CNRS - INPT - UPS - UT1 - UTM

RAPPORT DE RECHERCHE

IRIT/RR–2012-12–FR

La honte *versus* la culpabilité : analyse conceptuelle et formelle en logique modale

Dominique Longin

Equipe LILaC
Dominique.Longin@irit.fr

Mai 2012

UPS-IRIT, 118 route de Narbonne, 31062 Toulouse CEDEX 9
+33 (0) 561 55 67 65 info@irit.fr www.irit.fr

Résumé

Une étude de la honte présente un intérêt tant applicatif que théorique. Dans le premier cas, il s'agit de faire en sorte qu'une machine soit capable de détecter de tels sentiments chez l'usager afin de développer des stratégies palliatives et, par voie de conséquence, d'améliorer son efficacité (tutoring intelligent pas exemple). Dans le second cas, il s'agit de replacer la honte parmi les autres émotions, en étudiant non seulement ce qui fait le propre de la honte, mais également ce qui différencie cette dernière des autres émotions, en particulier la culpabilité. Après une brève présentation de ce qu'est l'émotion, le présent article présente dans un premier temps une analyse approfondie de la honte en philosophie et psychologie. Dans un second temps, un langage formel de type logique modale est présenté afin d'offrir un cadre de formalisation d'un ensemble d'émotions, dont la honte, celle-ci (et d'autres) étant ensuite formalisée au sein de ce cadre. Il en découle un cadre unifié propre à représenter des émotions simples telles que la joie ou la tristesse, ou des émotions plus complexes telles que la honte ou la culpabilité.

Mots-clés

Émotions, modal logic, shame, guilt, regret.

Table des matières

1 Introduction

La honte est une émotion très particulière et son étude se justifie à plusieurs niveaux. En premier lieu, c'est une émotion morale (elle est en relation avec les normes qu'un individu se fixe). Elster [15, p. 145] souligne combien les normes sociales ont une influence immensément puissante sur le comportement (« *an immensely powerful influence on behavior* »). De ce fait, la honte nous touche dans ce que nous avons de plus intime, de plus personnel. Comme le notent Tangney et Ronda [36], c'est une émotion qui a une influence certaine sur l'image que nous avons de nous-mêmes et sur la manière dont on pense être socialement perçu. C'est donc une émotion clé de notre comportement, notamment en situation de prise de décision, qui constitue pour Elster *le* support des normes sociales. [1]

Deuxièmement, il est fréquent de constater dans la littérature que la définition d'une émotion peut varier d'un auteur à l'autre [2] ou que certaines émotions sont amalgamées, rendant très difficile la mise en place d'une typologie cohérente des émotions. Par exemple, la honte est très souvent confondue avec la culpabilité (voir section suivante pour plus de détails). Il y a donc une justification fondamentale à étudier *une* émotion en relation avec ou par opposition à d'autres émotions. De ce point de vue, étudier la honte permet d'appréhender plus précisément la culpabilité (et vice-versa).

Enfin, une étude de l'émotion se justifie dans un cadre applicatif, en permettant à une machine de comprendre certaines inhibitions de ses utilisateurs, ou au contraire en lui fournissant un tel mécanisme afin de s'adapter à son environnement social. Lors de l'enseignement d'une langue par exemple, le système détectera certaines inhibitions liées à la honte de parler une certaine langue étrangère par exemple, et pourra ainsi utiliser des stratégies palliatives (discours rassurant, encourageant, *etc.*).

Dans ce qui suit, nous présentons une vision assez bien acceptée de l'émotion afin de clairement délimiter le présent travail de formalisation (Section 2). En particulier, il ne s'agit pas de donner une description parfaitement fidèle et psychologiquement plausible de la honte mais plutôt d'essayer d'en capturer la structure cognitive au sein de laquelle elle peut naître. Après une analyse des émotions de honte et de culpabilité (Section 3) nous présentons ensuite le cadre formel (Section 4) sur lequel nous nous appuierons pour caractériser (en particulier) ces deux émotions (Section 5).

2 Qu'est-ce qu'une émotion ?

« *What is an emotion?* » est le titre d'un article de William James, l'un des pères fondateurs de la psychologie expérimentale, dans lequel l'auteur tente de présenter pour la première fois de manière moderne et rigoureuse ce qu'est une émotion [23]. [3] Sa vision était que l'émotion est la *conséquence directe* d'un changement corporel résultant lui-même

1. Elster illustre son point de vue par l'exemple suivant : si l'agent i viole une norme sociale, je vais refuser de traiter avec lui, ce qui va peut-être engendrer chez lui une perte matérielle quelconque, mais qui va surtout marquer mon mépris ou mon dégoût, ce qui va engendrer chez lui de la honte. Et plus il me coûtera de refuser de traiter avec lui, plus sa honte sera importante [15, p. 146].

2. L'espoir par exemple est défini de manière duale chez Ortony, Clore et Collins [29] et Lazarus [24].

3. Cette question, des philosophes se la posaient déjà depuis l'antiquité puisque Platon [30] cherchait déjà à distinguer raison, passion et désir. De même, dans La rhétorique, Aristote analyse l'émotion au travers de son rôle dans la politique grecque. Il l'aborde de manière détaillée dans l'Ethique à Nicomaque.

d'un certain stimuli. Dans la lignée de ses travaux on peut citer : Tomkins [39] où certaines actions faciales, comme le froncement de sourcil par exemple, jouent un rôle central dans la régulation de nos émotions ; Ekman [14] qui a élaboré un grand nombre d'expérience tendant à prouver que certaines expressions motrices peuvent amplifier, voire déclencher, certaines émotions ; ou Damasio et ses marqueurs somatiques biaisant la prise de décision. Mais à la lumière d'autres expériences, et comme le soulignent Sander et Scherer dans [32, p. 8] : « Une question importante apparaît alors : la réaction corporelle est-elle une cause, une composante, ou une conséquence de l'émotion ? ».

Face à des expériences contradictoires étayant tantôt une vision tantôt une autre, ces auteurs (et d'autres) adoptent une vision multi-componentielle de l'émotion : *le sentiment* (le ressenti de l'émotion) ; *la réponse psychophysiologique* (accélération du rythme cardiaque, de la température corporelle, *etc*.) ; *l'expression motrice* (du visage, de la voix, des gestes) ; *la tendance à l'action* (c'est-à-dire les possibilités d'action mises en avant par l'émotion, sans que l'on puisse confondre ces tendances à l'action avec l'action elle-même) ; *l'évaluation cognitive* (ou *appraisal* dans la littérature anglo-saxonne). Dans les théories dites de l'évaluation cognitive, cette dernière composante est considérée comme étant celle qui détermine les quatre autres. Elle représente le processus cognitif d'évaluation d'un certain événement qui déclenche une réponse émotionnelle différenciée, c'est-à-dire qui détermine si c'est une émotion qui est déclenchée plutôt qu'une autre, les autres composantes n'étant alors que des sortes de canaux de manifestation dans notre corps et notre esprit de l'émotion déclenchée. Cette différenciation serait rendu possible par le fait que l'on évalue (consciemment ou non) un stimuli donné par rapport à notre état mental (incluant nos préférences, buts, idéaux, et connaissances acquises au cours d'expériences passées). Ainsi, une émotion correspond alors à une *variation épisodique* de certaines de ces composantes suite à l'évaluation d'un événement donné [32]. Cela permet de distinguer clairement l'*émotion* du *sentiment* (qui n'est qu'une composante de l'émotion) et de l'*humeur* (qui n'est pas un phénomène épisodique, pas nécessairement déclenchée par un événement particulier, et d'intensité moindre que l'émotion).

Dans ce suit, une émotion sera ainsi toujours *à propos de quelque chose* : on sera déçu *de voir son équipe préférée perdre* mais jamais « triste en général » (car c'est plutôt une humeur). En revanche, nous ne prendrons en compte ni la réponse psychophysiologique, ni l'expression motrice car nous nous focalisons sur l'aspect cognitif des émotions et non sur leur expression. Concernant le sentiment de l'émotion, nous le représentons par le fait que notre agent est introspectif et conscient de ses émotions. Mais cela ne constitue bien sûr qu'une partie du *sentiment de l'émotion* car celui-ci inclut également le sentiment des changements psychophysiologiques (on sent son cœur s'accélérer par exemple) ou moteur (on entend sa voix changer ou on sent ses traits se déformer sous le coup de la peur par exemple). Le sentiment est également lié à une notion d'intensité : on ressent des émotions plus ou moins fortement (elles nous « touchent » plus ou moins) que nous ne traitons pas ici pour ne pas compliquer le formalisme (bien que des solutions techniques existent, comme par exemple celle développée par Lorini [26] qui offre un cadre formel cohérent avec celui du présent article). Comme dans la littérature, nous considérons l'évaluation cognitive comme la (non) congruence entre une croyance de l'agent (conséquence d'une observation ou d'un raisonnement) et ses buts/désirs ou ses idéaux (selon l'émotion considérée), ce qui est en tout point conforme avec les théories psychologiques de l'émotion. Enfin, les tendances à l'actions pourront être capturées *via* l'ensemble des actions

rendues exécutables par une émotion donnée. Ce point ne sera que partiellement abordé car il n'est pas au centre du présent travail. En définitive, nous formalisons dans ce qui suit ce qui correspond davantage à des structures cognitives d'émotion que les émotions elles-mêmes, en tant qu'entités multi-componentielle.

3 La honte *versus* la culpabilité

Les sociétés orientales comme le Japon ou la Chine ont une approche très spécifique vis-à-vis de la honte et on parle de « cultures de la honte », par opposition à nos sociétés occidentales, que l'on qualifie de « cultures de la culpabilité » [4]. On peut également décrire la Grèce antique comme une « culture de la honte ». À ce titre les études de Ruth Benedict en 1946 sont particulièrement révélatrices [6]. Dans les « cultures de la culpabilité », on restreint le comportement des individus en les rendant coupables. Dans les « culture de la honte » les conséquences sociales d'un acte rendu public et considéré comme honteux sont bien plus importantes et déterminantes que les sentiments individuels. Ce sont des cultures où les rangs sociaux ont une importance capitale dans l'organisation et la vie de tous les jours. L'image que dégage une personne la définit, c'est pour cela que les individus y sont particulièrement sensibles, et qu'un acte rendu publique qui ternit leur image est si terrible pour eux. Un parfait exemple est le *seppuku*, le suicide rituel au Japon. Une des raisons qui entraînait cet acte était la volonté de laver son image d'un échec personnel, ou d'un mauvais comportement.

La honte et la culpabilité ont bien souvent été assimilées ou peu différenciées l'une de l'autre. La principale raison est que l'évaluation de ces deux émotions est basée sur la violation d'une norme sociale par un comportement inapproprié par rapport à une société donnée (voir par exemple [38, 24, 15, p. 145]) et qu'elles sont à ce titre toutes les deux des émotions morales (*moral emotions*). [5] Parmi ceux qui assimilent honte et culpabilité, Ortony *et al.* [29, p. 142–143] par exemple voient dans honte la violation d'un standard considéré comme important (comme une norme morale, par exemple) et dont la violation est inexcusable, ce qui n'est pas une condition nécessaire de la culpabilité. Pour eux, la culpabilité serait principalement une émotion composée à partir de la honte et du regret. On trouve dans la littérature en psychologie une très forte proportion de travaux assimilant honte et culpabilité. (Voir par exemple [36, p. 11–12] pour plus de détails à ce sujet.)

C'est très certainement à Lewis [25] que l'on doit d'avoir trouvé un critère discriminant la honte et la culpabilité, critère par la suite vérifié expérimentalement dans un nombre très important de travaux en psychologie. Lorsqu'un individu éprouve de la honte, c'est lui-même qu'il juge, sa propre personne dans son ensemble. Dans le cas de la culpabilité, ce sont ses actions, son comportement. Ainsi Elster [15, p. 143–144] définit la honte comme une émotion négative déclenchée par une croyance à propos de sa propre personne (« *a negative emotion triggered by a belief about one's own character* ») et la culpabilité comme une émotion négative déclenchée par un croyance à propos de ses propres actions (« *a negative emotion triggered by a belief about one's own action* »). (Voir aussi [13, 24, 36] par exemple.)

4. Par *culpabilité*, nous et non à la notion légale prononcée par la justice.

5. Elster [15, p. 149], citant en cela K. Dover (Greek Popular Morality, 1994), les appelle également des émotions de l'embarras (« *self-conscious emotions* »).

Cette distinction explique en particulier pourquoi la honte se ressent bien plus profondément que la culpabilité, pourquoi elle est bien plus douloureuse, et pourquoi il est beaucoup plus difficile de lutter contre elle. Elle explique aussi par conséquent pourquoi la honte conduit à vouloir systématiquement chercher à ce que l'objet de notre honte ne s'ébruite pas [29], à tenter de minimiser son exposition aux autres agents. Lazarus note que dans les cas extrêmes, on se sent incapable de vivre en société selon les normes établies, d'atteindre « l'ego idéal » [24, 28] ce qui peut conduire au suicide [16, p. 274]. Plusieurs méthodes sont possibles comme nier tout lien avec la transgression ou insister sur la nature privée des évènements [28]. Dans la cas de la culpabilité, on a plutôt tendance à adopter un comportement actif et réparateur [15, 28] dans le but de minimiser ou effacer les conséquences de notre action. Un corollaire à cela est que dans le cas de la culpabilité on se sent nécessairement responsable de la situation présente (sinon on ne pourrait pas se sentir coupable) alors que dans le cas de la honte toute responsabilité, quand elle est réelle, est non assumée [28]. Elster [15, p. 150], citant en cela [38], indique que la honte peut avoir une cause indépendante de notre bonne volonté, comme avoir des parents pauvres ou devenir vieux.

Cette distinction explique également que les idéaux mis en jeu soient un peu différents. Dans le cas de la culpabilité, il s'agit d'idéaux internalisés, que l'agent a fait siens. Si je me sens coupable de m'être garé sur une place pour personnes handicapées, c'est parce que je me reconnais dans le fait qu'il est mal de se garer sur de telles places si on n'est pas handicapé. Je considère ce principe comme devant être respecté. Si au contraire j'ai connaissance de ce principe mais que pour moi ce n'est pas important, alors je pourrai me garer sur une telle place sans me sentir coupable. Dans le cas de la honte, nous n'imposons pas que la norme violée soit une norme internalisée. Par exemple, supposons qu'un individu rentre pour la première fois dans un restaurant très chic et qu'il attache sa serviette autour du cou. Dès lors qu'il s'aperçoit (au travers du regard du serveur ou des autres invités par exemple) de l'inadéquation de son comportement par rapport à son environnement, il pourra éprouver de la honte d'avoir noué sa serviette autour du cou même s'il continue de penser que c'est mieux de le faire (parce que ça évite de salir sa chemise par exemple). Si toutefois ce n'est pas un idéal internalisé et qu'il ne reconnaît pas non plus que dans ce contexte son comportement peut être perçu comme inadapté ou décalé, il n'éprouvera pas de honte. Pour Lazarus [24, p. 240] la culpabilité comme la honte requièrent des normes internalisées. Nous pensons que dans la honte, des normes (internalisées) plus générales que celles violées peuvent suffire à déclencher de la honte et nous n'imposons pas que ces dernières soient nécessairement internalisées (tout en ne l'interdisant pas non plus).

Certains ont argué que la honte inclut nécessairement une dimension sociale, publique [15, 38, 28, 29, 13], ce qui ne serait pas le cas pour la culpabilité. Elster [15, p. 149] par exemple dit que « je ressens de la honte en votre présence parce que je sais que vous me désapprouvez » (« *I feel shame in your presence because I know you disapprove of me.* »). Encore faut-il préciser ce qu'on entend pas dimension sociale. Intuitivement, une personne ressentant de la honte mêle étroitement sa personne et un groupe (ou une institution) vis-à-vis de duquel (ou de laquelle) elle éprouve de la honte. Dans [37] les auteurs ont mené des expérimentations dont les résultats montrent que la honte ressentie en dehors de tout groupe témoin est au contraire légèrement plus fréquente que pour la culpabilité. Ce n'est donc pas un critère discriminant. Les auteurs citent l'exemple d'un

adulte racontant que lorsqu'il était enfant, il a vu son frère se faire réprimander par leur mère pour avoir fait quelque chose d'immoral. Lui-même avait fait la même chose mais sa mère l'ignorait. Pourtant il a ressenti de la honte. La dimension sociale ne se situe donc pas nécessairement au niveau du fait que l'objet de notre honte soit connu d'un certain groupe, mais plutôt au niveau du fait qu'on *croit* que cela constitue une violation d'ordre morale vis-à-vis de ce groupe. Darwin [12, p. 352] dit qu'un individu peut éprouver de la honte mais ne pas rougir pour autant ; que pour rougir, il faut que l'objet de sa honte ait été découvert. Cela signifie qu'on peut éprouver de la honte sans que l'objet de notre honte soit exposé publiquement. Autrement dit, le groupe face auquel on éprouve de la honte n'a pas besoin d'être participatif ou physiquement présent [38], il n'a pas besoin d'être même au courant de la violation de la norme en question : il suffit de penser que ce groupe a un certain idéal que je viole. Lazarus [24, p. 241] souligne même qu'on peut éprouver de la honte vis-à-vis d'une personne décédée. En revanche, il y a indéniablement une dimension sociale dans la honte (et la culpabilité) au sens elle met en jeu un (groupe d')individu(s) : celui à qui on a causé du tord (culpabilité) ou celui face auquel notre image a été perçue négativement (honte).

La dimension sociale touche également les réactions potentielles suite au ressenti de la honte (les tendances à l'action, ou *action tendencies*). Toute la littérature s'accorde à dire que quelqu'un qui éprouve de la honte a tendance à vouloir se faire tout petit, à se cacher, à minimiser les faits (voir [36] ou [24, p. 244] par exemple). Dans le cas de la culpabilité, l'individu a plutôt tendance à vouloir réparer les effets négatifs de son action. En particulier, dans le cas où la transgression n'est pas encore connue, l'individu ressentant de la honte ne souhaite pas qu'elle le soit et peut même essayer de la cacher.

3.1 En résumé

Objet. Selon notre définition des émotions, honte et culpabilité sont toutes les deux toujours à propos de quelque chose : on a honte d'avoir fait d'avoir cassé quelque chose dans un magasin ou on se sent coupable d'avoir fait une mauvaise action, mais on n'a pas honte *en général* ni un sentiment de regret *en général* (car il s'agit plutôt d'humeurs).

Idéal. Que ce soit dans le cadre de la honte ou de la culpabilité il y a violation d'un idéal et celui-ci n'est pas nécessairement internalisé dans le premier cas alors qu'il l'est dans le second.

Responsabilité. Dans la culpabilité, l'agent pense être responsable de la situation constituant la violation d'un idéal, c'est là l'essence même de la culpabilité puisque le focus est mis sur ses actions. Dans la honte, l'agent ne pense pas (à tort ou a raison) avoir fait quelque chose de mal et refuse d'assumer cette situation ou cherche à la minimiser. Il ne se sent donc pas responsable à proprement parler. Ceci représente bien le fait que dans la culpabilité, on se concentre sur l'acte et ses conséquences, et que dans la honte c'est l'individu en lui-même qui se juge, le fait qu'il n'ait pas pu faire autrement.

Aspect public. Ni la honte, ni la culpabilité ne requièrent la présence d'une audience témoin de la violation de la norme à l'origine de ces émotions. En revanche, ces deux émotions ne peuvent être ressenties que vis-à-vis d'une certaine audience (un (groupe d')individus(s)) non nécessairement présente ni même nécessairement au

courant de la situation à l'origine de la honte de la culpabilité.

Tendances à l'action. La culpabilité pousse un individu à réparer le tort qu'il a causé, alors que la honte le pousse à faire en sorte que la situation honteuse ne s'ébruite pas.

3.2 Exemples

Afin d'illustrer la différence entre la honte et la culpabilité, Elster [15] cite l'exemple de la Princesse de Clèves dont l'amour pour le Duc de Nemours la fait se sentir coupable, et de Mathilde de la Mole dont l'amour pour Julien Sorel lui fait honte. En trompant son mari, la princesse de Clèves accomplit une action qui va à l'encontre de ses idéaux. Bien que Mathilde de la Mole puisse se sentir coupable pour les mêmes raisons, elle éprouve en plus de la honte du fait d'être tombée amoureuse du fils d'un charpentier. Autrement dit, elle se sent honteuse vis-à-vis des personnes de son rang d'avoir violé un idéal de ce rang. C'est son image qui est en jeu.

Autre exemple : supposons qu'une personne a faisant du shoping dans un magasin oublie involontairement un vêtement sur son sac et se dirige vers la sortie du magasin, étant ainsi sur le point de commettre un vol involontaire. Si la personne r responsable du magasin demande à a d'ouvrir son sac, a ne pourra éprouver de la culpabilité pour une action qu'il n'a pas commise volontairement. En revanche, il est probable que a éprouvera de la honte face à cette situation car sa réputation est en jeu. En revanche, en supposant au contraire que l'action ait été volontaire, et en supposant que ne pas voler est une norme internalisée de a, alors a pourrait se sentir coupable.

Considérons finalement l'exemple d'une personne perdant son pantalon dans la rue. Il s'agit là de la transgression d'une norme sociale ou culturelle (on ne se promène pas en sous-vêtements dans la rue) involontaire. L'individu éprouvera donc probablement de la honte d'avoir perdu son pantalon [6], et ne pourra là encore rien se repprocher (donc il ne pourra culpabiliser vis-à-vis de cet événement). À l'inverse, si l'on suppose qu'il l'a fait sciemment (pour provoquer, par exemple) il ne pourra éprouver de la honte, même s'il pourra par ailleurs regretter son geste.

La honte ne doit pas être confondue avec l'humiliation, cette dernière étant généralement imposée par les autres et de manière publique. Supposons par exemple qu'un enfant ait par inadvertance « fait sur lui », il éprouvera de la honte. Si la maîtresse, croyant bien faire, le laisse en sous-vêtements afin de laver le reste, il peut se sentir humilié : il subit une situation imposée par une tierce personne et portant atteinte à son image où le groupe par rapport auquel il projette son image est spectateur direct de la scène.

4 Cadre formel

Le cadre formel ci-dessous est une extension de celui développé dans [20] sur la logique $\mathcal{DLA}$ (pour *Dynamic Logic of Agency*) par des opérateurs de croyance, de but, et d'idéalité afin de pouvoir caractériser les émotions, ainsi qu'un opérateur dynamique dans le passé (opérateur *Done*) afin de pouvoir parler du passé.

6. Bien sûr, certains n'éprouveront pas de honte en pareille situation. Cela signifie simplement qu'ils n'ont pas le sentiment de violer un idéal, ou que l'importance qu'ils accordent à cet idéal est trop faible pour que l'intensité de l'émotion ressentie soit accessible à la conscience.

Cette logique a l'intérêt de pouvoir redéfinir des opérateurs d'action non explicite de type STIT (*cf.* la suite pour plus de détails) à l'aide d'opérateurs de la logique dynamique, ainsi que des opérateurs temporels permettant de parler de ce qui s'est passé l'instant juste avant ou juste après, tout en conservant un cadre décidable (alors que celui du STIT ne l'est pas). L'opérateur STIT est utilisé dans la formalisation de certaines émotions complexes telles que le regret ou la culpabilité.

Dans ce qui suit, la logique qui constitue une extension de $\mathcal{DLA}$ est appelée $\mathcal{LOCE}$ (pour *logic of complex emotions*).

4.1 Syntaxe

Soit AGT l'ensemble fini des agents, ATM l'ensemble des formules atomiques et $ACT = \{a_1, a_2, \ldots a_n\}$ l'ensemble fini non vide des actions atomiques. Le langage $\mathcal{L}_{\mathcal{LOCE}}$ est défini comme suit :

$$\varphi :: p \mid \bot \mid \neg\varphi \mid \varphi \vee \varphi \mid Bel_i\,\varphi \mid Goal_i\,\varphi \mid Ideal_i\,\varphi \mid Done_{i:a}\,\varphi \mid Happens_{i:a}\,\varphi \mid \Diamond\varphi$$

où p appartient à ATM, a à ACT et i à AGT. Les autres connecteurs classiques ($\wedge$, $\rightarrow$, $\leftrightarrow$ et $\bot$) sont définis de manière usuelle.

$Bel_i\,\varphi$ se lit : « l'agent i croit que φ est vrai ». La notion de croyance est celle d'un savoir subjectif, au sens où l'agent ne doute pas que φ soit vrai, où il pense que φ est vrai dans le monde réel (c'est un savoir subjectif).

$Goal_i\,\varphi$ se lit : « l'agent i a pour but que φ » et correspond à la notion de but de [11], qui englobe non seulement les désirs (qui sont intrinsèquement endogènes à une personne) mais également des buts pouvant provenir des normes que l'agent s'impose, ou encore des buts exogènes qui s'imposent à lui (voir [33] pour plus de détails). Une conséquence de cela est que la satisfaction d'un but ne correspond pas nécessairement à un état valencé positivement (*i.e.* bon, plaisant) mais peut simplement correspondre à un état « moins mauvais » que celui atteint en ne satisfaisant pas ce but. Une seconde conséquence est que les buts ne sont pas nécessairement réalistes (au sens où un agent peut avoir un certain but sans pour autant croire que celui-ci pourra être un jour atteint). Il est nécessaire d'adopter une définition aussi large des buts dans la mesure où des émotions peuvent être générées à partir d'une (in)congruence entre l'état réel du monde et tout type de but. Ainsi, on peut aussi bien être satisfait d'avoir gagné le cœur de quelqu'un (qui correspond plutôt à la satisfaction d'un désir) que d'avoir réussi à trouver une solution au problème d'un client (qui correspond plutôt à la satisfaction d'un but adopté).

$Ideal_i\,\varphi$ se lit : « φ est un état de chose idéal pour l'agent i ». Les opérateurs $Ideal_i$ sont utilisés pour représenter les attitudes morales de l'agent i. Plus généralement, le fait que $Ideal_i\,\varphi$ soit vrai signifie que i se commande (s'ordonne) à lui-même de faire en sorte que φ soit vrai (quand φ est faux) ou de faire en sorte qu'il continue de l'être (quand φ est déjà vrai) [7]. En ce sens, il est moralement responsable de la réalisation de φ.

$Done_{i:a}\,\varphi$ se lit : « l'agent i vient juste d'accomplir l'action a avant quoi φ était vrai ». [7]
$Done_{i:a}\,\top$ se lit : « l'agent i vient juste d'accomplir l'action a » et l'abréviation

$$Before_{i:a}\,\varphi \stackrel{déf}{=} \neg Done_{i:a}\,\neg\varphi$$

7. Il correspond à l'opérateur $\langle i{:}a^{-1}\rangle\varphi$ de la logique dynamique de Harel [19].

se lit « φ est vrai avant toute exécution de l'action a par l'agent i ». $Before_{i:a}\,\bot$ se lit : « l'agent i ne vient pas d'accomplir l'action a ».

$Happens_{i:a}\,\varphi$ se lit : « l'agent i accomplit l'action a après quoi φ sera vrai ».[8] $Happens_{i:a}\,\top$ se lit : « l'agent i accomplit l'action a » et l'abréviation

$$After_{i:a}\,\varphi \stackrel{déf}{=} \neg Happens_{i:a}\,\neg\varphi$$

se lit « φ est vrai après toute exécution de l'action a par l'agent i ». $After_{i:a}\,\bot$ se lit : « l'agent i ne va pas accomplir l'action a ».

$\Diamond\varphi$ se lit : « φ est vrai dans au moins un état alternatif », ou plus simplement « il est possible que φ soit vrai ». L'opérateur $\Diamond$ représente la possibilité historique, c'est-à-dire qu'il représente l'existence d'au moins un état alternatif à l'état présent. Autrement dit, chaque monde accessible par la relation de possibilité historique représente un présent alternatif appartenant à une histoire parallèle, c'est-à-dire un déroulement des événements différent de celui qu'on considère comme étant l'histoire réelle. Cette construction permet ainsi de représenter un futur arborescent, où différents états peuvent être atteints selon l'action accomplie car d'un point de vue sémantique, les hypothèses sur les actions font que dans chaque état (ou monde), une et une seule action est accomplie.[9] Chaque histoire peut être vue comme une séquence de mondes reliés les uns aux autres par l'exécution d'une action unique de chaque agent. L'opérateur dual de nécessité historique

$$\Box\varphi \stackrel{déf}{=} \neg\Diamond\neg\varphi$$

se lit : « φ est nécessairement vrai (quel que soit l'état alternative considéré) ». Autrement dit, φ est nécessairement vrai (quoi que les agents fassent).

4.1.1 Définition d'opérateurs temporels

Herzig et Lorini [20] définissent l'opérateur *next* tel que (pour tout agent i donné[10]) :

$$X\varphi \stackrel{déf}{=} \bigwedge_{a\in ACT} After_{i:a}\,\varphi$$

qui se lit : « φ sera vrai l'instant suivant ssi quelle que soit l'action accomplie par un agent i donné, φ est vrai après cette action ». Nous lui adjoignons l'opérateur *next moins* 1 défini

8. Cette formule correspond à $\langle i{:}a\rangle\varphi$ en logique dynamique et sa signification est traditionnellement « $i{:}a$ est exécutable après quoi φ sera vrai ». Il convient de souligner dès à présent deux hypothèses justifiant la lecture adoptée. La première porte sur le caractère linéaire du temps qui fait que si une action est exécutable, alors c'est elle qui sera nécessairement accomplie (on peut alors lire « exécutable » comme « sur le point de se produire »). La seconde est que toute action a une durée instantanée (on peut alors lire « sur le point de se produire » comme « se produit » puisqu'il n'y a d'un point de vue temporel aucune différence). Ces hypothèses sont décrites dans la sémantique des opérateurs. À titre d'exemple, si l'agent « est sur le point d'allumer la lumière », c'est comme s'il avait déjà posé le doigt sur le bouton, et commencé suffisamment à appuyer pour que, même s'il se ravisait, il appuierait quand même.

9. Le fait qu'un état soit une alternative au monde actuel ne signifie pas pour autant que l'action qui est sur le point d'être exécutée dans cet état soit différente de celle qui est sur le point d'être exécutée dans le monde présent (l'histoire « passant » par chacun de ces deux mondes diffère à un instant ou à un autre).

10. Il suffit de choisir un agent arbitraire car, comme nous le verrons dans la sémantique des opérateurs d'action, les contraintes qu'on impose au cadre font que les actions de tous les agents dans un monde w sont accomplies simultanément et mènent au même monde.

tel que :

$$X^{-1}\varphi \stackrel{déf}{=} \bigwedge_{a \in ACT} Before_{i:a}\,\varphi$$

qui se lit : « φ était vrai l'instant précédent ssi quelle que soit l'action venant d'être accomplie par un agent i donné, φ était vrai juste avant cette action ».

4.1.2 Définition d'opérateurs STIT

Enfin, Herzig et Lorini [20] définissent l'opérateur $STIT$ tel que $STIT_C\,\varphi$ se lit : « Quelles que soient les actions accomplies par les agents ne faisant pas partie du groupe C, il existe des actions accomplies par les agents du groupe C qui font en sorte que φ soit vrai ». Plus simplement, cette formule peut se lire : « le groupe C fait en sorte que φ soit vrai ». (Voir B pour la définition formelle de cet opérateur.)

Il est utile pour la suite de définir l'opérateur $\overline{STIT}_i$ tel que :

$$\overline{STIT}_i\,\varphi \stackrel{déf}{=} \neg STIT_{AGT\setminus\{i\}}\,\varphi \qquad (\text{Déf}_{\overline{STIT}_i})$$

qui signifie : il n'est pas le cas que, quelle que soit l'action de l'agent i, il existe une action jointe de la coalition $AGT \setminus \{i\}$ qui fait en sorte que φ. Autrement dit, quelles que soient l'action jointe accomplie par la coalition $AGT \setminus \{i\}$, il existe une action de l'agent i dont l'exécution fait en sorte que φ soit faux. En raccourci, cela signifie encore que i peut faire en sorte d'empêcher que φ soit vrai.

4.2 Sémantique

La sémantique associée à la logique $\mathcal{LOCE}$ est une extension de celle de $\mathcal{DLA}$ (voir [20, Sections 2.2 et 2.3] pour une description détaillée). Les $\mathcal{DLA}-frames$ sont des tuples $F = \langle W,\mathscr{R},\mathscr{R}_\Box\rangle$ où :

- W est un ensemble non vide de mondes possible ou d'états ;
- $\mathscr{R} : AGT \times ACT \longrightarrow W \times W$ fait correspondre tout couple agent-action à une relation $\mathscr{R}_{i:a} \subseteq W \times W$ entre mondes possibles ;
- $\mathscr{R}_\Box$ est une relation d'équivalence sur W.

Les $\mathcal{LOCE}-frames$ constituent une extension des $\mathcal{DLA}-frames$ et correspondent à des tuples $F = \langle W,\mathscr{R},\mathscr{B},\mathscr{G},\mathscr{I},\mathscr{R}_\Box\rangle$ où :

- $\mathscr{B} : AGT \longrightarrow W \times W$ fait correspondre chaque agent i à une relation sérielle, transitive et euclidienne $\mathscr{B}_i \subseteq W \times W$ entre mondes possibles ;
- $\mathscr{G} : AGT \longrightarrow W \times W$ fait correspondre chaque agent i à une relation sérielle $\mathscr{G}_i \subseteq W \times W$ entre mondes possibles ;
- $\mathscr{I} : AGT \longrightarrow W \times W$ fait correspondre chaque agent i à une relation sérielle $\mathscr{I}_i \subseteq W \times W$ entre mondes possibles.

Dans ce qui suit, nous écrivons $\mathscr{R}_{i:a}(w) = \{w' \in W \mid (w,w') \in \mathscr{R}_{i:a}\}$ et $\mathscr{R}^{-1}_{i:a}(w) = \{w' \in W \mid (w',w) \in \mathscr{R}_{i:a}\}$. $\mathscr{R}_{i:a}(w)$ est l'ensemble des mondes que l'agent i peut atteindre en exécutant l'action a, et $\mathscr{R}^{-1}_{i:a}(w)$ est l'ensemble des mondes d'où l'agent i vient en exécutant l'action a. Nous ne détaillons pas les contraintes sémantiques sur les relations de $\mathscr{R}$ qui peuvent être trouvées dans [20]. Ces contraintes imposent :

- qu'à chaque monde w il y ait une unique action jointe de tous les agents accomplie en w ;

– qu'il existe exactement un monde successeur de w *via* cette action jointe ;
– qu'il en soit de même pour le monde prédécesseur, et l'action jointe ayant permis d'arriver en w ;

Il en découle que chaque agent accomplit exactement une et une seule action en w, et que celle-ci est exécutée en parallèle de l'action exécutée par chacun des autres agents. D'autre part, nous supposons que :

– si chaque action individuelle d'une action jointe est possible, alors cette dernière est elle aussi possible ;
– deux mondes alternatifs doivent avoir le même historique d'actions jointes de tous les agents (déterminisme du passé).

De même, nous écrivons $\mathscr{R}_{\Box}(w) = \{w' \in W \mid (w,w') \in \mathscr{R}_{\Box}\}$ qui représente l'ensemble des mondes alternatifs au monde w. $\mathscr{R}_{\Box}(w)$ est un ensemble de mondes équivalents représentant le même instant, mais dont chacun fait partie d'une histoire alternative (c'est-à-dire d'une séquence d'actions différente). Par exemple $\varphi \wedge \Diamond\neg\varphi$ signifie que φ est actuellement vrai mais il est possible qu'il soit faut.

De même, $\mathscr{B}_i(w) = \{w' \in W \mid (w,w') \in \mathscr{B}_i\}$, $\mathscr{G}_i(w) = \{w' \in W \mid (w,w') \in \mathscr{G}_i\}$, $\mathscr{I}_i(w) = \{w' \in W \mid (w,w') \in \mathscr{I}_i\}$.

$\mathscr{B}_i(w)$ est l'état de croyance de l'agent i dans le monde w : c'est l'ensemble des mondes qu'il considère depuis le monde w comme des alternatives possibles à ce monde w. Le fait que chaque relation $\mathscr{B}_i$ soit sérielle, transitive et euclidienne signifie respectivement qu'un agent ne peut avoir de croyances contradictoires, et qu'il est conscient de ses croyances et de ses non croyances.

$\mathscr{G}_i(w)$ est l'état des buts de l'agent i dans le monde w : c'est l'ensemble des mondes qu'il souhaite atteindre ou qu'il préfère depuis le monde w. Le fait que chaque relation $\mathscr{G}_i$ soit sérielle signifie que les buts de chaque agent ne sont pas contradictoires.

$\mathscr{I}_i(w)$ est l'état des idéaux de l'agent i dans le monde w : c'est l'ensemble des mondes que l'agent i considère comme idéaux (d'un point de vu moral) depuis le monde w. Le fait que chaque relation $\mathscr{I}_i$ soit sérielle signifie que l'agent n'a pas d'idéaux contradictoires.

Afin d'imposer que tout agent i est conscient de ses (non) buts et de ses (non) idéaux, nous imposons les contraintes supplémentaires suivantes aux $\mathcal{LOCE}-models$: Pour tout monde $w \in W$, pour tout $i \in AGT$:

(CSS1) si $w' \in \mathscr{B}i(w)$ then $\mathscr{G}_i(w') = \mathscr{G}_i(w)$;

(CSS2) si $w' \in \mathscr{B}i(w)$ then $\mathscr{I}_i(w') = \mathscr{I}_i(w)$;

La contrainte CSS1 capture une propriété d'introspection positive et négative pour les buts : les mondes préférés par l'agent i sont aussi ceux qu'il préfère depuis les mondes qu'il considère comme possibles. De même, la contrainte CSS2 capture une propriété d'introspection positive et négative pour les idéaux : les mondes idéaux pour l'agent i sont aussi ceux qui lui sont idéaux depuis les mondes qu'il considère comme possibles.

Un $\mathcal{LOCE}-model$ est une paire ordonnée $M = \langle F,V\rangle$ où F est un $\mathcal{LOCE}-frame$ et $V : ATM \longrightarrow 2^W$ est une fonction de valuation. Les conditions de vérité des différents opérateurs sont alors définis comme suit :

– $M,w \models p$ ssi $w \in V(p)$;
– $M,w \models \neg\varphi$ ssi il n'est pas le cas que $M,w \models \varphi$;
– $M,w \models \varphi \wedge \psi$ ssi $M,w \models \varphi$ et $M,w \models \psi$;

- $M,w \models Bel_i\,\varphi$ ssi $M,w' \models \varphi$ pour tout $w' \in \mathscr{B}_i(w)$;
- $M,w \models Goal_i\,\varphi$ ssi $M,w' \models \varphi$ pour tout $w' \in \mathscr{G}_i(w)$;
- $M,w \models Ideal_i\,\varphi$ ssi $M,w' \models \varphi$ pour tout $w' \in \mathscr{I}_i(w)$;
- $M,w \models Done_{i:a}\,\varphi$ ssi $M,w' \models \varphi$ pour au moins un $w' \in \mathscr{R}^{-1}_{i:a}(w)$;
- $M,w \models Happens_{i:a}\,\varphi$ ssi $M,w' \models \varphi$ pour au moins un $w' \in \mathscr{R}_{i:a}(w)$;
- $M,w \models \Diamond\varphi$ ssi $M,w' \models \varphi$ pour au moins un $w' \in \mathscr{R}_{\Box}(w)$.

4.3 Axiomatique

Comme dans toute logique modale, tous les principes de la logique classique propositionnelle sont satisfaits. Voir A pour la forme logique des principes énoncés ci-dessous et auxquels il est fait référence selon la terminologie utilisée dans [9].

En ce qui concerne les opérateurs de croyance Bel_i, pour chaque agent $i \in AGT$, ils vérifient (RN$_{Bel_i}$) et (K$_{Bel_i}$), ce qui signifie que tout agent i croit toutes les tautologies et toutes les conséquences logiques de ses croyances. C'est une idéalisation de la notion de croyance car dans l'absolue ce n'est pas le cas qu'un individu ait toutes ces croyances, mais dans le cas d'agents artificiels c'est une hypothèse acceptable. Les opérateurs de croyance vérifient également (D$_{Bel_i}$), (4$_{Bel_i}$) et (5$_{Bel_i}$) pour tout $i \in AGT$. Ces propriétés rendent compte du fait qu'un agent i ne peut pas avoir de croyances contradictoires, et qu'il est conscient de ses croyances et de ce qu'il ne croit pas. La logique ainsi définie permet de déduire l'équivalence inverse pour les deux dernières propriétés.

Concernant les buts, les relations sémantiques définies précédemment sont axiomatisées (pour tout $i \in AGT$) par les principes (RN$_{Goal_i}$), (K$_{Goal_i}$) et (D$_{Goal_i}$). Le fait qu'un agent ait pour but toute tautologie est due au fait que la notion de but adoptée inclut les buts à maintenir : l'agent souhaite que ce qui est toujours vrai le reste. Il est purement rationnel dans le sens où il a également pour but toutes les conséquences de ses buts, et ces derniers ne peuvent pas être contradictoires.

Pour ce qui est de ses idéaux, les propriétés sont similaires à celles des buts : pour tout $i \in AGT$, les opérateurs d'idéaux vérifient (RN$_{Ideal_i}$), (K$_{Ideal_i}$) et (D$_{Ideal_i}$). Cela signifie que toute tautologie est un ideal de l'agent i qui devrait être vrai, que toutes les conséquences de ses idéaux sont des idéaux, et qu'il n'a pas d'idéaux contradictoires.

Afin de respecter les contraintes sémantiques (CSS1) et (CSS2), nous définissons les axiomes suivants :

$$Goal_i\,\varphi \rightarrow Bel_i\,Goal_i\,\varphi \qquad (\text{IP}_{Bel_i})$$
$$\neg Goal_i\,\varphi \rightarrow Bel_i\,\neg Goal_i\,\varphi \qquad (\text{IN}_{Bel_i})$$
$$Ideal_i\,\varphi \rightarrow Bel_i\,Ideal_i\,\varphi \qquad (\text{IP}_{Ideal_i})$$
$$\neg Ideal_i\,\varphi \rightarrow Bel_i\,\neg Ideal_i\,\varphi \qquad (\text{IN}_{Ideal_i})$$

Du fait de (D$_{Bel_i}$), il est facile de prouver les propriétés inverses des quatres propriétés ci-dessus. Ainsi, ces propriétés signifient qu'un agent a un but (resp. un idéal) ssi il croit qu'il a ce but (resp. cet idéal) et qu'un agent n'a pas un but (resp. un idéal) ssi il croit qu'il n'a pas ce but (resp. cet idéal). Autrement dit, il est conscient des buts et idéaux qu'il a et qu'il n'a pas.

L'axiomatique des opérateurs d'action est définie dans [20] et une partie est présentée C. Des principes similaires peuvent être définis pour les opérateurs dynamiques dans le

passé que nous avons ajoutés. Il s'agit simplement de faire dans les cinq axiomes définis dans C les substitutions suivantes : $Happens$ par $Done$ et $After$ par $Before$. Il convient en outre de rendre compte du fait que les relations d'accessibilité associées aux opérateurs $Done_{i:a}$ et $Happens_{i:a}$ sont l'inverse l'une de l'autre. Cela est fait en introduisant les deux axiomes de conversion suivants :

$$\varphi \rightarrow \neg Happens_{i:a} \neg Done_{i:a} \varphi \quad \text{(Conv1)}$$
$$\varphi \rightarrow \neg Done_{i:a} \neg Happens_{i:a} \varphi \quad \text{(Conv2)}$$

qui signifient respectivement que si φ est vrai, c'est que nécessairement après toute exécution de $i{:}a$, $i{:}a$ vient juste d'être exécutée, et avant toute exécution de $i{:}a$, $i{:}a$ est sur le point de se produire.

Enfin, on peut montrer que les opérateurs X et X^{-1} sont des opérateurs normaux de type KD et qu'ils respectent (RN_X), (K_X) et (D_X) (et la même chose pour X^{-1}). Cela signifie que toute formule valide le reste l'instant suivant, que si une implication est vraie l'instant suivant alors si son antécédent est vrai à ce même instant alors son conséquent sera également vrai à ce même instant, et que si une formule est vraie dans l'instant suivant alors il est faut que son contraire est vraie à ce même instant.

4.4 Complétude

La procédure de complétude est une extension de celle faite pour la logique $\mathcal{DLA}$ dans [20]. Des axiomes similaires à ceux des opérateurs $Happens_{i:a}$ sont introduits pour les opérateurs $Done_{i:a}$ ainsi que les deux axiomes de conversion ci-dessus qui sont des axiomes de Sahlqvist [31].

5 Formalisation d'émotions complexes

L'ambition du langage utilisé est non seulement de pouvoir capturer les émotions de honte et de regret, mais également d'autres émotions que l'agent artificiel que l'on cherche à modéliser pourrait être amené à utiliser.

5.1 Formalisation d'émotions simples

Ce que nous appelons « émotions simples » correspond à des émotions ne faisant intervenir que des structures cognitives simples, essentiellement en termes de congruence ou d'incongruence entre les buts ou les normes internalisées d'une part, et les croyances d'autre part.

Par exemple, quand un agent croit que φ est vrai et qu'il a justement pour but qu'il le soit, alors il ressent de la joie à propos du fait que φ soit vrai (voir [29, 24, 17] par exemple). À l'opposé, s'il a pour but que φ soit faux, alors il éprouve de la tristesse. De façon similaire, on peut remplacer les buts par les normes : si l'agent croit que φ est vrai et qu'il pense que φ est un idéal qu'il pense devoir respecter (norme internalisée), alors il ressent de l'approbation. S'il pense au contraire que φ devrait être faux, il éprouve de la réprobation. Ces différentes émotions sont résumées dans le tableau suivant (extrait de

[18]) :

$\wedge$	$Goal_i\,\varphi$	$Goal_i\,\neg\varphi$	$Ideal_i\,\varphi$	$Ideal_i\,\neg\varphi$
$Bel_i\,\varphi$	$Joy_i\,\varphi$	$Sadness_i\,\varphi$	$Approval_i\,\varphi$	$Disapproval_i\,\varphi$

Il est important de noter que parfois certains utilisent l'intention [11] et non le but (ou de désir). L'argument avancé est qu'en utilisant une notion plus faible que l'intention un agent aurait en permanence des émotions générées par le fait qu'il a intrinsèquement des désirs et des buts. Néanmoins, nous pensons que nous avons beaucoup d'émotions non reliées à des intentions particulières. Par exemple, si un matin en ouvrant ses volets après plusieurs jours de pluie, un agent qui aime le beau temps découvre qu'il fait beau, il est probable qu'il va éprouver de la joie. Pourtant, il n'avait pas nécessairement l'intention qu'il fasse beau (et, au sens de Cohen et Levesque, c'est même impossible). Il semble donc que les émotions ne requièrent pas nécessairement des attitudes mentales motivationnelles aussi fortes que l'intention. En revanche, il est évident que nous ne sommes pas sans arrêt dans un état « d'incontinence émotionnelle », éprouvant en permanence des sentiments forts à propos de toute chose. L'explication à cela réside selon nous dans l'intensité des émotions. La théorie de l'activation d'Anderson *et al*. (voir [3, 4] par exemple) explique bien selon nous ce mécanisme : ne sont accessibles à la conscience que les informations ayant une intensité (ce qu'Anderson *et al*. appellent « l'activation ») supérieure à un certain seuil. En d'autres terme, nous connaissons des choses, nous avons certains buts, nous éprouvons certaines émotions... mais nous n'en sommes pas nécessairement conscients à partir du moment où ces connaissances, ces buts ou ces émotions ne sont pas dans le focus de notre conscience et ont une intensité trop faible pour y entrer.

5.2 Formalisation de la responsabilité et de la culpabilité

L'émotion de culpabilité a été largement étudiée en psychologie [37, 35, 36, 24] et il est établi comme nous l'avons dit plus haut qu'elle inclut une notion de responsabilité de celui qui éprouve l'émotion à propos d'un certain fait φ alors même que φ ou sa négation est un idéal (norme internalisée) de l'agent. Lorini et Schwartzentruber [27] introduisent un cadre formel complet pour le regret basé sur l'opérateur STIT et la notion de responsabilité. Dans [2], cette logique est présentée de manière didactique et d'autres émotions complexes sont définies par Guiraud *et al.* [18] ainsi que les actes de langage permettant de les exprimer. Dans ce qui suit, nous montrons que la logique présentée permet de formaliser une notion de responsabilité ainsi que la culpabilité.

Cette notion de responsabilité est une notion de responsabilité faible au sens où l'on est responsable du fait que φ soit vrai à partir du moment où nous pouvions faire en sorte que φ soit faux. Cela ne préjuge pas du caractère *a priori* bon ou mauvais de φ, et comme nous l'avons dit auparavant il s'agit donc d'une notion différente de la notion légale entraînant une punition. Cette notion de responsabilité peut être capturée de la façon suivante :

$$Resp_i\,\varphi \stackrel{déf}{=} \varphi \wedge X^{-1}\overline{STIT}_i\,X\varphi \qquad (\text{Déf}_{Resp_i})$$

Autrement dit, l'agent i est responsable du fait que φ soit vrai si et seulement si φ est vrai et que l'instant juste avant, i aurait pu empêcher le fait qu'il le devienne l'instant juste

11. La notion d'intention ici est celle définie dans [10], *i.e*. un but qu'on persiste à chercher à réaliser jusqu'à ce qu'il soit atteint ou qu'on découvre qu'il soit impossible à atteindre.

après.[12]

Supposons que Jean ait aidé François à sortir une table de ping-pong dehors, et que suite à un orage de grêle la table se soit abîmée. Dans ce cas, ils sont chacun responsables de ce qui est arrivé, puisqu'ils auraient pu choisir de ne pas aider l'autre à sortir la table dehors (qui aurait alors été à l'abri de l'orage).

Des travaux comme ceux d'Ortony *et al*. par exemple [29] ont déjà été formalisés en logique (voir notamment [1, 34]). Cependant, ces modèles ne définissent pas des émotions telles que la culpabilité, ou le font sans introduire cette dimension relative à ce que l'agent aurait pu faire d'autre.

Par ailleurs, un individu éprouve de la culpabilité par rapport à la violation d'une certaine norme telle que cet individu la conçoit. Elle ne fait pas appel au regard des autres et la notion de norme utilisée dans le cadre de la culpabilité est donc une notion de norme internalisée, c'est-à-dire une norme que l'agent a fait sienne, qu'il a adoptée [8]. Supposons qu'un agent croit que dans une certaine institution, ou au sein d'un certain groupe, il existe une norme indiquant que φ doit être vrai. Supposons en outre que cet agent s'identifie comme membre de cette institution ou de ce groupe. Dans ce cas, l'agent adopte la norme, c'est-à-dire que la norme externe devient un idéal de l'agent.[13]

Ainsi, il convient de définir la culpabilité comme le fait de ne pas avoir empêché la violation de ce qu'on considère comme une norme à respecter :

$$Guilt_i\,\varphi \stackrel{déf}{=} Bel_i\,Resp_i\,\varphi \wedge Ideal_i\,\neg\varphi \qquad (\text{Déf}_{Guilt_i})$$

Cette définition, modulo notre définition de la responsabilité qui est légèrement différente, correspond à la notion définie par ailleurs dans [2].

5.3 Formalisation de la honte

Afin de formaliser la honte, nous reprenons dans ce qui suit les différences essentielles entre honte et culpabilité telles que nous les avons définies plus haut (cf. Section 3).

Comme toute émotion, la honte est à propos de quelque chose. Formellement, il s'agit donc de définir un opérateur du type $Shame_i\,(\varphi,C)$ (qui se lit : « l'agent i à honte vis-à-vis du groupe d'agents C du fait que φ soit vrai ») plutôt que du type $Shame_i\,(C)$ (qui se lit : « l'agent i à honte vis-à-vis du groupe d'agents C »).

Le groupe C apparaît dans cette définition car, comme nous l'avons montré précédemment, la honte se manifeste vis-à-vis d'une audience (non nécessairement présente).

Cette audience se manifeste en premier lieu au travers des idéaux qui ont été violés. Comme nous l'avons souligné précédemment, et contrairement à la culpabilité, les idéaux mis en jeu ne sont pas nécessairement des normes internalisées de l'agent lui-même : il

12. Dans [27, 2], ainsi que dans [18] (bien que le langage soit différent), la responsabilité est définie par l'expression : $\varphi \wedge \overline{STIT}_i\,\varphi$. Nous trouvons cette définition moins intuitive, dans le sens où la réalisation de φ et l'accomplissement de l'action y conduisant sont vrais simultanément. Notre langage, qui est plus expressif, permet d'introduire une nuance temporelle supplémentaire : on est responsable du fait que φ après avoir fait en sorte, l'instant juste avant, que φ soit vrai l'instant suivant.

13. Quand il ne fait pas ce processus d'internalisation, l'agent se retrouve *de facto* au ban de cette institution ou de ce groupe puisqu'il n'en respecte pas les règles et n'éprouve aucun sentiment de culpabilité d'avoir violé cette norme puisque de son point de vue, il ne s'agit pas d'une norme à respecter. On retrouve ici la notion de société basée sur la culpabilité.

suffit que l'agent croit que ce sont des normes d'un certain groupe d'agent (celui vis-à-vis duquel il a honte). Nous définissons ainsi la notion de norme partagée par un certain groupe de la manière suivante :

$$SIdeal_C\varphi \stackrel{déf}{=} \bigwedge_{i\in C} Ideal_i \qquad (\text{Déf}_{SIdeal_C})$$

Autrement dit, φ est un idéal partagé (*shared ideal*) par le groupe C si et seulement si φ est un idéal internalisé pour chacun des individus de ce groupe C. Nous n'imposons aucune autre propriété concernant cette norme partagée, en particulier des propriétés souvent associées aux groupes tel que par exemple le fait que chaque agent sache que les autres agents du groupe partagent cet idéal. C'est une structure minimale qui ne nécessite pas de supposer que le groupe soit constitué de manière formelle (*i.e*. où chacun a un rôle particulier assigné par des règles institutionnelles constitutives du groupe). Cette définition n'impose pas non plus que l'agent qui éprouve de la honte fasse partie de C et ne l'interdit pas non plus.

Nous avons également vu que dans le cas de la honte, l'agent a tendance à nier sa responsabilité (indépendamment de sa responsabilité réelle). En d'autres termes, on pense ne pas être responsable de la situation honteuse. Formellement, cela revient à écrire qu'il était inévitable que cette situation se produise, c'est-à-dire qu'elle est vraie indépendamment de ce que chacun des agents a fait l'instant juste avant. Soit :

$$Inevitable\varphi \stackrel{déf}{=} X^{-1}\Box X\varphi \qquad (\text{Déf}_{Inevitable})$$

qui se lit « il était inévitable que φ soit vrai » si et seulement si l'instant juste avant, il était nécessairement vrai (c'est-à-dire, indépendamment des actions de tous les agents, lui inclus) que φ serait vrai l'instant suivant. Mais ce qui importe dans la honte, ce n'est pas tant la réalité *objective* que d'*avoir le sentiment* de ne pas être responsable de cette situation, qu'elle était inévitable *même si ce n'était réellement pas le cas*. Soit :

$$Inevitable_i\varphi \stackrel{déf}{=} X^{-1}Bel_i\Box X\varphi \qquad (\text{Déf}_{Inevitable_i})$$

Autrement dit, du point de vue de l'agent i il était inévitable que φ si et seulement si l'instant juste avant, i croyait qu'il était nécessairement vrai (c'est-à-dire, indépendamment des actions de tous les agents, lui inclus) que φ serait vrai l'instant suivant.

Le fait de définir la honte à l'aide de l'opérateur $Inevitable_i$ plutôt que de l'opérateur $Inevitable$ a l'avantage de ne pas conditionner le ressenti de la honte au fait qu'il sache aujourd'hui qu'en fait, la situation actuelle était réellement inévitable. Ce qui suffit, c'est qu'il pense que l'instant juste avant, il pensait que la situation était inévitable même si aujourd'hui il se rend compte que ce n'était pas vrai.

Comme nous l'avons montré auparavant, l'aspect public de la honte est relatif à une audience qui n'est pas nécessairement physiquement présente, ni nécessairement au courant des faits. Cet aspect a été pris en compte dans le fait que la honte se manifeste par rapport à un idéal internalisé par un groupe d'individus. Il serait trop fort d

5.3.1 Définition formelle de la honte.

En définitive, si on reprend chacun des éléments venant d'être développés, nous obtenons la définition suivante où $i \notin C$:

$$Shame_i\,(\varphi,C) \stackrel{déf}{=} Bel_i\,(SIdeal_C\neg\varphi \wedge Inevitable_i\varphi \wedge \varphi \wedge Dread_i(\varphi,C)) \qquad (\text{Déf}_{Shame_i})$$

Ainsi, l'agent i a honte du fait que φ soit vrai vis-à-vis du groupe C si et seulement si il croit que : i) idéalement pour le groupe C, φ devrait être faux ; ii) il était inévitable de son point de vue que φ soit vrai ; iii) φ est vrai ; iv) il redoute le fait que C l'apprenne.

5.3.2 Conséquences logiques

Une propriété intéressante est la relation entre ce qui est nécessairement vrai (donc, indépendant de ce que font les agents) et les actions des agents. Ainsi, selon [20], on peut montrer que pour tout groupe d'agents C donné :

$$\Box\varphi \rightarrow STIT_C\,\varphi \qquad (1)$$

qui se lit : « si φ est inévitable, alors n'importe quelle coalition C fait en sorte que φ soit vrai ».

Théorème 1. *Pour tout agent $i \in AGT$:*

$$\overline{STIT}_i\,\varphi \rightarrow \neg\Box\varphi$$

Autrement dit, si un agent i donné peut faire en sorte d'empêcher que φ soir vrai, alors nécessairement φ n'est pas inévitable.

Démonstration. Directement de ($\text{Déf}_{\overline{STIT}_i}$) et de la contraposée de (1). □

Théorème 2. *Pour tout agent $i \in AGT$ et coalition $C \in 2^{AGT}$:*

$$Resp_i\,\varphi \rightarrow \neg Inevitable\varphi \qquad (a)$$

$$X^{-1}Bel_i\,\overline{STIT}_i\,X\varphi \rightarrow \neg Inevitable_i\varphi \qquad (b)$$

$$Shame_i\,(\varphi,C) \rightarrow Bel_i\,\neg X^{-1}Bel_i\,\overline{STIT}_i\,X\varphi \qquad (c)$$

Le théorème (2a) signifie que pour tout agent i donné, s'il est responsable du fait que φ soit vrai alors c'est qu'il n'était pas inévitable que φ soit vrai. (Par contraposition, nous avons également que si φ était inévitable alors l'agent i n'est pas responsable du fait que φ soit vrai.) Le théorème (2b) signifie que pour tout agent i donné, si l'instant d'avant i croyait qu'il pouvait faire en sorte d'empêcher le fait que φ soit vrai l'instant juste après, alors de son point de vue il n'était pas inévitable que φ soit vrai. Il s'agit là d'un point de vue du présent sur ce qui était vrai pour l'agent i l'instant juste avant (indépendamment de ce qui était réellement vrai). La contraposition de ce théorème se lit : si l'instant d'avant il était inévitable du point de vue de l'agent i que φ soit vrai, alors il n'est pas le cas qu'à cet instant i croyait pouvoir faire en sorte d'empêcher que φ soit vrai l'instant suivant. Enfin, le théorème (2c) signifie que, pour tout agent i donné et tout groupe C d'agents, si l'agent i éprouve de la honte vis-à-vis de C par rapport à φ alors i croit qu'il n'est pas le cas

que l'instant d'avant il croyait pouvoir faire en sorte d'empêcher que φ soit vrai l'instant suivant. On retrouve là une propriété importante de la honte et qui a été discutée plus haut, à savoir que lorsqu'une personne éprouve de la honte, elle a le sentiment qu'au moment où a lieu la violation de la norme qui entraînera ensuite la honte, elle n'avait aucun moyen d'empêcher cette situation de se produire.

Démonstration. Théorème (2a) : par (RN$_{X^{-1}}$) puis (K$_{X^{-1}}$) sur Théorème 1 instancié par $X\varphi$, on obtient $X^{-1}\overline{STIT}_i\, X\varphi \rightarrow X^{-1}\neg\Box X\varphi$. D'une part, on sait que $Resp_i\,\varphi \rightarrow X^{-1}\overline{STIT}_i\, X\varphi$ par (Déf$_{Resp_i}$). D'autre part, par (D$_{X^{-1}}$) puis par (Déf$_{Inevitable}$) on obtient que $X^{-1}\overline{STIT}_i\, X\varphi \rightarrow \neg Inevitable\varphi$. Donc finalement, $Resp_i\,\varphi \rightarrow \neg Inevitable\varphi$.

Théorème (2b) : en appliquant (RN$_{Bel_i}$) puis (K$_{Bel_i}$) sur Théorème 1 on obtient que $Bel_i\,\overline{STIT}_i\, X\varphi \rightarrow Bel_i\,\neg\Box X\varphi$. Puis en appliquant sur ce résultat (RN$_{X^{-1}}$) puis (K$_{X^{-1}}$), on obtient que $X^{-1}Bel_i\,\overline{STIT}_i\, X\varphi \rightarrow X^{-1}Bel_i\,\neg\Box X\varphi$, ce qui implique par (D$_{Bel_i}$) puis par (D$_{X^{-1}}$), on a $X^{-1}Bel_i\,\overline{STIT}_i\, X\varphi \rightarrow \neg X^{-1}Bel_i\,\Box X\varphi$, dont le conséquent est par (Déf$_{Inevitable_i}$) égal à $\neg Inevitable_i\varphi$.

Théorème (2c) : d'une part, par (Déf$_{Shame_i}$) et (M$_{Bel_i}$) on obtient que $Shame_i\,(\varphi,C) \rightarrow Bel_i\; Inevitable_i\varphi$; d'autre part, par contraposition de (2b) puis par (RN$_{Bel_i}$) on obtient que $Bel_i\; Inevitable_i\varphi \rightarrow Bel_i\,\neg X^{-1}Bel_i\,\overline{STIT}_i\, X\varphi$. □

En fait, comme l'analyse informelle l'avait déjà laissé entendre, on voit bien ainsi que la honte, la culpabilité et le regret ne sont pas liés par quelque lien logique que ce soit. Ce sont donc des émotions distinctes qui peuvent cohabiter.

Enfin, pour reprendre l'example de Mathilde de la Molle dans les exemples Section 3.2, elle éprouve de la honte de constater qu'elle est tombée amoureuse du fils d'un charpentier, tout en éprouvant de la culpabilité de tromper son mari. Plusieurs émotions complexes peuvent donc co-exister simultanément.

Autre cas intéressant, supposons que l'agent éprouve de la honte vis-à-vis du groupe C à propos de φ, groupe pour lequel idéalement φ devrait être faux. Supposons en outre que le groupe C en vienne à apprendre que φ est vrai. De ce fait, nous avons que :

$Ideal_j\,\neg\varphi$	pour tout agent $j \neq i$
$Bel_j\,\varphi$	pour tout agent $j \neq i$

Ainsi, alors que i éprouve de la honte, les agents du groupe C éprouvent un sentiment de désapprobation (*cf.* Tableau des émotions simples page 15). Ce sera également le cas de l'agent i si $\neg\varphi$ est un idéal qu'il a internalisé.

6 Conclusion

Nous avons montré et formalisé les différences essentielles entre la honte et la culpabilité. Bien sûr, des notions comme la responsabilité d'un individu (au sens causal du terme) ou comme l'idéal d'un groupe sont des notions complexes qui méritent à elles seules une étude approfondie. Dans la réalité, nous adaptons même souvent ce concept à la situation présente. Par exemple, on peut avoir honte vis-à-vis d'une institution ou d'une personne morale (ce qui requiert un groupe de type structuré, avec des rôles, *etc*.) mais on peut aussi éprouver de la honte vis-à-vis d'un groupe non structuré (un simple ensemble d'agents). C'est pour cela que nous nous sommes contenté de notions simples.

7 Travaux futurs

Concernant les deux émotions formalisées, nous souhaitons nous focaliser maintenant sur les tendances à l'action (ce qu'un individu est tenté de faire lorsqu'il éprouve une telle émotion).

Les travaux réalisés sur la honte et la culpabilité ouvrent la voie à des études similaires sur d'autres émotions. A long terme, il est également possible de faire le lien entre les études sur les émotions et celles sur les actes de langages de type expressif. L'aboutissement pourrait être de coupler ces deux travaux avec des agents conversationnels animés de type GRETA capable d'exprimer des émotions de manière faciale, vocale et gestuelle.

On peut également orienter la suite de ce travail vers la mise en place d'expérimentations destinées à tester la validité des définitions données.

Remerciements

Ce travail a été soutenu par le contrat de recherche CECIL (Complex Emotions in Communication, Interaction and Language) N°ANR-08-CORD-005 obtenu auprès de l'ANR suite à l'appel à projet ContInt 2008. Site web du projet : `www.irit.fr/CECIL/`.

Références

[1] C. ADAM. *Emotions: from psychological theories to logical formalization and implementation in a BDI agent*. Ph.D. thesis, INP Toulouse, France, Jul. 2007.

[2] C. ADAM, B. GAUDOU, D. LONGIN, E. LORINI. *Logical modeling of emotions for Ambient Intelligence*. Dans *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives*, réds. F. Mastrogio, N.-Y. Chong, IGI Global, 2011.

[3] J. ANDERSON, C. LEBIERE. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ, 1998.

[4] J. R. ANDERSON, D. BOTHELL, M. D. BYRNE, S. DOUGLASS, C. LEBIERE, Y. QIN. *An integrated theory of the mind*. *Psychological Review*, vol. 111, p. 1036–1060, 2004.

[5] N. BELNAP, M. PERLOFF, M. XU. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York, 2001.

[6] R. BENEDICT. *The chrysanthemum and the sword*. Mariner Books, 1946.

[7] H. N. CASTANEDA. *Thinking and Doing*. D. Reidel, Dordrecht, 1975.

[8] C. CASTELFRANCHI, E. LORINI. *Cognitive Anatomy and Functions of Expectations*. Dans *IJCAI03 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions*. Morgan Kaufmann, Acapulco, Mexico, August 9-11 2003.

[9] B. F. CHELLAS. *Modal Logic: an Introduction*. Cambridge University Press, Cambridge, 1980.

[10] P. R. COHEN, H. J. LEVESQUE. *Intention is Choice with Commitment*. *Artificial Intelligence Journal*, vol. 42(2–3), p. 213–261, 1990.

[11] R. CONTE, C. CASTELFRANCHI. *Cognitive and social action*. London University College of London Press, London, 1995.

[12] C. R. DARWIN. *The expression of emotions in man and animals*. Murray, London, 1872.

[13] R. M. DE PISON. *Death by Despair: Shame And Suicide*. Peter Lang Pub Inc, 2006.

[14] P. EKMAN, R. LEVENSON, W. FRIESEN. *Autonomic nervous system activity distinguishes among emotions. Science*, vol. 221, p. 1208–1210, 1983.

[15] J. ELSTER. *Alchemies of the Mind: Rationality and the Emotions*. Cambridge University Press, Cambridge, 1999.

[16] N. H. FRIJDA. *The Emotions*. Cambridge University Press, 1986.

[17] R. GORDON. *The structure of emotions*. Cambridge University Press, New York, 1987.

[18] N. GUIRAUD, D. LONGIN, E. LORINI, S. PESTY, J. RIVIÈRE. *The face of emotions: a logical formalization of expressive speech acts (regular paper)*. Dans *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), Taipei, Taiwan, 02/05/2011-06/05/2011*. ACM, p. 1031–1038, 2011.

[19] D. HAREL, D. KOZEN, J. TIURYN. *Dynamic Logic*. MIT Press, Cambridge, 2000.

[20] A. HERZIG, E. LORINI. *A dynamic logic of agency I: STIT, capabilities, and powers. Journal of Logic, Language, and Information*, vol. 19, p. 89–121, 2009.

[21] J. F. HORTY. *Agency and Deontic Logic*. Oxford University Press, Oxford, 2001.

[22] J. F. HORTY, N. BELNAP. *The deliberative STIT: A study of action, omission, and obligation. Journal of Philosophical Logic*, vol. 24(6), p. 583–644, 1995.

[23] W. JAMES. *What is an emotion? Mind*, vol. 9, p. 188–205, 1884.

[24] R. S. LAZARUS. *Emotion and Adaptation*. Oxford University Press, 1991.

[25] H. B. LEWIS. *Shame and guilt in neurosis*. International Universities Press, New-York, 1971.

[26] E. LORINI. *A Dynamic Logic of Knowledge, Graded Beliefs and Graded Goals and Its Application to Emotion Modelling*. Dans *Proceedings of the LORI-III Workshop on Logic, Rationality and Interaction, Guangzhou, P.R.China, 10/10/2011-13/10/2011*, réds. H. van Ditmarsch, J. Lang, S. Ju. Springer-Verlag, vol. 6953 de *LNAI*, p. 165–178, 2011.

[27] E. LORINI, F. SCHWARZENTRUBER. *A logic for reasoning about counterfactual emotions. Artificial Intelligence*, vol. 175(3-4), p. 814–847, 2011.

[28] M. MICELI, C. CASTELFRANCHI. *How to silence one's conscience: Cognitive defenses against the feeling of guilt. Journal for the Theory of Social Behaviour*, vol. 28, p. 287–318, 1998.

[29] A. ORTONY, G. CLORE, A. COLLINS. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA, 1988.

[30] PLATON. *La République*. Flammarion, 2002.

[31] H. SAHLQVIST. *Completeness and correspondence in the first and second order semantics for modal logics*. Dans *Proceedings of the 3rd Scandinavian Logic Symposium*, réd. S. Kanger. vol. 82 de *Studies in Logic*, 1975.

[32] D. SANDER, K. SCHERER, réds. *Traité de psychologie des émotions*. Cognitive. Dunod, 2009.

[33] J. SEARLE. *Rationality in Action*. MIT Press, Cambridge, 2001.

[34] B. STEUNEBRINK, M. DASTANI, J.-J. MEYER. *The OCC model revisited*. Dans *Proc. of the 4th Workshop on Emotion and Computing*, réd. D. Reichardt, 2009.

[35] J. P. TANGNEY. *The self-conscious emotions: shame, guilt, embarrassment and pride*. Dans *Handbook of Cognition and Emotion*, réds. T. Dalgleish, M. Power, John Wiley & Sons, 1999.

[36] J. P. TANGNEY, R. L. DEARIN. *Shame and Guilt*. The Guilford Press, 2002.

[37] J. P. TANGNEY, R. S. MILLER, L. FLICKER, D. H. BARLOW. *Are shame, guilt, and embarrassment distinct emotions? Journal of Personality and Social Psychology*, vol. 70(6), p. 1256–1269, 1996.

[38] G. TAYLOR. *Pride, Shame, and Guilt: Emotions of Self-Assessment*. Oxford University Press, New-York, 1985.

[39] S. TOMKINS. *Affect theory*. Dans *Approaches to emotion*, réds. K. Scherer, P. Ekman, Erlbaum, Hillsdale, NJ, p. 163–196, 1984.

A Principaux schémas d'axiomes de la logique modale

Dans cette section sont détaillés les principaux schémas axiomes et règles d'inférence de la logique modale qui nous servent dans l'article. Ceux-ci sont donnés pour un opérateur générique $\bullet$ (*bullet*) qu'il suffit de substituer par l'opérateur voulu. Ainsi, (K_{Bel_i}) correspond à ($\mathrm{K}_\bullet$) dans lequel on a substitué $\bullet$ par Bel_i, ce qui donne pour (K_{Bel_i}) le schéma d'axiome $Bel_i\,(\varphi \to \psi) \to (Bel_i\,\varphi \to Bel_i\,\psi)$.

$$\frac{\varphi}{\bullet\varphi} \qquad (\mathrm{RN}_\bullet)$$

$$\bullet\,(\varphi \wedge \psi) \to (\bullet\varphi \wedge \bullet\psi) \qquad (\mathrm{M}_\bullet)$$

$$(\bullet\varphi \wedge \bullet\psi) \to \bullet(\varphi \wedge \psi) \qquad (\mathrm{C}_\bullet)$$

$$\bullet\,(\varphi \to \psi) \to (\bullet\varphi \to \bullet\psi) \qquad (\mathrm{K}_\bullet)$$

$$\bullet\,\varphi \to \neg \bullet \neg\varphi \qquad (\mathrm{D}_\bullet)$$

$$\bullet\,\varphi \to \bullet\bullet\,\varphi \qquad (4_\bullet)$$

$$\neg \bullet\,\varphi \to \bullet\neg \bullet\,\varphi \qquad (5_\bullet)$$

B Définition réductionniste du STIT

L'opérateur STIT (pour *seeing to it that*, c'est-à-dire faire en sorte que) est un opérateur permettant de décrire l'état du monde après l'exécution d'une action non explicitée, ce dont nous nous servirons pour définir la notion de responsabilité. La logique modale du STIT a été proposée pour la première fois en logique philosophique au dans les années 90 [5, 22, 21]. La logique STIT supporte le raisonnement sur les actions des agents ainsi que sur les actions jointes des groupes. Dans le reste de cette section, nous montrons comment l'opérateur $STIT_i$ tel qu'il est défini dans [20] nous permet de définir l'opérateur $\overline{STIT}_i$.

Soit $\delta = \langle 1{:}a_1, 2{:}a_2, \ldots n{:}a_n\rangle$ (où n est le cardinal de AGT) l'*action jointe accomplie par tous les agents de* AGT où $i{:}a_i$ est l'action individuelle a_i accomplie par l'agent i et notée δ^i (*i.e.*, $\delta = \langle \delta^1, \delta^2, \ldots \delta^n\rangle$) et où toutes les actions individuelles δ^i ($i \in AGT$) sont accomplies simultanément.

On note $\Delta \stackrel{déf}{=} \prod_{i\in AGT}\{i{:}a \mid a \in ACT\}$ l'ensemble de toutes les actions jointes possibles de tous les agents. Par exemple, si $ACT = \{a_1, a_2, a_3\}$ et $AGT = \{1,2\}$, alors Δ est le produit cartésien de $\{1{:}a_1, 1{:}a_2, 1{:}a_3\}$ (l'ensemble de toutes les actions possibles accomplies par l'agent 1) par $\{2{:}a_1, 2{:}a_2, 2{:}a_3\}$ (l'ensemble de toutes les actions possibles accomplies par l'agent 1), ce qui donne en tout 9 actions jointes possibles.

On appelle $C \subseteq AGT$ une coalition dès lors que les agents contenus dans C accomplissent une action jointe δ_C qui est la restriction de δ aux actions des agents appartenant à C. δ_C est donc l'*action jointe accomplie par la coalition* C. Par exemple, si les agents 1, 3 et 6 forment la coalition $\{1,3,6\}$ au sein de l'ensemble de tous les agents AGT, alors ils accomplissent l'action jointe $\delta_{\{1,3,6\}} = \langle \delta^1, \delta^3, \delta^6\rangle$ (restriction de δ à la coalition $\{1,3,6\}$). Il est important de noter que δ_C correspond à l'action jointe accomplie par la coalition C *durant l'accomplissement de l'action jointe* δ *accomplie par tous les agents de* AGT.

Nous sommes alors en mesure de définir les deux opérateurs supplémentaires sui-

vants :

$$Done_{\delta_C}\,\varphi \stackrel{déf}{=} \bigwedge_{i\in C} Done_{\delta_i}\,\varphi$$

qui se lit « l'action jointe accomplie par la coalition C vient juste d'être accomplie, avant quoi φ était vrai », et :

$$Happens_{\delta_C}\,\varphi \stackrel{déf}{=} \bigwedge_{i\in C} Happens_{\delta_i}\,\varphi$$

qui se lit « la coalition C fait l'action jointe δ_C après quoi φ sera vrai ».

Ainsi, pour toute coalition $C \subseteq AGT$ et action jointe $\delta \in \Delta$, [20] définissent le fait que la coalition C fait en sorte que φ lors de l'accomplissement de l'action jointe δ telle que :

$$STIT_C\,(\delta,\varphi) \stackrel{déf}{=} Happens_{\delta_C}\,\top \wedge \Box(Happens_{\delta_C}\,\top \rightarrow \varphi)$$

qui se lit : « indépendamment de ce que font les autres agents n'appartenant pas à la coalition C, l'action jointe δ_C est exécutée par la coalition C et nécessairement si δ_C est exécutée alors φ est vrai ». Autrement dit, quoi que fassent les agents extérieurs à la coalition C, celle-ci fait en sorte que φ soit vrai en exécutant l'action jointe δ_C lors de l'exécution de l'action jointe δ de tous les agents.

Finalement, pour tout $C \subseteq AGT$, le fait que la coalition C fait en sorte que φ soit vrai est défini par :

$$STIT_C\,\varphi \stackrel{déf}{=} \bigvee_{\delta\in\Delta} STIT_C\,(\delta,\varphi)$$

qui se lit : « il existe au moins une action jointe δ accomplie par tous les agents telle que, quoi que fassent les agents extérieurs à la coalition C, celle-ci fait en sorte que φ soit vrai lors de l'exécution de δ ». En d'autres termes, quoi que fassent les agents extérieurs à la coalition C, C fait en sorte que φ soit vrai.

On peut facilement démontrer les propriétés suivantes :

$$STIT_C\,(\delta,\varphi) \rightarrow \varphi \qquad (\mathrm{T}_{STIT_{C_\delta}})$$

$$STIT_C\,\varphi \rightarrow \varphi \qquad (\mathrm{T}_{STIT_C})$$

Ces résultats sont liés à l'hypothèse selon laquelle les actions ont une durée instantanée: faire en sorte que φ soit vrai a pour résultat immédiat que φ est vrai.

C Axiomatique des opérateurs dynamiques

Dans [20], les opérateurs dynamiques sont simplement définis dans une logique normale K. Soit, pour toute action $i{:}a$ telle que $i \in AGT$ et $a \in ACT$:

$$\frac{\varphi}{After_{i:a}\,\varphi} \qquad (\mathrm{RN}_{After_{i:a}})$$

$$After_{i:a}\,(\varphi \rightarrow \psi) \rightarrow (After_{i:a}\,\varphi \rightarrow After_{i:a}\,\psi) \qquad (\mathrm{K}_{After_{i:a}})$$

qui signifient que si une proposition est une tautologie, alors elle sera vraie après toute action. Par ailleurs, si $\varphi \rightarrow \psi$ est vrai après toute exécution d'une action, alors si φ est également vrai à ce moment-là, ψ l'est également.

Afin de capturer les autres propriétés sémantiques des actions qui ont été définies précédemment, il faut ajouter des contraintes supplémentaires :

$$Happens_{i:a}\,\varphi \rightarrow After_{j:b}\,\varphi \qquad (\text{Alt}_f)$$

$$\bigvee_{a \in ACT} Happens_{i:a}\,\top \qquad (\text{Act}_f)$$

$$Happens_{i:a}\,\top \rightarrow After_{i:b}\,\bot \qquad \text{si } a \neq b \qquad (\text{Sin}_f)$$

$$\left(\bigwedge_{i \in AGT} \Diamond Happens_{\delta_i}\,\top\right) \rightarrow \Diamond Happens_{\delta}\,\top \qquad (\text{Ind}_f)$$

$$Happens_{\delta}\,\Diamond\varphi \rightarrow \Diamond Happens_{\delta}\,\varphi \qquad (\text{Per}_f)$$

(Alt$_f$) signifie que si une action est accomplie après quoi φ sera vrai, alors n'importe quelle action de n'importe quel agent conduit à un état identique. Cela rend compte non seulement du fait que les actions sont exécutées en parallèle mais également que leur exécution est déterministe [14]. (Act$_f$) signifie que pour chaque agent, il y a toujours au moins une action sur le point d'être exécutée. (Sin$_f$) signifie que si une action est accomplie, alors toutes les autres actions du répertoire d'action de l'agent sont inexécutables. (Ind$_f$) signifie que s'il est possible que chaque agent i exécute une action individuelle δ_i au sein d'une action collective de tous les agents δ, alors il est possible que cette action collective de tous les agents δ soit exécutée. Enfin, (Per$_f$) signifie que si une action jointe est accomplie après quoi il sera possible que φ, alors il est possible (maintenant) que δ soit accomplie après quoi φ sera vrai. (C'est un axiome de permutation entre les opérateurs $\Diamond$ et $Happens_{\delta}$ d'action jointes accomplies par tous les agents.

14. Ainsi, si on prend $i = j$ et $a = b$, cet axiome signifie que si une action est accomplie conduisant à un état où φ est vrai, alors toutes exécution de cette action conduit à un état où φ est vrai. Cela reste vrai dès lors que cette action est accomplie par n'importe quel autre agent $j \neq i$, même si chacun de ces agents effectue une action différente ($b \neq a$).

CNRS - INPT - UPS - UT1 - UTM

Résumé

Une étude de la honte présente un intérêt tant applicatif que théorique. Dans le premier cas, il s'agit de faire en sorte qu'une machine soit capable de détecter de tels sentiments chez l'usager afin de développer des stratégies palliatives et, par voie de conséquence, d'améliorer son efficacité (tutoring intelligent pas exemple). Dans le second cas, il s'agit de replacer la honte parmi les autres émotions, en étudiant non seulement ce qui fait le propre de la honte, mais également ce qui différencie cette dernière des autres émotions, en particulier la culpabilité. Après une brève présentation de ce qu'est l'émotion, le présent article présente dans un premier temps une analyse approfondie de la honte en philosophie et psychologie. Dans un second temps, un langage formel de type logique modale est présenté afin d'offrir un cadre de formalisation d'un ensemble d'émotions, dont la honte, celle-ci (et d'autres) étant ensuite formalisée au sein de ce cadre. Il en découle un cadre unifié propre à représenter des émotions simples telles que la joie ou la tristesse, ou des émotions plus complexes telles que la honte ou la culpabilité.

Mots-clés

Émotions, modal logic, shame, guilt, regret.

UPS-IRIT, 118 route de Narbonne, 31062 Toulouse CEDEX 9
+33 (0) 561 55 67 65 info@irit.fr www.irit.fr