit très clairement une distinction entre raison, passion et désir.

En d'autre termes, ce dernier aspect signifie que cela doit conduire à une interaction naturelle et optimale car nous savons aujourd'hui que nous communiquons sans arrêt des informations sur notre état émotionnel (réel ou non) sans avoir du coup à verbaliser ce comportement. Par exemple, un « salut ! » accompagné d'un sourire est un moyen commun et verbalement économique d'exprimer ses salutations à quelqu'un ainsi que le fait qu'on est heureux de le voir (ce qui pourrait être verbalisé par « Salut, je suis content/heureux de te voir »).

1.4.1 Formalisation logique des émotions

Au niveau des modèles formels de l'émotion, on cherche à développer des cadres logiques pour la formalisation de certaines émotions spécifiques, leurs propriétés, les liens qu'elles entretiennent les unes par rapport aux autres, *etc.* (voir par exemple les travaux de (Adam et al., 2009; Turrini et al., 2010)). Le but principal est d'exploiter des méthodes logiques pour spécifier de manière rigoureuse comment les émotions pourraient être implémentées au sein d'un agent artificiel. La conception de tels systèmes d'agent, capables de raisonner et d'exprimer certaines émotions, peut de plus bénéficier du fait que la logique est un outil particulièrement adapté au raisonnement tout en obligeant le concepteur du système à désambiguïser les différentes dimensions des émotions identifiées au sein de différents modèles psychologiques de l'émotion.

Les définitions logiques des émotions caractérisent le plus souvent ce qu'il est plutôt convenu d'appeler des structures cognitives d'émotion plutôt que les émotions elles-mêmes. Suivant les théories de l'évaluation cognitive (Lazarus, 1991), la structure cognitive d'une émotion est la configuration de l'état mental qu'un agent a à l'esprit quand il ressent cette émotion, et qui est responsable de son sentiment. La structure cognitive est juste une partie du phénomène affectif dans son entier. Par la suite, nous utilisons le terme *emotion* pour référer à *la structure cognitive d'une émotion.*

Parmi les émotions, nous distinguons les émotions simples de ce que nous appelons ici les émotions complexes (Adam et al., 2011; Lorini and Schwarzentruber, 2011). Les premières sont celles pouvant être décrites uniquement avec des attitudes mentales telles que la croyance ainsi que les buts ou les idéaux. Les secondes sont celles nécessitant des raisonnements plus complexes du type contrefactuel : « j'aurais pu faire en sorte que φ soit vrai (resp. faux) alors qu'actuellement φ est faux (resp. vrai). » En ce sens, les émotions complexes sont associées à des raisonnements contre-factuels portant sur des normes, des responsabilités.

Par exemple, le fait que l'agent i ressente de la joie à propos du fait φ (ou se réjouisse que φ soit vrai) peut être exprimé de la façon suivante :

$$Joy_i\,\varphi \stackrel{def}{=} Bel_i\,\varphi \wedge Choice_i\,\varphi$$

En accord avec cette définition, l'agent i ressent de la joie à propos de φ si et seulement si i croit que φ est vrai et qu'il souhaite que φ soit vrai. Par exemple, Tom ressent de la joie d'avoir réussi son examen parce qu'il pense avoir réussi son examen et que c'est ce qu'il souhaitait. Ainsi, Tom est content dans le sens où il croit que l'état de chose présent correspond à ce qu'il souhaitait qu'il fut. La joie a une valence positive, c'est-à-dire que lorsqu'elle est éprouvée, elle est associée à un état de fait congruent

avec les désirs. Ce n'est pas le cas de la tristesse par exemple, dont l'état de fait associé est incongruent avec les désirs.

En ce qui concerne les émotions complexes, nous nous limitons à celles en relation avec la notion de responsabilité. Celle-ci peut incomber à l'agent ressentant l'émotion en question, ou à un autre agent différent. La responsabilité de l'agent i pour le fait que φ soit vrai peut être définie par le fait que φ soit vrai et qu'il aurait pu faire en sorte qu'il ne le soit pas. Soit :

$$\mathbf{Resp}_i\varphi \stackrel{def}{=} \varphi \wedge \mathbf{Cd}_i\neg\varphi$$

Ici $\mathbf{Cd}_i$ (l'agent fait en sorte que) est un opérateur de base du langage formel mais peut-être défini à partir de l'opérateur d'action implicite STIT vu précédemment. (Pour plus de détails, voir (Lorini and Schwarzentruber, 2011).)

Ainsi (voir par exemple (Zeelenberg et al., 1998)) quand l'agent i croit qu'il est responsable du fait que φ alors qu'il avait $\neg\varphi$ comme but, il ressent du regret (et réciproquement). Soit, formellement :

$$\mathit{Regret}_i\varphi \stackrel{def}{=} \mathit{Goal}_i\,\neg\varphi \wedge \mathit{Bel}_i\,\mathbf{Resp}_i\varphi.$$

D'autres émotions peuvent ainsi être définies de manière similaire. Les émotions sont un domaine en pleine expansion du fait que l'informatique n'exploite encore que peu les possibilités offertes, et avec difficulté. Dans les systèmes implémentés, les modèles se réduisent souvent à de simples étiquettes qui sont activées ou désactivées selon les besoins. Les modèles formels basés sur la logique obligent à expliciter la nature des émotions et, de ce fait, nous aide à les comprendre.

1.4.2 Les modèles numériques des émotions

Il existe des modèles numériques des émotions qui étudient les aspects quantitatifs des phénomènes affectifs. Par exemple, El-Nasr et al. (El-Nasr et al., 2000) ont proposé le modèle numérique des émotions FLAME (Fuzzy Logic Adaptive Model of Emotions) basé sur la logique flue. La contribution principale de ce travail est une quantification de l'intensité des émotions, à partir des variables d'évaluation (appraisal variables) comme la désidérabilité ou la probabilité d'un événement. Par exemple, en se basant sur le modèle psychologique des émotions d'Ortony, Clore et Collins (Ortony et al., 1988), dans le modèle FLAME l'intensité de l'espoir par rapport à un certain événement dépend du degré de désidérabilité de cet événement et de la probabilité subjective que l'événement aura lieu. Plus récemment, plusieurs chercheurs en IA ont étendu des modèles formels des émotions avec des aspects quantitatifs. Par exemple, Meyer et al. (Steunebrink et al., 2008) ont proposé un modèle décrivant comment l'intensité des émotions décline dans le temps. Lorini (Lorini, 2011) a proposé un étude systématique de l'intensité des émotions basées sur les attentes (espoir, crainte, déception, soulagement) et de la relation entre ces émotions et le mécanisme de révision des croyance d'un agent cognitif.

Il existe également des modèles numériques de l'émotion qui représentent celle-ci comme un vecteur de nombres où chaque nombre correspond à une composante

de l'émotion. Par exemple, Mehrabian capture l'humeur au travers de la mesure du plaisir, de l'excitation, et de la dominance (la capacité de l'individu à dominer le stimuli auquel il est soumis). Selon la valeur de chacune de ces composantes, l'humeur associée est différente. On peut citer également des travaux sur le robot à tête humaine WE-4R développé à l'université de Waseda (Japon) par Hiroyasu Miwa et son équipe. Le modèle des émotions est un vecteur orienté dans l'espace et calculé à partir de trois composantes : le plaisir, l'activation et la détermination.

1.4.3 Applications liées au concept d'émotion

Au niveau applicatif, il a été développé des systèmes d'enseignement gérant les émotions afin d'augmenter la persévérance et l'engagement des étudiants, mais également des simulateurs ou des jeux vidéos, et des systèmes d'intelligence ambiante (voir par exemple (Adam et al., 2011) pour un survol de la littérature et des applications sur les émotions dans ce domaine). Parmi la très large variété d'ACA existants, EM [11] est un système assez typique qui simule le déclin des émotions à travers le temps pour un ensemble spécifique d'émotions en accord avec les buts qui les ont générés. Un autre exemple est le système Affective Reasoner de Gratch & Marsella où les agents utilisent des représentations d'eux-mêmes et des autres. Enfin, GRETA (de Rosis et al., 2003) est un ACA 3D pouvant être animé temps-réel et capable d'exprimer des états émotionnels.

1.5 Conclusion

Dans ce chapitre, nous avons tout d'abord abordé la formalisation des systèmes d'agent cognitif. Un tel agent est capable d'agir de manière autonome selon les buts qu'il s'est fixé, et est généralement caractérisé *a minima* par des attitudes mentales (croyances, désirs, normes, *etc.*), le temps et l'action. Après un bref historique des grands courants qui animent ce domaine, nous avons présenté les concepts fondamentaux des systèmes BDI ainsi que les outils pour gérer un problème bien connu en IA qui est l'évolution des connaissances. Enfin, nous avons utilisé les concepts précédemment définis dans le but de formaliser deux concepts utilisés dans ces systèmes : la confiance, et l'émotions. Nous avons également montré que ces dernières sont également capturées au sein de formalismes numériques, bien moins fins au niveau des définitions des concepts manipulés, mais plus aisés à mettre en œuvre dans le cadres d'applications concrètes.

Bien entendu, il existe également de très nombreuses branches en IA concernant la formalisation de systèmes d'agent cognitifs. Mais certains ne se basent pas sur les états mentaux, d'autres se limitent à un langage de représentation. La particularité des systèmes présentés ici est qu'ils correspondent à des logiques (avec une axiomatique et une sémantique) dont les propriétés (en termes de complexité, de décidabilité, de complétude) sont également un objet d'étude. Plus particulièrement, il s'agit de logiques modales qui sont particulièrement adaptées pour représenter les états mentaux ainsi que

11. Il s'agit d'un système basé sur l'architecture Tok du projet Oz. Voir http://www.cs.cmu.edu/afs/cs.cmu.edu/project/oz/web/.

les relations entre ces différents états (croyances sur les croyances des autres agents, ou sur leur buts, *etc.*). Il s'agit pour ces systèmes de représenter finement des concepts qui seront utilisés par les agents tout en trouvant un formalisme logique ayant de « bonnes » propriétés logiques. Les enjeux sont donc à la fois informatiques et mathématiques tout en se nourrissant abondamment des SHS *via* la philosophie et la psychologie notamment.

En dehors des deux concepts présentés, il existe non seulement des études sur l'influence de l'un sur l'autre (voir par exemple (Bonnefon et al., 2009)), mais également sur un grand nombre d'autres concepts. Nous n'avons notamment pas abordés la formalisation de concepts sociaux non réductionnistes. Par exemple, il existe des notions de croyances ou d'acceptance de groupe non réductibles à la somme des croyances ou des acceptances de chacun des agents constituant ce groupe. Il faut alors être capable de capturer le groupe comme une entité unique constituant une institution particulière régie par des règles sociales particulières.

Bibliographie

Adam, C., Gaudou, B., Longin, D., and Lorini, E. (2011). Logical modeling of emotions for Ambient Intelligence. In Mastrogiacomo, F. and Chong, N.-Y., editors, *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives*. IGI Global.

Adam, C., Herzig, A., and Longin, D. (2009). A logical formalization of the OCC theory of emotions. *Synthese*, 168(2):201–248.

Amgoud, L. and Rahwan, I. (2006). An Argumentation-based Approach for Practical Reasoning. In Weiss, G. and Stone, P., editors, *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006)*, pages 347–354. ACM Press.

Austin, J. L. (1962). *How To Do Things With Words*. Oxford University Press.

Baltag, A. and Moss, L. S. (2004). Logics for epistemic programs. *Synthese*, 139(2):165–224.

Belnap, N., Perloff, M., and Xu, M. (2001). *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York.

Bonnefon, J.-F., Longin, D., and Nguyen, M. H. (2009). A Logical Framework for Trust-Related Emotions. *Electronic Communications of the EASST, Formal Methods for Interactive Systems 2009*, 22:1–16.

Bratman, M. (1987). *Intentions, plans, and practical reason*. Harvard University Press, Cambridge.

Burgess, J. P. (2002). Basic tense logic. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 7, pages 1–42. Kluwer, 2nd edition.

Castaneda, H. N. (1975). *Thinking and Doing*. D. Reidel, Dordrecht.

Castelfranchi, C. and Paglieri, F. (2007). The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese*, 155:237–263.

Castelfranchi, C. and Tan, Y. H., editors (2001). *Trust and Deception in Virtual Societies*. Kluwer Academic Publishers, Dordrecht.

Chellas, B. F. (1980). *Modal Logic: an Introduction*. Cambridge.

Cohen, P. R. and Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence Journal*, 42(2–3):213–261.

Cohen, P. R., Morgan, J., and Pollack, M. E., editors (1990). *Intentions in Communication*. MIT Press, Cambridge, MA.

Conte, R. and Castelfranchi, C. (1995). *Cognitive and social action*. London University College of London Press, London.

de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., and De Carolis, B. (2003). From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59:81–118.

Ditmarsch, H. v., der Hoek, W. v., and Kooi, B. (2007). *Dynamic Epistemic Logic*. Kluwer Academic Publishers.

El-Nasr, M. S., Yen, J., and Ioerger, T. R. (2000). FLAME: Fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems*, 3(3):219–257.

Gilbert, M. (1989). *On Social Facts*. Routledge, London and New York.

Gochet, P. and Gribomont, P. (2006). Epistemic Logic. In Gabbay, D. and Woods, J., editors, *Handbook of the History of Logic*, volume 7, pages 99–195. Elsevier.

Gordon, R. (1987). *The structure of emotions*. Cambridge University Press, New York.

Gratch, J. and Marsella, S. (2005). Lessons from emotion psychology for the design of lifelike characters. *Journal of Applied Artificial Intelligence (special issue on Educational Agents - Beyond Virtual Tutors)*, 19(3-4):215–233.

Harel, D., Kozen, D., and Tiuryn, J. (2000). *Dynamic Logic*. MIT Press, Cambridge.

Herzig, A., Lorini, E., Hübner, J. F., and Vercouter, L. (2010). A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1):214–244.

Kamvar, S. D., Schlosser, M. T., and Garcia-Molina, H. (2003). The Eigentrust Algorithm for Reputation Management in P2P Networks. In *12th International Conference on World Wide Web (WWW)*, pages 640–651. ACM.

Kooi, B. (2007). Expressivity and completeness for public update logic via reduction axioms. *Journal of Applied Non-Classical Logics*, 17(2):231–253.

Lane, R. and Nadel, L., editors (2000). *The cognitive neuroscience of emotions*. Oxford.

Langville, A. N. and Meyer, C. D. (2005). Deeper Inside PageRank. *Internet Mathematics*, 1(3):335–400.

Laverny, N. and Lang, J. (2005). From knowledge-based programs to graded belief-based programs, part ii: off-line reasoning. In *Proceedings of IJCAI'05*, pages 497–502. Professional Book Center.

Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford University Press.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *American Economic Review*, 90(2):426–432.

Lorini, E. (2011). The cognitive anatomy and functions of expectations revisited. In Paglieri, F., Tummolini, L., Falcone, R., and Miceli, M., editors, *The Goals of Cognition: Festschfit for Cristiano Castelfranchi*. College Publications, London. to appear.

Lorini, E., Longin, D., Gaudou, B., and Herzig, A. (2009). The logic of acceptance: grounding institutions on agents' attitudes. *Journal of Logic and Computation*, 19(6):901–940.

Lorini, E. and Schwarzentruber, F. (2011). A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175:814–847.

Marsh, S. (1994). *Formalising Trust as a Computational Concept*. PhD thesis, Department of Computing Sciece and Mathematics, University of Sterling.

Ortony, A., Clore, G., and Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

Pelachaud, C. (2009). Modelling multimodal expression of emotion in a virtual agent. *Philosophical transactions of the Royal society B*, 364:3539–3548.

Rao, A. S. and Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. In *Proceedings of KR'91*, pages 473–484. Morgan Kaufmann Publishers.

Sabater, J. and Sierra, C. (2005). Review on Computational Trust and Reputation Models. *Artificial Intelligence*, 24:33–60.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge.

Segerberg, K. (1992). Getting started: Beginnings in the logic of action. *Studia Logica*, 51(3-4):347–378.

Segerberg, K. (1995). Belief revision from the point of view of doxastic logic. *Logic Journal of IGPL*, 3(4):535–553.

Singh, M. P. (1999). An ontology for commitments in multiagent systems. *Artificial Intelligence and Law*, 7:97–113.

Steunebrink, B. R., Dastani, M., and Meyer, J.-J. C. (2008). A formal model of emotions: integrating qualitative and quantitative aspects. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008)*, pages 256–260. IOS Press.

Turrini, P., Meyer, J.-J. C., and Castelfranchi, C. (2010). Coping with shame and sense of guilt: a dynamic logic account. *Journal of AAMAS*, 20(3).

van Benthem, J. (1991). *The Logic of Time*. D. Reidel Publishing Company.

van Benthem, J. and Liu, F. (2007). Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 17(2):157–182.

van der Hoek, W., Jamroga, W., and Wooldridge, M. (2007). Towards a theory of intention revision. *Synthese*, 155(2):265–290.

van Ditmarsch, H., van der Hoek, W., and Kooi, B. (2007). *Dynamic Epistemic Logic*. Kluwer Academic Publishers.

van Linder, B., van der Hoek, and W., Meyer, J.-J. C. (1998). Formalising abilities and opportunities. *Fundamenta Informaticae*, 34:53–101.

Wooldridge, M. (2000). *Reasoning about Rational Agents*. MIT Press.

Zeelenberg, M., van Dijk, W. W., and Manstead, A. S. R. (1998). Reconsidering the relation between regret and responsibility. *Organizational Behavior and Human Decision Processes*, 74:254–272.