Honte ou culpabilité ? (That is the question)

Carole Adam, Dominique Longin

Honte ou culpabilité ? (*That is the question*)

C. Adam♣ carole.adam@imag.fr

D. Longin♠ Dominique.Longin@irit.fr

♣Laboratoire d'Informatique de Grenoble
Université Joseph Fourier, équipe MAGMA
Maison J. Kuntzmann, 110 avenue de la Chimie, 38400 Saint-Martin-d'Hères

♠Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, équipe LILaC
118 route de Narbonne, 31062 Toulouse cedex 9

Résumé :
Sous ce titre quelque peu humoristique se cache une vraie question : qu'est-ce qui différencie la honte de la culpabilité ? Cette question est d'autant plus importante que de nombreuses études en psychologie montrent qu'on a bien souvent tendance à les assimiler et à nommer l'une pour l'autre. Mais l'intérêt d'une telle étude réside également dans le fait qu'en caractérisant une notion par ce qu'elle est, on caractérise l'autre par ce qu'elle n'est pas. À terme, notre objectif est de fournir une typologie des émotions où chacune se définirait selon ces deux types de (non)propriétés, donnant une cohérence nouvelle à des classifications bien souvent très subjectives. On se propose de formaliser ces deux émotions dans un même langage de type logique modale des états mentaux.

Mots-clés : émotions, honte, culpabilité, logique modale.

1 Introduction

De nombreuses études auprès de sujets humains montrent que honte et culpabilité sont souvent confondues. (Nous montrerons que ce n'est pas sans raison.) Il est intéressant d'étudier ces émotions pour elles-mêmes mais également d'un point de vue méthodologique.

Pour elles-mêmes, car ce sont des émotions morales, en relation avec les normes qu'un individu se fixe. Elster [9, p. 145] souligne combien les normes sociales ont une influence immensément puissante sur le comportement (« *an immensely powerful influence on behavior* »). En particulier, la honte nous touche dans ce que nous avons de plus intime, de plus personnel. Comme le notent Tangney et Ronda [25], 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]

D'un point de vue méthodologique ensuite, car il est fréquent de constater dans la littérature l'existence de deux définitions différentes pour une même émotion (c'est le cas par exemple de l'espoir dans [21] et [14]) ou l'inverse (c'est le cas de la honte et de la culpabilité). Étudier une émotion en mettant en relief ce qu'elle a en commun avec une autre, et ce qui l'en différencie permet en définitive de mieux cerner chacune d'elles.

Enfin, ces deux émotions jouent aussi un rôle prépondérant en situation de décision stratégique : que va penser l'autre de moi si je fais telle ou telle chose ? Et l'autre justement, quel comportement doit-il adopter ? Que faire si un tuteur intelligent détecte que son élève a honte de parler anglais car il pense mal parler ? Ne devrait-il pas mettre en place quelque stratégie visant à le rassurer ?

Dans ce qui suit, nous présentons brièvement ce que nous appelons une (structure cognitive d')émotion (Section 2). Après une analyse de la littérature (Section 3) nous présentons le cadre formel (Section 4) propre à les caractériser (Section 5).

2 Qu'est-ce qu'une émotion ?

« *What is an emotion ?* » est un article de William James [13], l'un des 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. Dans la lignée de Scherer nous adoptons une vision multi-componentielle de l'émotion : *le sentiment* (le ressenti de l'émotion) ; *la réponse psychophysiologique* (accélération

1. Il prend l'exemple suivant : si un agent viole une norme sociale, nous allons refuser de traiter avec lui, ce qui va peut-être engendrer chez lui une perte matérielle quelconque, mais qui va surtout marquer notre mépris ou notre dégoût et engendrer chez lui de la honte. Et plus il nous coûtera de refuser de traiter avec lui, plus sa honte sera importante [9, p. 146].

du rythme cardiaque, de la température corporelle, *etc.*) ; *l'expression motrice* (du visage, de la voix, des gestes) ; *la tendance à l'action* (à ne pas confondre avec l'action elle-même) ; *l'évaluation cognitive* (*appraisal* en anglais). Dans les théories de l'évaluation cognitive cette dernière composante est vue comme le déclencheur des 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 rendue possible par le fait que l'on évalue (consciemment ou non) un stimulus 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 [22].

Dans ce suit, nous nous intéressons aux émotions et non aux humeurs. Une distinction communément admise est que les premières, contrairement aux secondes, sont toujours *à propos de quelque chose* : on sera déçu *de voir son équipe préférée perdre* mais jamais *triste en général* (ce qui correspond plutôt à une humeur de tristesse). Le sentiment de l'émotion transparaît dans le fait que notre agent est introspectif (conscient) de ses émotions. Ce n'est bien sûr qu'une description partielle du phénomène : il est aussi lié à une notion d'intensité de l'émotion non traitée ici pour ne pas compliquer le formalisme (bien que des solutions techniques existent [17]). 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. En définitive, nous formalisons dans ce qui suit ce qui correspond davantage à la structure cognitive de l'émotion, *i.e.* l'état mental nécessaire à son déclenchement.

3 La honte *versus* la culpabilité

D'après [4], des sociétés orientales comme le Japon ou la Chine ont une « culture de la honte » alors que nos sociétés occidentales ont une « culture de la culpabilité » (au sens de *se sentir coupable*). Selon cet auteur, dans les cultures de la culpabilité on restreint le comportement des individus en les rendant coupables. Dans celles 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 public qui ternit leur image est si terrible pour eux.

La honte et la culpabilité ont été largement étudiées en psychologie [26, 24, 25, 14, 21]. Pourtant, ces émotions ont bien souvent été assimilées, ou peu différenciées l'une de l'autre (voir par exemple [25, p. 11–12] pour plus de détails). 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 [27, 14, 9, p. 145]). Ainsi, Ortony *et al.* [21, p. 142–143] voient ces deux émotions comme le fait que l'agent qui les éprouve s'attribue la responsabilité [2] de la violation de l'un de ses propres idéaux (d'où le fait qu'ils classent ces deux émotions comme des *attribution emotions*). La seule chose qui les différencie est l'importance de l'idéal en question, dont la violation est jugée comme inexcusable dans le cas de la honte, et non dans le cas de la culpabilité (qui serait principalement composée à partir de la honte et du regret).

C'est très certainement à H. B. Lewis [15] que l'on doit d'avoir trouvé un critère discriminant (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. De même Elster [9, p. 143–144] définit la honte comme une émotion négative déclenchée par une croyance à propos de sa personne, et la culpabilité comme une émotion négative déclenchée par une croyance à propos de ses actions [3]. (Voir aussi [8, 14, 25] par exemple.)

Cette distinction explique en particulier pour-

2. Il s'agit ici d'une responsabilité causale, *i.e.* au sens où l'agent vient d'accomplir une action ayant causé l'état présent.

3. En ce sens, l'agent est causalement responsable de par ses actions de la situation présente où un de ses idéaux est violé.

quoi 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 [21], 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 » [14, 20] ce qui peut conduire au suicide [10, p. 274]. Plusieurs méthodes sont possibles comme nier tout lien avec la transgression ou insister sur la nature privée des évènements [20]. Dans la cas de la culpabilité, on a plutôt tendance à adopter un comportement actif et réparateur [9, 20] 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 [20]. Elster [9, p. 150], citant en cela [27], indique que la honte peut avoir une cause indépendante de notre bonne volonté, comme avoir des parents pauvres ou devenir vieux.

Dans la culpabilité comme dans la honte, les idéaux mis en jeu ont été internalisés : se sentir coupable de s'être garé sur une place pour personnes handicapées par exemple, c'est se reconnaître dans le fait qu'il est mal de se garer sur de telles places si on n'est pas handicapé ; on considère ce principe comme devant être respecté. Si au contraire on a connaissance de ce principe mais qu'il ne nous paraît pas important de le respecter (idéal non internalisé) alors on pourra me garer sur une telle place sans pour autant se sentir coupable. (Voir [14, p. 240] par exemple.)

Il a été dit que la honte inclut nécessairement une dimension sociale, publique [9, 27, 20, 21, 8], ce qui ne serait pas le cas de la culpabilité. Elster [9, p. 149] par exemple dit que « je ressens de la honte en votre présence parce que je sais que vous me désapprouvez ». Mais des expérimentations [26] montrent que la honte ressentie en dehors de tout groupe témoin est au contraire légèrement plus fréquente que pour la culpabilité. 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 (dont l'objet n'est pas précisé par l'adulte). Ce qui est important n'est donc pas tant que l'objet de notre honte soit connu d'un certain groupe, mais plutôt qu'on *croie* que cela constitue une violation d'ordre moral vis-à-vis de ce groupe. Darwin abonde en ce sens lorsqu'il dit qu'un individu peut éprouver de la honte mais ne pas rougir pour autant ; pour rougir, il faut que l'objet de sa honte ait été découvert [6, p. 352]. Ainsi, le groupe face auquel on éprouve de la honte n'a ni besoin d'être participatif ou physiquement présent [27], ni même d'être au courant de la violation de la norme en question. Lazarus [14, p. 241] souligne même qu'on peut éprouver de la honte vis-à-vis d'une personne décédée. S'il y a une dimension sociale dans la honte et la culpabilité, elle se situe au niveau du groupe d'individus par rapport auquel on se projette (soit-même dans le cas de la honte, ou ses action dans le cas de la culpabilité).

En résumé. 1) Nous nous intéressons aux émotions et non aux humeurs, les notions de honte et de culpabilité que nous capturons sont donc nécessairement à propos de quelque chose.

2) La honte comme la culpabilité nécessitent la violation d'un idéal internalisé.

3) Dans la culpabilité, l'agent pense être causalement responsable de la situation présente (il se focalise sur ses actions) alors que dans la honte, il ne le pense pas (à tort ou à raison) et refuse d'assumer cette responsabilité ou cherche à la minimiser ; c'est lui-même en tant qu'individu qu'il juge dans son ensemble.

4) Ni la honte ni la culpabilité ne requièrent la présence d'une audience témoin de la violation, ni même qu'elle soit au courant des faits. Il suffit qu'on imagine qu'elle le soit.

Exemple 1 (extrait de [9]). La Princesse de Clèves se sent coupable de l'amour qu'elle éprouve pour le Duc de Nemours mais Mathilde de la Mole a honte d'être amoureuse de Julien Sorel : en trompant son mari la princesse de Clèves accomplit une action qui va à l'encontre de ses idéaux mais bien que Mathilde de la Mole puisse se sentir coupable pour les mêmes raisons, elle a surtout honte être tombée amoureuse du fils d'un charpentier ce qui lui ferait perdre la face si des personnes de son rang l'apprenaient.

Exemple 2. Une personne a faisant du shopping dans un magasin oublie involontairement un vê-

tement sur son sac et se dirige vers la sortie du magasin, étant ainsi sur le point de commettre involontairement un vol. Si la personne r responsable du magasin demande à a d'ouvrir son sac, a ne pourra pas 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.

Exemple 3. Une personne perd son pantalon dans la rue. Il s'agit là de la transgression involontaire d'une norme sociale ou culturelle (on ne se promène pas en sous-vêtements dans la rue). L'individu éprouvera donc probablement de la honte d'avoir perdu son pantalon (à condition bien sûr que cette norme ait été internalisée par l'agent : dans le cas contraire c'est qu'il ne la reconnaît pas comme une norme à suivre et il ne peut alors pas éprouver de la honte). À l'inverse, si l'on suppose qu'il l'a fait volontairement (pour provoquer par exemple) il ne devrait normalement pas non plus éprouver de honte. [4]

4 Cadre formel

Notre cadre formel correspond à une extension de celui développé dans [12]. Nous limitons sa présentation à ce qui est nécessaire à la compréhension de la suite. En particulier, nous présentons les différents opérateurs utilisés, mais nous ne présenterons ni la sémantique, ni même l'axiomatique complète associée. (Pour cela, se reporter à [16].)

4.1 Langage de base et attitudes mentales

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 de base est défini comme suit : $\varphi :: p \mid \bot \mid \neg\varphi \mid \varphi \vee \varphi \mid Bel_i\,\varphi \mid Ideal_i\,\varphi \mid SIdeal_i\,\varphi \mid \Diamond\varphi$ 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 mais pense au contraire que φ est vrai dans le monde réel.

$Ideal_i\,\varphi$ 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) [5]. En ce sens, il est moralement responsable de la réalisation de φ. $SIdeal_i\,\varphi$ (*strong ideal*) représente le fait que φ est un idéal particulièrement important pour i et dont la violation est susceptible de lui faire perdre la face (au sens de [21, p. 142–143]). Nous imposons juste que $SIdeal_i\,\varphi \rightarrow Ideal_i\,\varphi$. Nous avons ainsi deux notions d'idéal dont l'une est plus forte que l'autre en gardant un langage simple (pas besoin d'introduire des degrés qui ne seraient qu'un raffinement supplémentaire).

$\Diamond\varphi$ 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 si la succession d'actions qui ont été accomplies jusqu'à présent n'avait pas été celle qu'elle a été (et qui a conduit 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. 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 alternatif considéré) ». Autrement dit, φ est nécessairement vrai (quoi que les agents aient fait). Par définition, nous avons $\Box\varphi \rightarrow \varphi$.

4.2 Opérateurs temporels

En étendant le langage précédent avec deux opérateurs dynamiques (voir [16]), on peut définir $X\varphi$ qui se lit « next φ » et qui signifie que φ sera

4. Il suffit qu'il n'ait pas internalisé l'idéal qu'il viole. Bien sûr, dans certains cas, on peut être prêt à choquer même si on doit ensuite éprouver de la honte pour ce que l'on a fait : cela signifie juste que dans ce cas, on a attribué plus d'importance à son but de choquer qu'à son idéal (qui est alors violé).

vrai l'instant juste après (quelle que soit l'action accomplie par chacun des agents) ; et $X^{-1}\varphi$ qui se lit « next-moins-un φ » et qui signifie que φ était vrai l'instant juste avant (quelle que soit l'action accomplie par chacun des agents).

On impose que $X\Box\varphi \leftrightarrow \Box X\varphi$ ce qui entraîne que $X^{-1}\Box X\varphi \leftrightarrow \Box\varphi$ (ce qui est nécessairement vrai persiste à l'être dans le futur).

4.3 Opérateurs d'agentitude

Les opérateurs d'agentitude (*agency*) [3] servent à capturer le fait qu'un état a été causé par un agent. On trouve initialement cette notion chez Davidson qui tente de définir ce qu'est une action [7, Essay 3]. Formellement (cf. [12, 16]) $[C]\varphi$ se lit : « il existe des actions accomplies conjointement par les agents du groupe C pour lesquelles, quelles que soient les actions accomplies conjointement par les agents ne faisant pas partie de ce groupe C, φ est vrai ». Plus simplement, cette formule peut se lire : « le groupe C fait en sorte que φ soit vrai ». D'où :

$$\langle i\rangle\varphi \stackrel{déf}{=} \neg[AGT \setminus \{i\}]\varphi$$

qui se lit : « il n'est pas le cas [qu'il existe des actions accomplies conjointement par les agents autres que i pour lesquelles, quelles que soit l'action accomplie par i, φ est vrai] ». Autrement dit : « quelles que soient les actions accomplies conjointement par les agents autres que i, il existe une action accomplie par i telle que φ est faux ». En raccourci, cela signifie encore que i peut faire en sorte d'empêcher que φ soit vrai.

5 Formalisation

5.1 Formalisation de la responsabilité

Comme nous l'avons montré précédemment, la culpabilité fait intervenir une notion de responsabilité causale, alors qu'au contraire la honte la rejette. Généralement, cette responsabilité peut prendre deux formes : la responsabilité **directe** (l'agent a accompli une action qui a causé directement la situation présente comme dans « L'agent i casse une tasse ») et la responsabilité **indirecte** (l'agent aurait pu intervenir pour empêcher une situation d'arriver mais il n'a rien fait, comme dans « L'agent aurait pu empêcher la tasse de se casser (mais il n'en a rien fait) »). La première correspond trivialement à l'opérateur d'agentitude ($\varphi \wedge X^{-1}[i]X\varphi$), et nous formalisons donc dans ce qui suit la seconde (bien que ces deux formes de responsabilité puissent intervenir au niveau émotionnel). [5] Soit :

$$Resp_i\,\varphi \stackrel{déf}{=} \varphi \wedge X^{-1}\langle i\rangle X\varphi$$

qui signifie que l'agent i est responsable (indirectement) du fait que φ soit vrai si et seulement si φ est vrai et qu'à l'instant juste avant, i aurait pu empêcher que l'instant juste après φ devienne vrai. [6]

Supposons que Jean autorise François à sortir une plante et que celle-ci meure brûlée par le soleil. Au sens où nous l'avons défini, François a une responsabilité directe dans la mort de la plante et Jean une responsabilité indirecte.

5.2 Formalisation de l'inévitable

Contrairement à la culpabilité, la honte suppose plutôt qu'on nie toute responsabilité dans ce qui arrive. Formellement, cela revient à écrire que l'instant juste avant il **était** inévitable (*i.e.* indépendant des actions des agents) que l'instant juste après cette situation se produise. Soit :

$$Inevitable\varphi \stackrel{déf}{=} X^{-1}\Box X\varphi$$

qui se lit « (l'instant juste avant) il *était* inévitable que φ devienne vrai maintenant ». (Il découle des propriétés précédentes et de cette définition que $Inevitable\varphi \rightarrow \varphi$.)

5.3 Formalisation des idéaux de groupe

Culpabilité et honte peuvent être ressenties (condition non nécessaire) vis-à-vis d'un groupe. Cela signifie qu'on se projette par rapport à ce groupe dont on partage les idéaux. On peut formaliser les idéaux d'un groupe comme suit (pour tout ensemble d'agents $C \in 2^{AGT}$) :

$$Ideal_C\,\varphi \stackrel{déf}{=} \bigwedge_{i\in C} Ideal_i$$

5. En fait, la nature de la responsabilité (directe ou indirecte, intentionnelle ou non, *etc.*) joue davantage un rôle au niveau de l'intensité de l'émotion qu'au niveau sa structure cognitive. Parce que nous nous intéressons ici uniquement à cette dernière la formalisation de la responsabilité n'est en soi pas central ici.

6. Dans [19, 2] et (avec un langage différent) dans [11] la responsabilité est définie par $\varphi \wedge \langle i\rangle\varphi$, ce qui ne nous semble pas intuitif car φ devrait être vrai *après* l'accomplissement de l'action et non pendant.

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. Par définition, il s'ensuit que $Ideal_{\{i\}}\,\varphi = Ideal_i\,\varphi$.

Comme la notion de responsabilité, celle de groupe fait l'objet de nombreuses études en épistémologie. C'est une notion complexe dont la formalisation entraîne nécessairement la formalisation de sa structure (ou au contraire de son absence dans le cas de groupes informels), de ses propriétés, *etc*. (voir par exemple [18] pour plus de détails) et celles-ci n'entrent pas nécessairement en ligne de compte dans la définition des émotions. Seul importe le fait qu'on considère les idéaux d'un groupe, quelle que soit la nature de ce dernier.

Enfin, de manière similaire (pour tout ensemble d'agents $C \in 2^{AGT}$) :

$$SIdeal_C\,\varphi \stackrel{déf}{=} \bigwedge_{i \in C} SIdeal_i$$

se lit : « φ est un idéal fort du groupe C ».

5.4 Formalisation de la culpabilité

Dans [19] un cadre formel complet est introduit pour la responsabilité et le regret. Dans [2], cette logique est présentée de manière didactique et d'autres émotions structurellement proches sont définies dans [11]. Par ailleurs, des travaux comme par exemple ceux d'Ortony *et al*. [21] ont déjà été formalisés en logique (voir notamment [1, 23]). 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.

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 :

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

Il est important de noter que comme $Resp_i\,\varphi \rightarrow \varphi$ alors $Guilt_i\,\varphi \rightarrow Bel_i\,\varphi$ (la culpabilité implique nécessairement qu'on croit que φ est vrai –même si ce n'est pas réellement le cas). Cette définition, modulo notre définition de la responsabilité qui est légèrement différente, correspond à la notion définie par ailleurs dans [2]. Il est intéressant de noter qu'à l'inverse de la honte, on peut culpabiliser même si nous n'étions pas conscient que nous pouvions empêcher ce qui est arrivé (*i.e*. il n'est pas nécessaire que $X^{-1}Bel_i\,\langle i\rangle X\varphi$ soit vrai, à comparer avec la définition de $Resp_i\,\varphi$).

Comme nous l'avons souligné précédemment, même si c'est légèrement plus fréquent pour la honte, il arrive qu'on éprouve un sentiment de culpabilité par rapport à un groupe d'agents (non nécessairement présent ni même au courant de la violation de l'idéal). Nous pouvons donc généraliser notre définition de la manière suivante (pour tout agent $i \in C$) :

$$Guilt_i\,(\varphi, C) \stackrel{déf}{=} Bel_i\,(Ideal_C\,\neg\varphi \wedge Resp_i\,\varphi)$$

5.5 Formalisation de la honte

Là encore, la honte peut être vis-à-vis de soi-même ou d'un groupe auquel on s'identifie et dont on a internalisé les idéaux (voir l'opérateur $Ideal_C$ ci-dessus).

Nous avons également montré qu'un agent qui éprouve de la honte a tendance à nier sa responsabilité (qui peut être réelle ou non). En d'autres termes, il pense qu'au moment où la situation honteuse s'est produite, elle était inévitable.

Nous obtenons la définition suivante ($i \in C$) :

$$Shame_i\,(\varphi, C) \stackrel{déf}{=} Bel_i\,(SIdeal_C\,\neg\varphi \wedge Inevitable\varphi)$$

Ainsi, l'agent i a honte du fait que φ soit vrai vis-à-vis du groupe C (qui peut se réduire à lui-même) si et seulement si il croit que :

1. φ devrait idéalement être faux pour le groupe C et que la nature de cet idéal est telle que sa violation peut lui faire perdre la face ;
2. il était inévitable au moment où la violation de l'idéal était sur le point de se produire que φ devienne vrai ;
3. φ est actuellement vrai (puisque $Inevitable\varphi \rightarrow \varphi$).

5.6 Propriétés intéressantes

Dans le reste de cette section, nous présentons quelques propriétés intéressantes dont la démonstration peut être trouvée dans [16]. Une première propriété concerne la relation entre ce

qui est nécessairement vrai (donc, indépendant de ce que font les agents) et les actions des agents. Ainsi, selon [12], on peut montrer que pour tout groupe d'agents C donné

$$\Box\varphi \rightarrow [C]\varphi \quad (1)$$

qui se lit : « si φ est inévitable, alors n'importe quelle coalition C fait en sorte que φ soit vrai, et ce quelles que soient les actions des agents extérieurs à cette coalition ».

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

$$\langle i\rangle\varphi \rightarrow \neg\Box\varphi$$

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

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

$$Resp_i\,\varphi \rightarrow \neg Inevitable\varphi \quad \text{(a)}$$
$$Shame_i\,(\varphi, C) \rightarrow Bel_i\,\neg Resp_i\,\varphi \quad \text{(b)}$$

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.)

Enfin, le théorème (2b) 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 responsable du fait que φ soit maintenant vrai. 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 croit qu'il n'est pas le cas que l'instant d'avant elle pouvait faire en sorte d'empêcher que φ soit vrai l'instant suivant (c'est-à-dire maintenant).

Émotions miroir. Il n'est pas possible d'aborder en profondeur cet aspect pour des raisons de place, mais il intéressant de souligner que, en tant qu'émotion sociale, il existe des *émotions miroir* qui répondent en quelque sorte à celle (honte ou culpabilité dans notre cas) éprouvée par l'agent responsable de la situation. Nous illustrons cette notion par l'exemple suivant. Supposons que $Ideal_C\,\varphi$ est vrai et que tout agent $j \in C$ croit que φ est faux (*i.e.* $\bigwedge_{j\in C} Bel_j\,\neg\varphi$). Supposons en outre que tous les agents du groupe C croient que la responsabilité du fait que φ soit faux est à imputer à un agent i quelconque. Cette incongruence entre attitude morale et attitudes épistémiques correspond traditionnellement au reproche ou à la désapprobation morale à propos de $\neg\varphi$ (voir [11, p. 6] par exemple). Supposons maintenant que $i \in C$. Il en découle que i éprouve de la culpabilité à propos de $\neg\varphi$ (qui peut être vue comme de la réprobation morale envers lui-même). Autrement dit, ces deux émotions (désapprobation et culpabilité) se différencient par le fait que nous sommes d'un côté du miroir ou de l'autre (on est responsable, ou c'est quelqu'un d'autre qui l'est), mais que dans tous les cas on regarde la même chose (la valeur de φ par rapport à ses idéaux).

6 Conclusion

Nous avons montré et formalisé les différences essentielles entre la honte et la culpabilité. Certaines notions (comme la responsabilité ou l'idéal d'un groupe) ont été formalisées de manière volontairement simple car c'est le concept qui est important, plus que sa forme particulière dans la situation considérée.

Nous avons rappelé combien ces émotions étaient importantes au niveau social : exprimer du regret lorsqu'on a fait quelque chose qui a violé une norme, ou faire attention à ne pas mettre quelqu'un dans une situation où il éprouverait de la honte sont des comportements qui trouveront, selon nous, une place naturelle au sein des agents conversationnels en ayant un rôle central sur leurs décisions en situation d'action.

Cette étude préliminaire mérite bien sûr d'être affinée, notamment en prenant en compte les tendances à l'action (ce qu'un individu est tenté de faire lorsqu'il éprouve une telle émotion). Cela permettrait ainsi de prendre en compte une autre composante (au sens de Sherer, cf. Section 2) de l'émotion. Il convient également de caractériser plus finement les différents idéaux mis en jeu.

Remerciements

Ce travail a été soutenu par le contrat de recherche CECIL (Complex Emotions in Communication, Interaction and Language) No. 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] Carole Adam. *Emotions : from psychological theories to logical formalization and implementation in a BDI agent*. PhD thesis, INP Toulouse, France, July 2007.

[2] Carole Adam, Benoit Gaudou, Dominique Longin, and Emiliano Lorini. Logical modeling of emotions for Ambient Intelligence. In Fulvio Mastrogio and Nak-Young Chong, editors, *Handbook of Research on Ambient Intelligence and Smart Environments : Trends and Perspectives*. IGI Global, 2011.

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

[4] R. Benedict. *The chrysanthemum and the sword*. Mariner Books, 1946.

[5] H. N. Castaneda. *Thinking and Doing*. D. Reidel, Dordrecht, 1975.

[6] Charles R. Darwin. *The expression of emotions in man and animals*. Murray, London, 1872.

[7] D. Davidson. *Essay on Actions and Events*. Oxford University Press, Oxford, 2nd edition, 2001.

[8] Ramon Martinez de Pison. *Death by Despair : Shame And Suicide*. Peter Lang Pub Inc, 2006.

[9] Jon Elster. *Alchemies of the Mind : Rationality and the Emotions*. Cambridge University Press, Cambridge, 1999.

[10] N. H. Frijda. *The Emotions*. Cambridge University Press, 1986.

[11] Nadine Guiraud, Dominique Longin, Emiliano Lorini, Sylvie Pesty, and Jérémy Rivière. The face of emotions : a logical formalization of expressive speech acts (regular paper). In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1031–1038. ACM, 2011.

[12] A. Herzig and E. Lorini. A dynamic logic of agency I : STIT, capabilities, and powers. *Journal of Logic, Language, and Information*, 19 :89–121, 2009.

[13] W. James. What is an emotion ? *Mind*, 9 :188–205, 1884.

[14] Richard S. Lazarus. *Emotion and Adaptation*. Oxford University Press, 1991.

[15] Helen Block Lewis. *Shame and guilt in neurosis*. International Universities Press, New-York, 1971.

[16] Dominique Longin. La honte versus la culpabilité : analyse conceptuelle et formelle en logique modale. Rapport de recherche IRIT/RR–2012-12–FR, IRIT, Université Paul Sabatier, Toulouse, mai 2012. 27 pages.

[17] Emiliano Lorini. A Dynamic Logic of Knowledge, Graded Beliefs and Graded Goals and Its Application to Emotion Modelling. In H. van Ditmarsch, J. Lang, and S. Ju, editors, *Proceedings of the LORI-III*, volume 6953 of *LNAI*, pages 165–178. Springer-Verlag, 2011.

[18] Emiliano Lorini, Dominique Longin, Benoit Gaudou, and Andreas Herzig. The logic of acceptance : grounding institutions on agents' attitudes. *Journal of Logic and Computation*, 19(6) :901–940, 2009.

[19] Emiliano Lorini and François Schwarzentruber. A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3-4) :814–847, 2011.

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

[21] Andrew Ortony, G.L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, 1988.

[22] D. Sander and K. Scherer, editors. *Traité de psychologie des émotions*. Cognitive. Dunod, 2009.

[23] B.R. Steunebrink, M. Dastani, and J.-J. Meyer. The OCC model revisited. In D. Reichardt, editor, *Proc. of the 4th Workshop on Emotion and Computing*, 2009.

[24] June Price Tangney. The self-conscious emotions : shame, guilt, embarrassment and pride. In *Handbook of Cognition and Emotion*. John Wiley & Sons, 1999.

[25] J. P. Tangney and R. L. Dearin. *Shame and Guilt*. The Guilford Press, 2002.

[26] J. P. Tangney, R. S. Miller, L. Flicker, and D. H. Barlow. Are shame, guilt, and embarrassment distinct emotions ? *Journal of Personality and Social Psychology*, 70(6) :1256–1269, 1996.

[27] Gabrielle Taylor. *Pride, Shame, and Guilt : Emotions of Self-Assessment*. Oxford University Press, New-York, 1985.