L'émotion dans la cognition : vers une formalisation logique

Dominique Longin

Version antérieure à la version publiée

1 Introduction

Le présent article n'a pas pour vocation de porter sur les relations entre art et émotion. Non qu'un tel travail manquerait d'intérêt, mais parce que la tâche semble encore bien trop complexe pour être menée à bien, du moins en l'état actuel de nos connaissances. En revanche, comme le souligne Barwell dans [2, p. 175], on ne peut comprendre ce que cela signifie que de qualifier une peinture de triste ou de reconnaître de la tristesse dans une peinture, si on ne sait pas auparavant ce que cela signifie de dire qu'une personne est triste et si on ne sait pas comment reconnaître cette tristesse. En ce sens, le présent travaille apporte une contribution à la compréhension des liens entre art et émotion : il se situe dans le cercle très restreint des travaux formels sur l'émotion, où l'on cherche aussi bien à caractériser les émotions elles-mêmes que le rôle qu'elles jouent dans le processus de raisonnement et de décision d'un individu (l'espoir étant d'appliquer ensuite ces résultats à des agents artificiels dits « intelligents »). Il s'agit de montrer qu'on est d'ores et déjà capable de capturer certains aspects des émotions. Par bien des côtés, le présent ouvrage présente également de manière plus large les travaux sur l'art, et il est permis d'avoir quelque espoir que les deux domaines soient suffisamment mûrs dans un futur très proche pour faire émerger une théorie des interactions entre l'un et l'autre. En attendant, parallèlement à cela, des recherches portant sur l'émotion se poursuivent dans le domaine de l'informatique.

Ces dernières années, il est aisé de constater que différentes technologies (téléphones portables, bornes interactives, télé-services *via* Internet, *etc*.) deviennent de plus en plus envahissantes (et tout laisse à penser que ce phénomène va encore s'accélérer) au point que ce constat est aujourd'hui banal. Devant l'importance du

phénomène, les services augmentent en même temps que leur propre complexité : rendre de plus en plus de services, traiter des demandes de plus en plus fines, fournir une réponse de plus en plus proche de ce que l'utilisateur attend... le tout en facilitant l'utilisation de ces systèmes et l'accès aux services qu'ils offrent. L'enjeu ultime commun à tous ces systèmes est donc aujourd'hui d'offrir aux utilisateurs des services de plus en plus importants (aussi bien quantitativement que qualitativement), et de les leur offrir de la manière la plus intuitive possible. Cela implique que ces systèmes soient crédibles aux yeux des utilisateurs et disposent dans ce but de capacités communicatives évoluées, ce qui passe immanquablement par une capacité à comprendre, raisonner sur et prédire, les émotions des utilisateurs. Certains poussent cette crédibilité jusqu'à exiger d'elle que les agents ne soient pas seulement honnêtes et fiables, mais fournissent l'illusion de la vie [3].

Nombreux sont les travaux en psychologie ayant montré le rôle central de l'émotion dans la cognition et l'interaction sociale [17] et son rôle dans chaque étape du processus de raisonnement. Du point de vue de l'informatique, le défi semble grand, et l'exemple de la chambre chinoise élaboré par Searle n'est point pour rassurer.[1] Mais pour reprendre les mots d'Ortony *et al*. eux-même, « *le but (...) n'est pas de créer des machines dotées d'émotions, mais de créer des modèles informatiques pouvant comprendre quelles émotions les gens peuvent éprouver, et sous quelles conditions. De tels systèmes seraient ainsi capables de prédire et d'expliquer les émotions humaines, pas de les ressentir.* » [16, p. 17]

Certes, le rêve de disposer de machines capables de ressentir des émotions n'est pas nouveau et date (au moins) des prémisses de l'informatique et de son désir de rendre les programmes intelligents. Du point de vue historique, ce sont John McCarthy, Marvin Minsky, Claude Shannon, Alan Newell et Herbert Simon qui, lors de la conférence de Dartmouth en 1956, se sont fixés un tel but, marquant ainsi la naissance de l'intelligence artificielle.

Après avoir brièvement présenté les fondements psychologiques de notre approche (Section 2) ainsi que le langage de formalisation des émotions que nous utiliserons (Section 3), nous modélisons quelques émotions et exhibons quelques propriétés les concernant (Section 4).

1. Dans [25, p. 42], l'auteur illustre le fonctionnement interne d'un ordinateur qui parlerait le Chinois, en montrant que ce fonctionnement ne requiert ni n'induit pour l'ordinateur aucune capacité à « comprendre » les phrases qu'il « prononce ».

2 Qu'est-ce qu'une émotion ?

Comme le rapporte Solomon [26] « *What is an emotion?* » est une question qui fut précisément posée en ces termes par William James comme titre d'un article qu'il écrivit pour *Mind* en 1884. (En fait, la philosophie s'intéresse à l'émotion — et par voie de conséquence, se pose cette question — depuis l'antiquité.) Selon le Robert électronique, l'émotion est définie comme un « état de conscience complexe, généralement brusque et momentané, accompagné de troubles physiologiques (pâleur ou rougissement, accélération du pouls, palpitations, sensation de malaise, tremblements, incapacité de bouger ou agitation). » Du point de vue psychologique, les définitions qui suivent sont relativement courantes (et de ce fait nécessairement vagues), mais elles permettent de donner une idée intuitive des concepts que nous manipulerons par la suite, et que nous préciserons au fur et à mesure :

- « émotion » une expérience forte de durée relativement courte, dont la cause est généralement identifiable et possédant un contenu cognitif (colère, peur, joie, *etc.*) ;
- l'« humeur » (*mood*) réfère, par opposition à l'émotion, à une expérience d'intensité plus basse, plus diffuse mais plus persistante, dont la cause n'est pas réellement identifiée (ni identifiable, dans certains cas) ni saillante ;
- l'« affect » (*affect*) est généralement un terme générique employé pour désigner indifféremment une émotion ou une humeur ;
- enfin, les « sentiments intérieurs » ou « sentiments » (*inner feelings*) désignent souvent l'expérience subjective d'états affectifs.

Un peu comme la question de savoir quelle est la relation entre l'esprit (la pensée) et le cerveau, la question de la relation entre l'émotion et la cognition s'est immédiatement posée. Solomon [26] rappelle à ce propos que l'une des plus anciennes métaphores associées à la raison et à l'émotion est celle du maître et de l'esclave : alors que la raison est reliée à l'idée de contrôle, l'émotion est reliée à celle d'impulsion dangereuse que l'on peut soit ignorer, soit canaliser, soit (idéalement) accorder à la raison. Socrate s'intéressait déjà à l'émotion et après lui son élève Platon établit une distinction claire entre raison (ce que l'on appelle désormais la cognition), passion (*i.e.* les émotions) et désir (appelé *motivation* de nos jours). Dans [18] ce dernier avance que les sentiments sont l'ennemi de la raison et que les citoyens devraient bannir les émotions de leurs décisions « cognitives ». Par la suite, Aristote défendit l'idée selon laquelle les émotions dépendent de la raison qui peut ainsi les contrôler [1] : la peur, par exemple, découle de pensées

particulières. Autrement dit, les émotions sont post-cognitives, elles interviennent après un certain raisonnement (conscient ou non).

Comme le souligne Lazarus [14], le fait que les anciens Grecs aient séparé, voire opposé, émotion, cognition et motivation, non seulement implique l'existence indépendante de ces trois concepts dans notre esprit, mais également a forcé la philosophie moderne à spécifier les relations fonctionnelles entre eux.

Quoi qu'il en soit, jusqu'à la fin des années 70, la psychologie contemporaine adoptait ce point de vue post-cognitiviste des émotions jusqu'à ce que dans [29], Zajonc cite un poème de Cummings [8, p. 160] prenant comme établi que les sentiments sont primitifs, fondamentaux, et note non sans humour que bien peu de chercheurs seraient affligés d'apprendre que leurs théories sont en conflit avec un poète controversé des années 20, eux pour qui l'affect est post-cognitif. (« *Affect is postcognitive* » [29, p. 151].) Mais Zajonc ajoute plus sérieusement que ces théories rentrent également en conflit avec celle d'un des pères fondateurs de la psychologie contemporaine, Wundt, qui écrivait dès 1905 que la claire aperception des idées dans des actes de connaissance et de reconnaissance est toujours précédée par des sentiments. (« *(...) the clear apperception of ideas in acts of cognition and recognition is always preceded by feelings.* » [28, p. 243–244].)

Cette thèse énoncée la première fois dans [29] et élaborée par la suite [30, 31] donna naissance à un débat qui dura 20 ans et dont Lazarus [12, 13, 15, 14] fut l'autre principal protagoniste (mais pas le seul). Le point de vue de Zajonc est exactement à l'opposé de celui qui consiste à dire que l'émotion n'est qu'une perturbation de la cognition : selon lui, l'émotion peut évoluer indépendamment de la cognition, ou même la précéder. Sa principale thèse est que l'évaluation affective d'un *stimulus* peut se produire immédiatement après sa perception et avant tout traitement cognitif. Cette évaluation est donc prioritaire et effectuée en terme de valence (*i.e.* on évalue le caractère positif ou négatif du *stimulus*). Il s'agit d'un processus bas niveau conférant aux émotions un caractère généralement inconscient et constituant une réaction immédiate au stimulus. L'émotion est donc une préférence binaire de type « on aime/on n'aime pas ». Pour Lazarus, c'est le contexte ainsi que la disposition dans laquelle se trouve un individu qui déterminent l'évaluation de la situation, évaluation nécessaire à l'existence des émotions. Pour Lazarus, une émotion se produit donc toujours après une évaluation cognitive.

Certes, il paraît difficilement contestable que le processus (inconscient) de préférence de Zajonc peut être mené à bien plus rapidement qu'un processus cognitif conscient, et que celui-là est très différent de celui-ci. Néanmoins, la définition que Zajonc prend de la cognition quand il l'oppose à l'affect est très contestable : il

oppose émotion et cognition au même titre que processus conscient et inconscient (ceci est particulièrement clair dans [31]). Or il apparaît de façon quasi unanime que la cognition, en tant que processus calculatoire, peut inclure des calculs de bas niveau dans le subconscient ce qui affaiblit considérablement la position de Zajonc. Dans le même temps les expérimentations menées semblent corréler la thèse de Lazarus. (Voir [21] pour plus de détails par exemple.)

Plus récemment, Lazarus estima pour sa part que ces trois concepts (émotion, cognition, et motivation) sont « plus ou moins des fictions de l'analyse scientifique, dont l'indépendance n'existe pas réellement dans la nature » [14]. En tout état de cause, l'émotion ne peut se manifester sans la présence *simultanée* de processus cognitifs et motivationnels. L'émotion est une réponse à une certaine croyance, et cette croyance est nécessairement en relation avec nos buts pour qu'elle nous soit pertinente et provoque en nous une émotion. Mais bien qu'une pensée puisse apparaître sans émotion associée, l'inverse n'est pas vrai : une émotion ne peut se manifester complètement détachée de toute pensée (elle a nécessairement un contenu). Cela ne signifie pas pour autant que l'émotion précède la cognition (contrairement à ce que peut laisser supposer le titre de [13]). Il faut considérer que les émotions et les pensées se produisent de manière permanente, et que l'une peut tout à fait être à l'origine de l'autre. Se demander si l'émotion précède la cognition ou si c'est l'inverse, c'est un peu essayer de répondre à la question similaire pour l'œuf et la poule !

En ce qui nous concerne, notre vision est résolument guidée par cette dernière remarque de Lazarus et nous concevons les émotions *dans* la cognition, ce qui n'exclut pas que les émotions puissent être à l'origine de pensées ou de comportement. (Voire par exemple [24, 9] où des expériences montrent comment certaines émotions sont directement inclues en tant qu'informations pertinentes dans le processus de jugement — et donc, l'influencent.)

En somme, que ce soit en philosophie ou en psychologie, l'idée selon laquelle émotion et cognition sont intimement liées n'est pas nouvelle. Mais paradoxalement, sitôt énoncée cette idée semble avoir été sitôt oubliée, et une distinction claire (voir une opposition) entre ces deux notions a longtemps constitué le point de vue largement prédominant. Le présent article se veut être au contraire une contribution où émotions et attitudes mentales (les *composants cognitifs* à l'origine de nos raisonnements, tels la croyance, les désirs, *etc*.) s'interpénètrent, où les unes ne sauraient exister sans les autres.

La question des émotions basiques *vs* complexes. Comme nous venons de le montrer, plusieurs auteurs, en un sens ou un autre, retiennent l'idée de l'existence d'émotions de base. À l'instar d'Ortony *et al*. [16, p. 25–32] nous trouvons cette notion « inacceptablement vague » (« *unacceptably vague* » [16, p. 26]) : il n'est pas évident si elles sont basiques parce qu'à l'origine de toutes les autres, si c'est parce qu'elles apparaissent dans toutes les cultures, si parce qu'apparemment certaines émotions peuvent être aussi ressenties par certains animaux, ou si c'est parce qu'elles sont à la base de comportements assurant une part de notre survie.

Cette distinction est, selon eux, fondée sur une illusion (dont ils détaillent les origines dans [16, Table 2.1 page 27]), en dépit du fait qu'un certain nombre d'auteurs soutiennent cette hypothèse.

Ortony *et al*. [16]. La théorie des émotions d'Ortony *et al*. est fondée sur trois concepts structurants : les types d'émotion, les groupes d'émotions, et les classes d'émotions. Un « type d'émotion » est un genre d'émotion qui peut se réaliser selon une variété de formes reconnaissables. Il est possible de classer des types d'émotion, pas les (instances de types d')émotions elles-mêmes car il y en a trop et chacune est différente. En ce sens, une émotion est vue comme une instance de type d'émotion ayant un certain degré d'intensité et pouvant avoir différents accents d'intensité pour un même type. Ortony *et al*. citent à titre d'exemple le type d'émotion de crainte (*fear*) qui peut se manifester sous différents degrés d'intensité (inquiétude, frayeur, *etc*.) ou différents accents tels l'angoisse. En définitive, pour Ortony *et al*. un type d'émotion est une famille d'émotions de natures relativement proches.

Il est important de noter que selon Ortony *et al*., les noms d'émotions fournis en FIG. 1 ne constituent que des étiquettes linguistiques pour nommer les types d'émotion. À ce titre, un étiquette n'est qu'un terme censé rendre compte le mieux possible de l'ensemble des émotions du type qu'elle décrit, sans pour autant qu'on puisse l'élever au rang d'émotion paradigmatique de ce type.

Les types d'émotion (et par voie de conséquence, les émotions se rapportant à chacun de ces types) sont structurés au sein de petits ensembles cohérents partageant des propriétés communes : les « groupes d'émotions ». Ces propriétés correspondent à des conditions d'obtention (*eliciting conditions*). Par exemple, le *well-being group* (*cf*. FIG. 1) est un groupe d'émotions contenant les deux types d'émotion : *joy* et *distress*. Son *eliciting condition* est le plaisir/déplaisir immédiat pour soi associé à l'occurrence d'un événement.

Enfin, les groupes d'émotions sont structurés en « classes d'émotions » sur

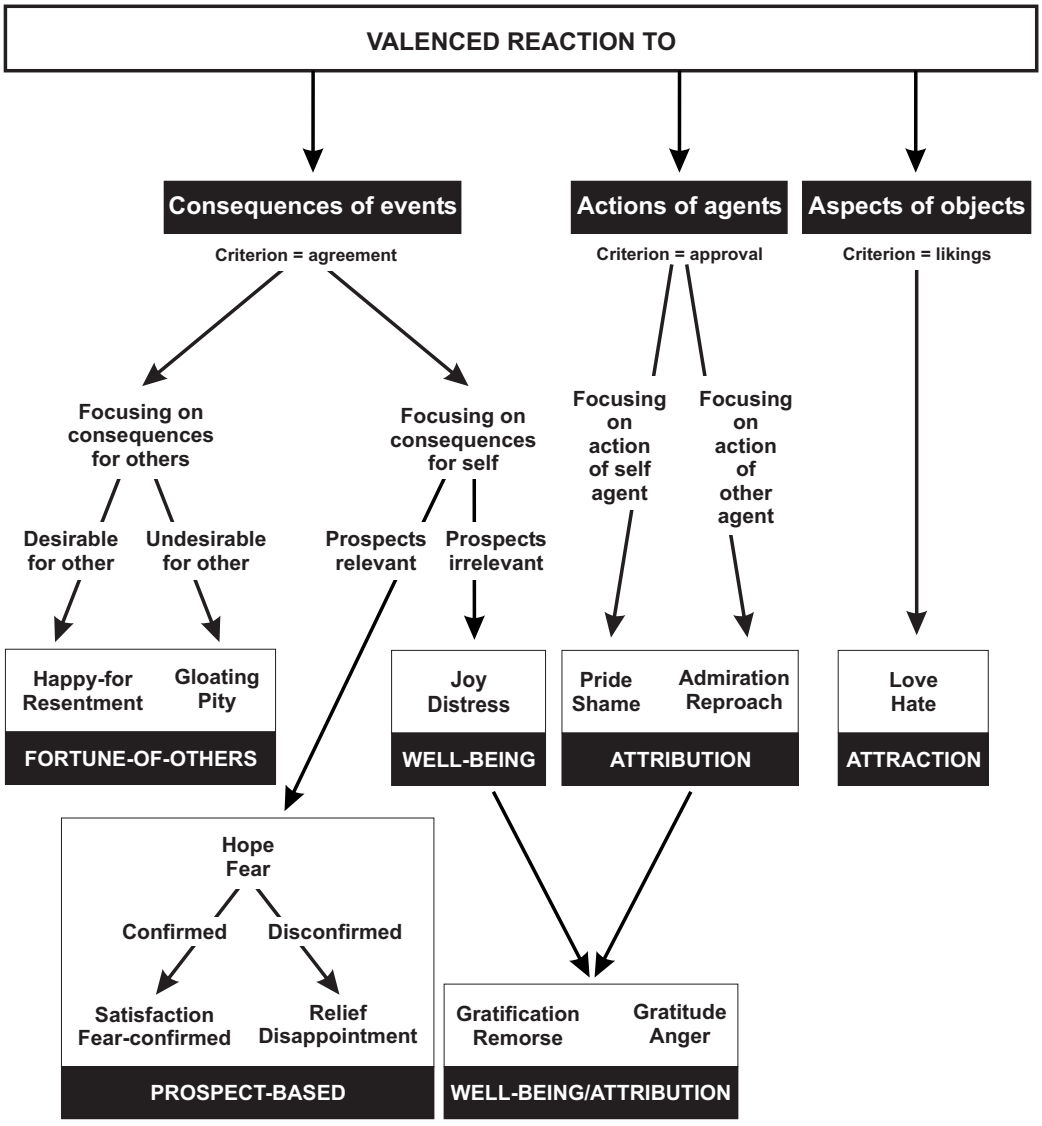

FIG. 1 – *Structure globale des types d'émotion chez Ortony* et al. *[16, p. 19]*

la base du principe selon lequel les individus perçoivent le monde au travers (FIG. 1) :

– des événements qui s'y produisent (classe des *consequences of events*) ;
– des actions que les agents y effectuent (classe des *actions of agents*) ;
– des objets qui le composent (classe des *aspects of objects*).

La notion centrale dans la théorie des émotions d'Ortony *et al*. est qu'une émotion est une réaction valencée (*i.e*. positive ou négative) à l'un de ces trois types de changement dans le monde. Cette perception induit ainsi trois classes d'émotions. (Dans la suite, nous n'étudierons que la première.) Il est utile de noter que la notion d'agent est très large, et inclut les institutions, les animaux, ou même des objets inanimés que l'on considère comme des agents. C'est le cas par exemple d'une voiture qui, en principe, est un objet, mais à qui on peut faire des reproches suite à des pannes répétées (personnification).

Les types d'émotion sont structurés à l'intérieur des classes et groupes d'émotions par des *variables d'évaluation*. Chaque groupe d'émotions est soumis à des « variables globales » (dont dépendent tous les types d'émotion) et à des « variables locales » (dont seules les émotions du groupe considéré dépendent). Les variables communes à toute une classe d'émotions sont appelées « variables (d'évaluation) centrales » (ou encore « variables principales »).

Ainsi, la classe *consequences of events* décrit les réactions d'un agent par rapport au plaisir ou au déplaisir (*pleased*/*displeased*) qu'il éprouve face aux conséquences d'événements qu'il juge désirables ou indésirables. Ce jugement se fait en évaluant la variable centrale de « désirabilité » (*desirability*). La désirabilité d'un événement se calcule par rapport au nombre de buts (sous-buts...) atteints ou au contraire bloqués suite à l'occurrence de cet événement. (En considérant toutes choses égales par ailleurs, la désirabilité — respectivement, l'indésirabilité — est évaluée basiquement en comptant le nombre de buts atteints — respect. contrariés.) Un événement peut ainsi être à la fois désirable et indésirable, mais chacune des conséquences de cet événement ne peut qu'être soit plaisante, soit déplaisante (mais pas les deux). Par exemple, on est content pour une personne lorsqu'on éprouve du plaisir suite à l'occurrence d'un événement désirable pour cette personne.

La classe *actions of agents* décrit les réactions d'un agent par rapport au fait que cet agent approuve ou désapprouve (*approving*/*disapproving*) les conséquences d'actions (méritantes ou non) accomplies par d'autres agents. Ce mérite est établi *via* l'évaluation de la variable centrale de « mérite » (*praiseworthiness*) que l'on calcule en évaluant les actions des agents par rapport à des normes, des

standards, que l'on a internalisés. Par exemple, l'admiration que l'on peut éprouver envers quelqu'un est fonction du mérite qu'il a eu à faire une certaine action, mérite calculé en fonction des standards.

Enfin, la classe *aspects of objects* décrit les réactions d'un agent par rapport à ses goûts (*liking*/*disliking*) à propos d'objets qui l'attirent ou non. Cette attirance est mesurée par la variable centrale d'attirance (*appealingness*). Par exemple, l'amour et la haine sont évalués selon ces critères. (Comme nous l'avons souligné auparavant, les « objets » considérés ici peuvent être des objets à proprement parler, mais également des animaux ou même des individus : ce n'est que lorsqu'on les *considère* comme des objets que l'on peut éprouver ces types d'émotion — *sic* Ortony *et al*.)

Comme nous l'avons dit précédemment, Ortony *et al*. rejettent le concept d'émotion complexe. En revanche, ils stipulent que s'il faut accepter l'idée d'émotions basiques (et, par voie de conséquence, complexes) c'est au niveau du nombre de variables qu'elles font intervenir, des différentes dimensions qu'elles revêtent. Certaines émotions sont plus simples à spécifier que d'autres. Par exemple, FIG. 1 montre que la colère (*Anger*) est plus complexe que le reproche (*Reproach*), les conditions d'obtention de la première incluant celles de la seconde (mais pas seulement).

Les auteurs vont donc dans le sens d'une classification hiérarchique de la complexité des émotions où, tout en haut, se trouveraient les deux types de réaction affective « positive » et « négative ». C'est en ce sens là seulement que les auteurs acceptent de parler d'émotions de base, ce qui rejoint le point de vue de Spinoza [27] et avant lui celui d'Aristote [1] qui réduit toutes les émotions à une forme ou à une autre de plaisir ou de douleur.

Un corollaire à cela est que toute émotion doit être valencée : il n'y a pas d'émotion neutre ! Cela induit notamment que la surprise, par exemple, qui peut être bonne, mauvaise, ou ni l'un ni l'autre, n'est pas une émotion. C'est un état qui *peut* induire un état émotionnel, mais pas nécessairement. De même que *se sentir abandonné* : on peut se trouver dans un tel état mais ne pas se laisser abattre, voire trouver cela positif pour une raison ou une autre. En définitive, de tels états sont des *états cognitifs* : puisqu'ils peuvent être valencés, ils peuvent être affectifs, mais comme ils ne le sont pas nécessairement, ce ne sont pas des émotions.

Peut-on éprouver des émotions contradictoires ? Ce que nous entendons ici par « émotions contradictoires » ce sont des émotions portant sur le même événement, une même action ou un même objet (selon le cas) mais ayant des valences

opposées. D'une manière générale, les psychologues acceptent le fait que l'on peut éprouver des émotions contradictoires.

Pour nuancer quelque peu cette position générale, nous disons que cela dépend de la granularité avec laquelle on considère la cause de l'émotion. Par exemple, le décès d'un proche peut nous rendre très triste et peut également nous faire ressentir un certain soulagement (parce que la mort a abrégé d'intenses souffrances, par exemple). Néanmoins, si on se focalise sur les causes particulières de chacune de ces émotions, on peut considérer que l'on est triste que la personne soit morte et on est soulagé de voir ses souffrances abrégées (ce qui constitue plutôt une conséquence de la mort). En ce sens, ces émotions ne sont pas contradictoires puisqu'elles correspondent à des buts (bloqués) différents : celui de voir son proche continuer à vivre, et celui de ne pas le voir souffrir.

Terminologie. Selon les différents travaux, le nom associé à l'ensemble des propriétés identifiant une émotions (ou un type d'émotion) particulière (les *eliciting conditions* d'Ortony *et al*.) peut varier.

Par exemple, l'espoir chez Ortony *et al*. correspond à l'attente d'un événement désirable très probable (dont la probabilité est supérieure à $0{,}5$), alors que chez Lazarus cela correspond plutôt à l'attente d'un événement peu probable. De même, dans l'exemple du paragraphe précédent, le soulagement est décrit comme la satisfaction (*confirmed*) d'un événement désirable, alors que chez Ortony *et al*. il correspond au blocage (*disconfirmed*) d'un événement indésirable.

En ce qui nous concerne, la formalisation logique conduit naturellement à une désambiguïsation des définitions des émotions considérées (les *eliciting conditions* d'Ortony *et al*.) : le nom affecté à telle ou telle définition n'est alors plus qu'une question terminologique et n'a que peu d'influence pour les questions qui nous occupent puisque ce nom correspond par définition à une certaine formule dont la sémantique est clairement identifiée.

3 Langage formel

Le langage utilisé une extension de la logique classique avec des opérateurs modaux. Ce qui caractérise un opérateur modal $\Box$ (qui se lit « *box* ») par rapport à un opérateur de la logique classique, c'est qu'il n'est généralement pas vérifonctionnel, c.-à-d. que la formule $\Box\varphi$ peut être vraie indépendamment du fait que la formule φ soit vraie ou non. Typiquement, la croyance est une modalité puisqu'on peut croire qu'il fait beau, par exemple, indépendamment du fait qu'il fasse

réellement beau (on peut se tromper !).

On note p, q, r... les formules atomiques (*i.e.* les formules primitives, non construites à partir d'autres formules) ; on note φ, ψ... les formules complexes construites à partir de formules atomiques ou d'autres formules complexes selon les règles suivantes :

1. si p est une formule atomique, alors p est une formule complexe ;
2. si φ et ψ sont des formules complexes, alors $\neg\varphi$ (qui se lit : « φ est faux »), $\varphi \vee \psi$ (qui se lit : « φ ou ψ est vrai »), et $\Box\varphi$ ($\Box$ est un des opérateurs modaux définis par la suite et représentant soit une croyance, soit le temps, soit des préférences) sont des formules ;
3. toute formule s'obtient en appliquant un nombre fini de fois les règles 1 et 2 ci-dessus.

On définit les abréviations suivantes :

$$\varphi \wedge \psi \stackrel{déf}{=} \neg(\neg\varphi \vee \neg\psi) \qquad (\varphi \text{ et } \psi \text{ sont vrais})$$

$$\varphi \rightarrow \psi \stackrel{déf}{=} \neg\varphi \vee \psi \qquad (\text{si } \varphi \text{ est vrai alors } \psi \text{ est vrai})$$

Dans la suite, nous définissons les différents opérateurs que nous utiliserons pour formaliser des émotions. Ces opérateurs sont tous définis dans une logique modale spécifique dont nous ne préciserons que les aspects destinés à faciliter la compréhension de la formalisation des émotions. Nous n'énonçons donc pas toutes les propriétés logiques (axiomatique) ni même toutes les contraintes sémantiques correspondantes (sémantique). La sémantique correspondant à ces opérateurs est une sémantique de Kripke en termes de mondes possibles. (Voir ci-dessous pour plus de détails.)

La croyance. La croyance représente la vision que l'agent a du monde. En ce sens, la croyance représente le savoir subjectif de Kant, par opposition au savoir objectif couramment appelé « connaissance » : il ne s'agit donc pas ici de croyance au sens de quelque chose dont on ne serait pas sûr, mais au contraire au sens où ce que l'on croit représente ce qui, de notre point de vue, est vrai dans le monde. Nous notons $Bel_i\,\varphi$ le fait que l'agent i croit que la formule φ est vraie.

D'un point de vue sémantique, dire que $Bel_i\,\varphi$ est vrai signifie que φ est vrai dans tous les mondes épistémiques que l'agent envisage comme des alternatives possibles à la réalité, et ce sans que cet agent ne puisse distinguer lequel de ces mondes est le monde réel, et sans même avoir la certitude que le monde réel appartient à cet ensemble de mondes épistémiques. (Il peut se tromper.)

Ainsi, dire que l'agent i croit que φ est vrai dans le monde réel noté w_0 s'écrit : $w_0 \Vdash Bel_i\,\varphi$. Sémantiquement, cela signifie que φ est vrai dans tous les mondes accessibles depuis w_0 *via* la relation représentant la croyance de l'agent i et notée $\mathscr{B}_i$.

De façon duale, dire que l'agent *envisage* que φ est vrai dans le monde w_0 signifie qu'il existe au moins un monde épistémique accessible depuis le monde réel w_0 où φ est vrai. (On ne sait rien de la valeur de vérité de φ dans les autres mondes accessibles.) Autrement dit, cela signifie qu'il est faux que φ est faux dans tout monde accessible. « L'agent i envisage φ » se formalise donc par : $\neg Bel_i\,\neg\varphi$. D'une manière similaire, l'agent i envisage que φ est faux se formalise : $\neg Bel_i\,\neg(\neg\varphi)$ qui est logiquement équivalent à $\neg Bel_i\,\varphi$.

Si nous supposons que l'ensemble des mondes épistémiques (*i.e.* l'ensemble des mondes accessibles par la relation de croyance de l'agent i depuis le monde réel w_0) est limité à quatre mondes w_1 à w_4, alors nous représentons graphiquement le fait que dans le monde réel w_0 :

– l'agent i *croit* que p est vrai (*i.e.* $w_0 \Vdash Bel_i\,p$) ;

– l'agent i *envisage* le fait que q soit vrai (*i.e.* $w_0 \Vdash \neg Bel_i\,\neg q$) ;

– l'agent i *ne sait pas si* r est vrai ou non (*i.e.* $w_0 \Vdash \neg Bel_i\,r \wedge \neg Bel_i\,\neg r$).

par le schéma FIG. 2. L'ensemble des mondes accessibles depuis w_0 est noté $\mathscr{B}_i(w_0)$.

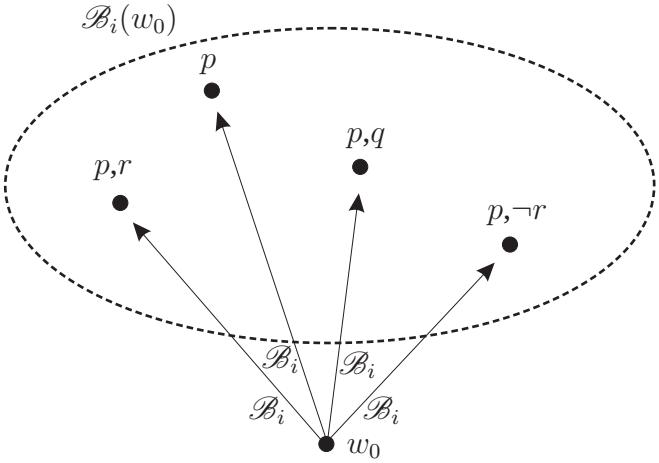

FIG. 2 – *Dans le monde w_0 l'agent i croit p, envisage q, et ne sait pas si r*

Du point de vue technique, les opérateurs de croyance sont définis dans une logique modale de type KD45 (selon la classification de [5]), qui est une logique

relativement standard (*cf.* [10] pour un article complet sur la formalisation de la connaissance et de la croyance en logique modale). Cela signifie qu'un agent est conscient de ce qu'il croit et de ce qu'il ne croit pas. De plus, la propriété suivante est valide :

$$Bel_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi \qquad (\mathrm{D}_{Bel_i})$$

ce qui signifie que les agents considérés sont rationnels : un agent ne peut croire simultanément une chose et son contraire. Nous verrons plus loin les implications de cette propriété.

Le temps. La notion de temps que nous utilisons est celle de temps linéaire. Cela signifie que les croyances de l'agent portent en fait non pas sur des mondes épistémiques mais sur des ensembles de mondes ordonnés linéairement dans le temps appelés des « histoires ».

Pour référer à tout instant dans le futur, nous utilisons l'opérateur G tel que $G\varphi$ signifie : « désormais et à chaque instant du futur de l'histoire considérée, φ est vrai ». De façon duale, pour référer à un instant dans le futur, nous utilisons l'opérateur F tel que $F\varphi \stackrel{déf}{=} \neg G\neg\varphi$ signifie : « pour une histoire donnée, il y a au moins un instant dans le futur où φ sera vrai dans cette histoire » (indépendamment du fait qu'il l'est actuellement ou non, ou l'a été ou non).

De même, nous disposons de l'opérateur H tel que $H\varphi$ signifie : « φ a toujours été et est encore vrai actuellement dans une histoire donnée ». (Cet opérateur est symétrique à G et parle du passé.) De façon duale, on définit $P\varphi \stackrel{déf}{=} \neg H\neg\varphi$ qui signifie : « il existe au moins un monde dans le passé sur l'histoire considérée où φ était vrai ».

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

Ainsi, FIG. 3 représente les quatre histoires crues par l'agent i. Les points situés sur les histoires indiquent le présent, et les tirets les instants avant ou après le présent. Dans cet exemple :

- il est toujours vrai que l'agent croit (dans le présent) que p est vrai ($Bel_i\,p$) ;
- il envisage également le fait que r soit actuellement vrai et devienne faux par la suite ($\neg Bel_i\,\neg(r \wedge F\neg r)$) ;

– il envisage aussi une histoire possible où désormais q est vrai ;
– il croit que s a été vrai à un instant dans le passé ($Bel_i\, Ps$) ;
– il envisage enfin que t sera vrai l'instant juste après.

Ces cinq formules sont représentées graphiquement en FIG. 3.

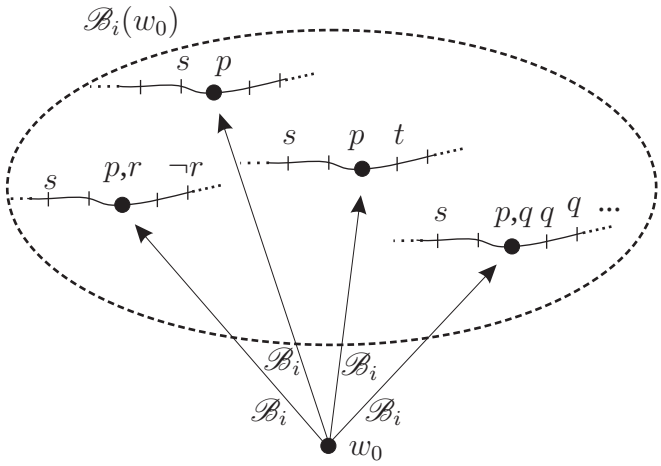

FIG. 3 – *Dans w_0 l'agent i croit p, envisage que désormais q est vrai, envisage que r soit vrai et devienne faux, croit que s a été vrai, et envisage que t soit vrai l'instant juste après*

Techniquement, le temps est défini dans une logique modale du temps linéaire de type S4.3$_t$ (*cf.* [4] pour plus de détails). On suppose donc que le futur est unique. Il existe un débat où certains défendent l'idée que le temps est ramifié dans le futur, mais comme ce qui nous intéresse avant tout c'est la perception subjective du temps (la façon dont un agent le perçoit), un temps linéraire convient : les croyances d'un agent portent sur des histoires (*i.e.* des successions de points temporels ordonnés linéairement les uns par rapport aux autres) et un agent peut envisager différentes histoires où le déroulement des événements est différent d'une histoire à l'autre. Une telle représentation du temps suffit donc à nos besoins puisqu'elle permet à l'agent d'envisager des futurs différents et donc de « simuler épistémiquement » un temps ramifié dans le futur.

Remarque 1 *Selon la présence ou non d'un opérateur de croyance dans la portée d'un opérateur du passé, on illustre des phénomènes différents :*

– $Bel_i\, X^{-1} Bel_i\, p$ *(mémoire)*
signifie que l'agent i croit actuellement que l'instant d'avant il croyait que

p était vrai (indépendamment de ce qu'il pense maintenant et de ce qui est ou était vrai alors).

– $Bel_i\, X^{-1} p$ *(point de vue actuel sur ce qui était vrai dans le passé) signifie que l'agent i croit actuellement que l'instant d'avant p était vrai (indépendamment de ce qu'il pensait alors ou pense maintenant) ;*

Ainsi, $Bel_i \neg p \wedge Bel_i\, X^{-1}(Bel_i\, p \wedge \neg p)$ signifie qu'actuellement i croit que p est faux, que p était déjà faux l'instant d'avant alors qu'à cet instant là il croyait que p était vrai.

La préférence. Nous aurons également besoin d'exprimer les préférences. Au sens où nous l'entendons, « l'agent i préfère φ » signifie que l'agent i préfère que φ soit actuellement vrai, indépendamment du fait que φ soit réellement vrai ou cru comme tel par l'agent i, et sans que cet agent ait nécessairement le but de le rendre vrai s'il ne l'est pas déjà. Ainsi, on peut préférer qu'il fasse soleil tout en sachant qu'il pleut et sans souhaiter agir (ce serait difficile !) pour qu'il fasse beau.

Comme il ne s'agit pas ici d'aborder l'intensité de ces préférences, nous nous contenterons d'un opérateur de préférence binaire (on préfère, ou on ne préfère pas), sans distinguer si on préfère beaucoup ou non. Cela aura une influence sur l'*intensité* des émotions modélisées mais pas sur la définition qu'on en donne. Comme nous ne nous préoccuperons pas de l'intensité des émotions, ce choix se trouve ici tout à fait justifié.

Formellement, nous introduisons à cet effet les opérateurs $Pref_i$ tels que $Pref_i\, \varphi$ signifie : « l'agent i préfère que φ soit actuellement vrai » (indépendamment du fait que φ soit actuellement vrai ou non, ou qu'il veuille le rendre vrai ou non).

Du point de vue de la sémantique, préférer φ signifie que φ est actuellement vrai dans toutes les histoires accessibles depuis le monde actuel w_0 par la relation de préférence $\mathscr{P}_i$. De façon similaire à la croyance, nous notons $\mathscr{P}_i(w_0)$ l'ensemble des histoires préférées par l'agent i dans le monde w_0.

La préférence se représente graphiquement de la même façon que la croyance. Une exemple de formules préférées est donné FIG. 4.

Techniquement, la préférence est définie dans une logique de type KD45 et possède donc des propriétés similaires à celles de la croyance (*cf.* les notions similaires de *goal* chez Cohen et Levesque [6, 7] ou Rao et Georgeff [19, 20], ou de choix chez Sadek [22, 23]). En particulier, les préférences sont elles aussi ration-

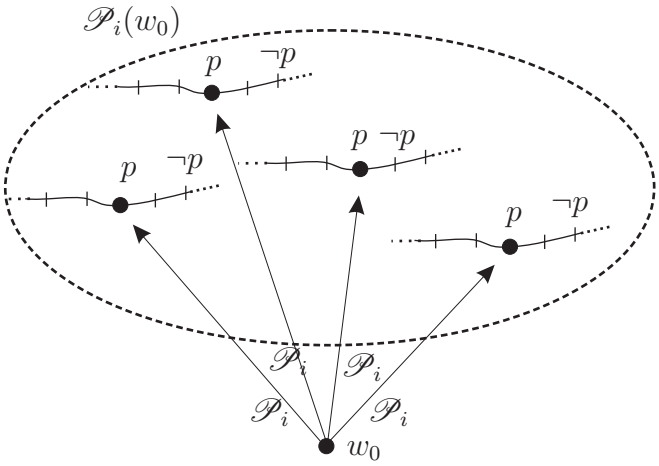

FIG. 4 – *Dans le monde w_0 l'agent i préfère que p soit actuellement vrai et devienne faux un jour :* $Pref_i\,(p \wedge F\neg p)$

nelles et obéissent à l'axiome suivant :

$$Pref_i\,\varphi \rightarrow \neg Pref_i\,\neg\varphi \qquad (\mathrm{D}_{Pref_i})$$

qui signifie qu'un agent ne peut préférer simultanément une chose et son contraire.

4 Formalisation d'émotions basées sur des événements selon Ortony *et al.*

Dans cette section, nous allons mettre à profit les opérateurs précédents afin de formaliser quelques émotions. Comme nous l'avons dit précédemment, nous nous fondons pour cela sur la théorie des émotions d'Ortony *et al*. [16] telle que nous l'avons décrite dans le paragraphe concerné en Section 2. Nous nous focalisons exclusivement sur la classe d'émotions *consequences of events* qui suffit à elle seule à donner un aperçu de la formalisation. Comme le montre FIG. 1 cette classe contient trois groupes principaux d'émotions qui sont détaillés ci-dessous.

Du point de vue formel, l'idée est de capturer la notion de plaisir par notre opérateur de préférence : dire que p nous plaît c'est dire qu'on préfère que p soit vrai (indépendamment du fait qu'on pense que p est actuellement vrai ou même qu'il le sera un jour). De façon similaire, dire que p nous déplaît c'est dire qu'on

préférerait que $\neg p$ soit vrai (*i.e.* que p soit faux). Soit :

$$Pleased_i\,\varphi \stackrel{déf}{=} Pref_i\,\varphi \qquad (\text{Déf}_{Pleased_i})$$

$$Displeased_i\,\varphi \stackrel{déf}{=} Pref_i\,\neg\varphi \qquad (\text{Déf}_{Displeased_i})$$

La notion de désirabilité (resp. indésirabilité) d'un événement, quant à elle, est relative de manière très générale à la satisfaction (resp. l'insatisfaction) des buts par cet événement. À partir du moment où un événement satisfait au moins en partie un de nos buts, alors il est désirable. Ainsi, nous dirons qu'un événement est désirable (resp. indésirable) si et seulement si, par définition, on préfère qu'une de ses conséquences soit vraie (resp. fausse) dans le futur. Soit :

$$Desirable_i\,\varphi \stackrel{déf}{=} Pref_i\,F\varphi \qquad (\text{Déf}_{Desirable_i})$$

$$Undesir_i\,\varphi \stackrel{déf}{=} Pref_i\,F\neg\varphi \qquad (\text{Déf}_{Undesir_i})$$

Le fait de placer la conséquence de l'événement dans le futur permet qu'un événement puisse être désirable même s'il n'a pas encore eu lieu. Par ailleurs, il découle de nos principes logiques que si on préfère quelque chose dans le présent, alors c'est qu'on le préfère dans le futur (il faut voir le futur au sens large, c.-à-d. incluant le présent) : par conséquent, un événement venant juste d'avoir lieu pourra être désirable à partir du moment où cet effet est plaisant, soit :

$$Pleased_i\,\varphi \rightarrow Desirable_i\,\varphi \qquad (1)$$

$$Displeased_i\,\varphi \rightarrow Undesir_i\,\varphi \qquad (2)$$

Comme nous l'avons indiqué dans le paragraphe intitulé « Peut-on éprouver des émotions contradictoires ? » (Section 2), les psychologues sont majoritairement d'avis que la réponse à cette question est positive. Selon nous, cela dépend si notre émotion porte sur un événement ou sur une des conséquences précises de cet événement : dans ce dernier cas, il semble peu intuitif de considérer des émotions contradictoires. En revanche pour Ortony *et al.*, la notion de désirabilité porte plutôt sur des événements et ceux-ci peuvent engendrer des effets positifs et des effets négatifs. Du point de vue de la logique, la conséquence d'un événement peut être désirable et indésirable mais à condition que ce soit à des instants différents. Subséquemment, comme les notions de plaisir et de déplaisir sont situées dans le présent (donc à un même instant temporel), la présence de l'un impose l'absence de l'autre, ce dont rend compte l'axiome (D_{Pref_i}) qui peut se réécrire : $Pleased_i\,\varphi \rightarrow \neg Displeased_i\,\varphi$.

Les émotions engendrées par ces notions de (dé)plaisir et de (in)désirabilité ne pourront pas être des émotions opposées (au sens de Ortony *et al.*). Par exemple, on ne pourra pas être en même temps joyeux et peiné du fait que φ soit vrai, sachant que la joie et la peine appartiennent à un même groupe d'émotions (les *well-being emotions*).

Enfin, si on suppose que $\Box$ représente respectivement chacun des quatre opérateurs de plaisir, déplaisir, désirabilité et indésirabilité, alors nous pouvons prouver que $\Box\varphi \leftrightarrow Bel_i\,\Box\varphi$ est une propriété valide (*cf.* [22, p. 84] par exemple pour la relation croyance/préférence). Cela signifie qu'un agent est conscient de ce qui lui est plaisant, déplaisant, désirable et indésirable. (Il ne peut ni se tromper sur ce qu'il ressent, ni ne pas être au courant.)

4.1 Les *well-being emotions*

Dans ce premier groupe d'émotions (comme dans le suivant, *cf.* Section 4.2, et contrairement au dernier, *cf.* Section 4.3), les réactions des agents aux événements sont plus ou moins indépendantes des attentes que cet agent pouvait initialement (*i.e.* avant que ces événements ne se produisent) avoir à propos de ces événements. Cela ne signifie pas que les *well-being emotions* sont indépendantes de ce caractère inattendu (ce dernier influe sur l'intensité de toutes les émotions), mais que ce caractère ne fait pas partie de leurs conditions d'obtention.

APPRAISAL OF EVENT	
DESIRABLE	UNDESIRABLE
pleased about a desirable event (*e.g.*, joy)	displeased about an undesirable event (*e.g.*, distress)

TAB. 1 – *Les* well-being emotions *selon [16]*

Selon Ortony *et al.*, quand un événement (et plus particulièrement son caractère désirable) nous concerne en propre, alors le plaisir qu'on retire de son occurrence est nécessairement congruent avec sa désirabilité (ce qui explique pourquoi TAB. 1 n'explicite pas les cas du plaisir face à l'indésirable ni du déplaisir face au désirable).

Dans le cas où la durée de l'événement considéré est très courte, il apparaît peu plausible que le fait que cet événement était (in)désirable n'était pas le cas juste avant que cet événement se produise (à moins de supposer que c'est l'évé-

nement lui-même qui a provoqué cette (in)désirabilité ou qu'elle s'est produite de manière fortuite). Il apparaît donc nécessaire que la conséquence de l'événement soit désirable au plus tard l'instant juste avant.

De plus, les définitions des émotions telles qu'elles sont données par Ortony *et al*. ne disent rien à propos de l'information les ayant déclenchées : pour que l'agent i soit joyeux (respect. peiné) que φ il faut nécessairement qu'il croit que φ soit actuellement vrai. Nous supposons que les croyances sont plus « versatiles » que les préférences : en ce sens, il est peu probable qu'un agent éprouve de la joie parce qu'il croyait φ depuis un certain temps et qu'il vient juste de se mettre à préférer que φ soit vrai alors que l'instant d'avant il n'avait pas cette préférence. (Un tel mécanisme pourrait être assimilé à un processus d'auto-déclenchement des émotions qui ne correspond pas au type de déclenchement que nous souhaitons capturer.) C'est donc la croyance qui joue ici le rôle de déclencheur, et pour qu'un agent soit joyeux ou peiné que φ cela suppose qu'il croit en cet instant φ (seconde condition de déclenchement) alors que l'instant juste avant il ne le croyait pas (ce que nous appelons « la première condition de déclenchement »).

D'après ce qui précède, nous avons donc les définitions formelles suivantes :

$$Joy_i\,\varphi \stackrel{déf}{=} Pleased_i\,\varphi \wedge Bel_i\,\varphi \wedge Bel_i\,X^{-1}(\neg Bel_i\,\varphi \wedge Desirable_i\,\varphi) \qquad (\text{Déf}_{Joy_i})$$

$$Distress_i\,\varphi \stackrel{déf}{=} Displeased_i\,\varphi \wedge Bel_i\,\varphi \wedge Bel_i\,X^{-1}(\neg Bel_i\,\varphi \wedge Undesir_i\,\varphi) \qquad (\text{Déf}_{Distress_i})$$

Autrement dit, l'agent i éprouve de la joie (resp. de la peine) à propos du fait que φ est vrai dès lors qu'il lui est plaisant (resp. déplaisant) que φ soit vrai et qu'il croit que φ est actuellement vrai alors que l'instant juste avant il envisageait la possibilité que φ soit faux et φ lui était désirable.

Par ailleurs, la conjonction des deux conditions de déclenchement a également pour conséquence que l'instant qui suit juste le déclenchement de l'émotion, il n'éprouvera plus cette émotion. Bien sûr, ce n'est en général pas réellement *juste l'instant d'après* qu'une émotion disparaît (cette limite nous est imposée par le fait qu'on ne gère pas les degrés d'intensité des émotions), mais les psychologues s'accordent à dire que cette intensité décroît très rapidement (selon une courbe de type exponentielle inverse) et cela constitue donc une idéalisation de ce phénomène et non quelque chose qui va à son encontre.

Enfin, par (D_{Bel_i}) et (D_{Pref_i}), il est aisé de montrer que les propriétés suivantes

sont valides :

$$Joy_i\,\varphi \rightarrow \neg Joy_i\,\neg\varphi \tag{3}$$

$$Distress_i\,\varphi \rightarrow \neg Distress_i\,\neg\varphi \tag{4}$$

$$Joy_i\,\varphi \vee Joy_i\,\neg\varphi \rightarrow \neg Distress_i\,\varphi \wedge \neg Distress_i\,\neg\varphi \tag{5}$$

(3) et (4) illustrent que le fait qu'être joyeux (resp. peiné) que φ soit vrai implique nécessairement de ne pas être joyeux (resp. peiné) que φ soit faux. Ces propriétés sont tout à fait intuitives puisqu'être joyeux ou peiné que φ soit vrai suppose que l'agent croit que φ est vrai. Il serait donc irrationnel d'être joyeux ou peiné du contraire puisque cela supposerait que l'agent croit également que φ est faux.

(5) indique que si on est joyeux que φ soit vrai ou joyeux que φ soit faux, alors nécessairement on ne peut être ni triste que φ soit vrai ni triste que φ soit faux. Intuitivement, si on est joyeux que φ est vrai, on ne peut :

- ni être triste que φ soit vrai
  (cette propriété est relative à notre hypothèse selon laquelle, quand on évalue la conséquence d'un événement plutôt que l'événement lui-même, alors on ne peut éprouver deux émotions contradictoires à propos de cette conséquence — *cf.* Section 2 le paragraphe intitulé « Peut-on éprouver des émotions contradictoires ? »)
- ni être triste que φ soit faux
  (cette propriété est relative à la rationalité des croyances de l'agent : être joyeux que φ implique que l'agent croit actuellement que φ est vrai et ne peut donc croire en même temps que φ est faux, ce que suppose le fait qu'il soit peiné que φ soit faux).

Un raisonnement similaire conduit à expliquer intuitivement qu'être joyeux que φ soit faux implique les mêmes contraintes sur la tristesse à propos de φ et de $\neg\varphi$.

Enfin, signalons que $\neg Joy_i\,(p \vee \neg p)$ (nous ne sommes pas joyeux à propos des tautologies — ces propositions qui sont toujours vraies) est une propriété valide, et que notre opérateur de joie n'est pas fermé par implication matérielle (*i.e.* $Joy_i\,\varphi \wedge Joy_i\,(\varphi \rightarrow \psi)$ n'implique pas que $Joy_i\,\psi$).

Par ailleurs, par contraposition de (5) la propriété suivante est également valide : $Distress_i\,\varphi \vee Distress_i\,\neg\varphi \rightarrow \neg Joy_i\,\varphi \wedge \neg Joy_i\,\neg\varphi$.

4.2 Les *fortune-of-others emotions*

Comme le groupe précédent, le groupe des *fortune-of-others emotions* concerne des événements se produisant sans que l'agent considéré ait une quel-

conque attente à propos de l'occurrence de cet événement. En reprenant les mêmes définitions que précédemment pour le plaisir et la désirabilité, les types d'émotion présentés en TAB. 2 peuvent être définis formellement de la manière suivante :

REACTION OF SELF	PRESUMED VALUE FOR OTHER	
	DESIRABLE	UNDESIRABLE
PLEASED	pleased about an event desirable for someone else (*e.g.*, happy-for)	pleased about an event undesirable for someone else (*e.g.*, gloating)
DISPLEASED	displeased about an event desirable for someone else (*e.g.*, resentment)	displeased about an event undesirable for someone else (*e.g.*, sorry-for)

TAB. 2 – *Les* fortunes-of-others emotions *selon [16]*

Ainsi, selon TAB. 2, l'émotion ***happy-for*** se formalise de la manière suivante :

$$HappyFor_{i,j}\varphi \stackrel{déf}{=} Bel_i\,\varphi \wedge Pleased_i\,\varphi \wedge Bel_i\,X^{-1}(Bel_i\,Desirable_j\,\varphi \wedge \neg Bel_i\,\varphi) \qquad (\text{Déf}_{HappyFor_{i,j}})$$

Autrement dit, i est content pour j que φ si et seulement si i croit que φ est vrai, que ça lui est plaisant, et qu'il pensait l'instant juste avant que c'était désirable pour j et que φ n'était pas encore vrai.

Il est important de remarquer que i peut être content pour j que φ soit vrai indépendamment du fait que j soit au courant ou non que φ est vrai. Il est donc nécessaire que la désirabilité de j soit une croyance passée de i (processus de mémoire au sens de Remarque 1) de manière à marquer le fait que : i) c'est (en partie) parce que i croyait l'instant juste avant[2] que φ était désirable pour j et qu'il pense maintenant que φ est vrai qu'il éprouve cette émotion de ***happy-for*** ; ii) ce déclenchement est indépendant du fait qu'il était désirable pour j et/ou qu'il est toujours actuellement désirable pour j que φ soit vrai (la croyance de i à ce sujet suffit, même si cette croyance est erronée).

De plus, comme dans le groupe d'émotions précédent, les deux conditions de déclenchement assurent que l'agent i éprouve cette émotion à cet instant et qu'il ne l'éprouvait ni l'instant juste avant, ni celui juste après.

De façon similaire, les autres émotions décrites en TAB. 2 se formalisent de la manière suivante :

2. *i.e.* l'instant juste avant de croire que φ est vrai

$$Gloating_{i,j}\varphi \stackrel{déf}{=} Bel_i\,\varphi \wedge Pleased_i\,\varphi \wedge Bel_i\,X^{-1}(Bel_i\,Undesir_j\,\varphi \wedge \neg Bel_i\,\varphi) \quad (\text{Déf}_{Gloating_{i,j}})$$

$$Resentment_{i,j}\varphi \stackrel{déf}{=} Bel_i\,\varphi \wedge Displeased_i\,\varphi \wedge Bel_i\,X^{-1}(Bel_i\,Desirable_j\,\varphi \wedge \neg Bel_i\,\varphi) \quad (\text{Déf}_{Resentment_{i,j}})$$

$$SorryFor_{i,j}\varphi \stackrel{déf}{=} Bel_i\,\varphi \wedge Displeased_i\,\varphi \wedge Bel_i\,X^{-1}(Bel_i\,Undesir_j\,\varphi \wedge \neg Bel_i\,\varphi) \quad (\text{Déf}_{SorryFor_{i,j}})$$

($\text{Déf}_{Gloating_{i,j}}$), qui représente la jubilation malveillante, se déclenche lorsque l'agent i trouve que φ est plaisant tout en pensant que, l'instant juste avant que φ devienne vrai, φ était indésirable pour l'agent j.

($\text{Déf}_{Resentment_{i,j}}$) signifie que l'agent i éprouve du ressentiment envers j à propos de φ si et seulement si φ est actuellement vrai et déplaisant pour l'agent i, et que celui-ci croyait l'instant juste avant que φ devienne vrai que φ était désirable pour l'agent j.

Enfin, ($\text{Déf}_{SorryFor_{i,j}}$) signifie que l'agent i est désolé pour l'agent j à propos de φ si et seulement si φ est actuellement vrai et déplaisant pour l'agent i, et que celui-ci croyait l'instant juste avant que φ devienne vrai que φ était également indésirable pour l'agent j.

D'après les définitions, les propriétés suivantes sont valides :

$$(HappyFor_{i,j}\varphi \vee Gloating_{i,j}\varphi) \wedge Bel_i\,X^{-1}Desirable_i\,\varphi \rightarrow Joy_i\,\varphi \quad (6)$$

$$(Resentment_{i,j}\varphi \vee SorryFor_{i,j}\varphi) \wedge Bel_i\,X^{-1}Undesir_i\,\varphi \rightarrow Distress_i\,\varphi \quad (7)$$

Autrement dit, si l'agent i est content pour l'agent j ou qu'il jubile de façon malveillante par rapport à l'agent j du fait que φ soit vrai, et que l'instant d'avant φ était désirable pour i, alors i est content que φ soit vrai. De façon similaire, si l'agent i éprouve du ressentiment vis-à-vis de l'agent j ou est désolé pour lui que φ soit vrai, et qu'il se rappelle que l'instant juste avant φ lui était indésirable, alors il éprouve de la peine que φ soit vrai. La condition selon laquelle l'instant d'avant, φ était désirable pour l'agent i est importante : elle permet de discerner les cas où cet agent est simplement content pour quelqu'un du fait que φ est vrai, des cas où en plus, à titre personnel, il est content de ce fait. (Cette distinction s'applique bien sûr pour les trois autres *fortunes-of-others emotions*.)

4.3 Les *prospect-based emotions*

Ce groupe d'émotions concerne celles ressenties en réponse à des événements attendus, ou à des événements confirmés ou infirmés [16, Chap. 6]. En général, ces derniers événements sont passés, mais pas nécessairement : un agent peut éprouver les émotions associées de façon contrefactuelle, c.-à-d. en pensant à ce qui serait arrivé si tel événement avait (ou n'avait pas) eu lieu. De même, les émotions relatives à des événements attendus se produisent généralement avant ces événements, mais pas toujours : on peut être effrayé (émotion particulière de type *fear*) après un accident à l'idée de ce qui aurait pu nous arriver si on avait roulé plus vite, ou si on était arrivé sur les lieux une seconde plus tôt ou plus tard *etc*. Pour des questions de simplicité du formalisme, nous ne traitons pas dans les définitions que nous donnons ci-dessous les cas de raisonnement contrefactuel.

Un événement pouvant être d'une part attendu, confirmé ou infirmé, et d'autre part désirable ou indésirable, Ortony *et al*. classent dans ce groupe six types d'émotion différents récapitulés en TAB. 3.

STATUS OF EVENT	APPRAISAL OF PROSPECTIVE EVENT	
	DESIRABLE	UNDESIRABLE
UNCONFIRMED	pleased about the prospect of a desirable event (*e.g.*, hope)	displeased about the prospect of an undesirable event (*e.g.*, fear)
CONFIRMED	pleased about the confirmation of the prospect of a desirable event (*e.g.*, satisfaction)	displeased about the confirmation of the prospect of an undesirable event (*e.g.*, fear-confirmed)
DISCONFIRMED	displeased about the disconfirmation of the prospect of a desirable event (*e.g.*, disappointment)	pleased about the disconfirmation of the prospect of an undesirable event (*e.g.*, relief)

TAB. 3 – *Les* prospect-based emotions *selon [16]*

Les types d'émotion présents sur la première ligne de TAB. 3 (labellisées *unconfirmed*) sont relatifs à des attentes particulières et dont l'agent concerné ne sait pas encore si elles sont réalisées ou non (et ce, indépendamment du fait qu'elles le soient réellement ou non : seules les croyances de l'agent à ce sujet sont importantes).

Comme pour les autres groupes, les types d'émotion décrits représentent des

ensembles d'émotions qui diffèrent en partie par leur intensité. C'est par exemple le cas de la crainte et de la peur qui sont toutes deux des instances (*token*) du type d'émotion *fear*.

Les émotions portant sur une information ni confirmée ni infirmée (statut *unconfirmed*) sont au nombre de deux :

$$Hope_i\,\varphi \stackrel{déf}{=} \neg Bel_i\,\varphi \wedge \neg Bel_i\,\neg\varphi \wedge Pleased_i\,\varphi \wedge \neg Bel_i\,\neg F\varphi \qquad (\text{Déf}_{Hope_i})$$

$$Fear_i\,\varphi \stackrel{déf}{=} \neg Bel_i\,\varphi \wedge \neg Bel_i\,\neg\varphi \wedge Displeased_i\,\varphi \wedge \neg Bel_i\,\neg F\neg\varphi \qquad (\text{Déf}_{Fear_i})$$

Autrement dit, l'agent i espère (resp. redoute) que φ soit vrai si et seulement si il ne sait pas si φ est actuellement vrai ou non (*i.e.* il envisage simultanément au moins une histoire possible où φ est vrai et au moins une histoire possible où φ est faux), que φ lui est plaisant (resp. déplaisant), et qu'il envisage au moins une histoire où φ sera vrai (resp. faux) dans le futur. (D'après (D_{Pref_i}), le fait que φ soit plaisant (resp. déplaisant) implique que φ lui est désirable (resp. indésirable.)

Ce dernier point caractérise l'espoir : si l'on supposait que l'agent croit que, dans le futur de toutes les histoires qu'il envisage, φ est faux, il serait difficile de comprendre ce qui le fait espérer... D'autre part, l'existence d'au moins une histoire suffit : le fait qu'il y ait plus d'histoires où φ sera vrai dans le futur que d'histoires où φ ne sera pas vrai, n'est pas caractéristique de l'espoir en lui-même et ne joue que sur son intensité.

Il est également important que l'agent ne sache pas si φ est vrai ou non (*i.e.* $\neg Bel_i\,\varphi \wedge \neg Bel_i\,\neg\varphi$) pour avoir l'espoir (ou la crainte) car cela rend compte de l'attente de l'agent. Par exemple, supposons que Jean regarde un match de foot à la télévision et que son équipe favorite est en train de gagner, mais que le match n'est pas fini. Supposons en outre que Hubert arrive à ce moment là, et qu'il n'est pas au courant du score. Quand Hubert demande si son équipe favorite gagne, Jean répondra certainement quelque chose du genre : « Pour l'instant, elle mène. J'espère vraiment qu'elle va gagner. ». L'espoir de Jean porte clairement sur le fait que l'équipe va gagner, et même s'il a de bonnes raisons de croire que ça va être le cas, il ne peut pas affirmer que son équipe a gagnée tant que le match n'est pas fini : il ne peut donc croire que son équipe *a gagné* (ou même perdu) en même temps qu'il espère qu'elle *va gagner*.

Les types d'émotion de la seconde ligne de TAB. 3 concernent des attentes confirmées par les événements qui viennent de se produire.

$$Satisfaction_i\,\varphi \stackrel{déf}{=} X^{-1}Hope_i\,\varphi \wedge Bel_i\,\varphi \wedge Pleased_i\,\varphi \qquad (\text{Déf}_{Satisfaction_i})$$

$$FearConfirmed_i\,\varphi \stackrel{déf}{=} X^{-1}Fear_i\,\varphi \wedge Bel_i\,\varphi \wedge Displeased_i\,\varphi \quad (\text{Déf}_{FearConfirmed_i})$$

($\text{Déf}_{Satisfaction_i}$) définit la satisfaction que φ soit vrai comme le fait qu'on espérait l'instant juste avant que φ allait être vrai et qu'il est actuellement vrai, ce qui est plaisant. De façon similaire, ($\text{Déf}_{FearConfirmed_i}$) définit la confirmation d'un crainte comme le fait que l'instant juste avant on craignait que φ ne soit vrai et que l'on croit actuellement que c'est le cas, ce qui est déplaisant.

Enfin, les types d'émotion de la dernière ligne de TAB. 3 concernent des attentes contrariées (*disconfirmed*) par les événements qui viennent de se produire.

$$Disappointment_i\,\varphi \stackrel{déf}{=} X^{-1}Hope_i\,\varphi \wedge Bel_i\,\neg\varphi \wedge Displeased_i\,\neg\varphi \qquad (\text{Déf}_{Disappointment_i})$$

$$Relief_i\,\varphi \stackrel{déf}{=} X^{-1}Fear_i\,\varphi \wedge Bel_i\,\neg\varphi \wedge Pleased_i\,\neg\varphi \qquad (\text{Déf}_{Relief_i})$$

($\text{Déf}_{Disappointment_i}$) signifie que l'agent i est désappointé que φ soit vrai quand il espérait l'instant juste avant que φ soit vrai et qu'il croit désormais que φ est faux, ce qui lui est déplaisant. Enfin, (Déf_{Relief_i}) signifie que l'agent i est soulagé quand il craignait l'instant juste avant que φ soit vrai et qu'il croit maintenant que φ est faux, ce qui lui est plaisant.

5 Conclusion

Sans prétendre à l'exhaustivité ni à l'unicité de la représentation adoptée, notre but était de montrer comment nous pouvions appréhender formellement des concepts aussi complexes que ceux présents dans les émotions. En cela, notre but était d'apporter une contribution indicative des perspectives ouvertes par la logique pour une représentation rigoureuse de certaines formes d'émotion et de la rationalité qui peut les sous-tendre. Une telle approche pourrait contribuer à donner une base théorique à la compréhension des relations qu'entretiennent l'art et l'émotion, sachant que cette compréhension passe par celle de l'un et de l'autre.

Dans [11] Goodman aborde cette question des liens entre art et émotion, et souligne sa difficulté. Celle-ci semble due en partie à « la distinction entre le scientifique et l'esthétique », elle-même « enracinée dans la différence entre connaître et ressentir, entre le cognitif et l'émotif. » Il souligne combien cette dichotomie est

douteuse car pour lui « l'expérience esthétique tout comme l'expérience scientifique a fondamentalement un caractère cognitif » [11, p. 287]. (Voir Section 2 pour une discussion entre émotion et cognition.) L'évaluation d'une œuvre artistique ne peut se faire que dans la comparaison avec d'autres, par l'analyse, bref, par des processus cognitifs de pensées qui sont totalement étrangers à une perception de l'œuvre exclusivement en termes d'émotions ressenties.

Mais au delà de ce constat, nulle réponse : la relation entre une œuvre d'art et les émotions qu'elle suscite semble revêtir différentes propriétés contradictoires, autrement dit aucune. L'expression de la colère peut ne susciter que le mépris, celle de peur ou de haine peut faire naître des sentiments positifs. Il n'est même pas question de degré d'intensité quand certaines œuvres, par un manque flagrant d'émotion vont inspirer de puissants sentiments, ou au contraire par un étalage émotionnel outrancier inspirent le plus froid détachement. À l'extrême, certaines œuvres peuvent être totalement dénuées d'émotion : en sont-elles moins artistiques pour autant ? Certes non, et cela semble supposer que l'émotion est plus la propriété (ou non) associée à une œuvre artistique plutôt qu'une de ses caractéristiques qui en font une œuvre en soi. Peut-être une connaissance plus précise et plus profonde des caractéristiques de l'émotion pourrait-elle contribuer à dénouer ce puzzle.

Références

[1] ARISTOTE : *Rhétorique*. Gf. Flammarion, 2007. Trad. Pierre Chiron.

[2] Ismay BARWELL : How Does Art Express Emotion? *The Journal of Aesthetics and Art Criticism*, 45(2):175–181, 1986.

[3] J. BATES : The role of emotion in believable agents. *Communications of the ACM*, 37(7), 1994.

[4] John P. BURGESS : Basic tense logic. *In* Dov GABBAY et Franz GUENTHNER, éditeurs : *Handbook of Philosophical Logic*, volume 7, pages 1–42. Kluwer Academic Publishers, 2nd édition, 2002.

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

[6] Philip R. COHEN et Hector J. LEVESQUE : Intention is choice with commitment. *Artificial Intelligence Journal*, 42(2–3):213–261, 1990.

[7] Philip R. COHEN et Hector J. LEVESQUE : Persistence, intentions, and commitment. *In* Philip R. COHEN, Jerry MORGAN et Martha E. POLLACK, éditeurs : *Intentions in Communication*. MIT Press, 1990.

[8] E.E. CUMMINGS : *Complete poems*, volume I. McGibbon & Kee, 1973.

[9] J.P. FORGAS : Mood and judgment: The affect infusion model (aim). *Psychological Bulletin*, 117:39–66, 1995.

[10] Paul GOCHET et Pascal GRIBOMONT : Epistemic Logic. *In* Dov GABBAY et John WOODS, éditeurs : *Twentieth Century Modalities*, volume 7 de *Handbook of the History of Logic*, pages 99–195. Elsevier, amsterdam édition, 2006.

[11] Nelson GOODMAN : *Langages de l'art*. Éditions Jacqueline Chambon, Nîmes, France, 1990.

[12] Richard LAZARUS : Thoughts on the Relation between Emotion and Cognition. *American Psychologist*, 37(9):1019–1024, 1982.

[13] Richard LAZARUS : On the Primacy of Cognition. *American Psychologist*, 39(2):124–129, 1984.

[14] Richard LAZARUS : The Cognition–Emotion Debate: a Bit of History. *In* Tim DALGLEISH et Mick POWER, éditeurs : *Handbook of Cognition and Emotion*, pages 3–20. John Wiley & Sons, New York, 1999.

[15] Richard S. LAZARUS : *Emotion and Adaptation*. Oxford University Press, 1991.

[16] Andrew ORTONY, G.L. CLORE et A. COLLINS : *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA, 1988.

[17] J. PIAGET : Les émotions. *In* B. RIMÉ et K. SCHERER, éditeurs : *Les relations entre l'intelligence et l'affectivité dans le developpement de l'enfant*, pages 75–95. Delachaux et Niestlé, Neuchâtel-Paris, 1989.

[18] PLATON : *La République*. Flammarion, 2002.

[19] Anand S. RAO et Michael P. GEORGEFF : Modeling rational agents within a BDI-architecture. *In* J. A. ALLEN, R. FIKES et E. SANDEWALL, éditeurs : *Proc. Second Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann Publishers, 1991.

[20] Anand S. RAO et Michael P. GEORGEFF : An abstract architecture for rational agents. *In* Bernhard NEBEL, Charles RICH et William SWARTOUT, éditeurs : *Proc. Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pages 439–449. Morgan Kaufmann Publishers, 1992.

[21] Tone ROALD : *Cognition in Emotion. An investigation through Experiences with Art*. Numéro 10 de Consciousness Literature and the Arts. Editions Rodolpi B.V., Amsterdam, 2007.

[22] M. D. SADEK : *Attitudes mentales et interaction rationnelle : vers une théorie formelle de la communication*. Thèse de doctorat, Université de Rennes I, Rennes, France, 1991.

[23] M.D. SADEK : A study in the logic of intention. *In* Bernhard NEBEL, Charles RICH et William SWARTOUT, éditeurs : *Proc. Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pages 462–473. Morgan Kaufmann Publishers, 1992.

[24] N. SCHWARZ et G. L. CLORE : Mood, misattribution, and judgments of well-being. *Journal of Personality and Social Psychology*, 45:513–523, 1983.

[25] John R. SEARLE : *Du cerveau au savoir*. Hermann, 1985.

[26] Robert C. SOLOMON : The Philosophy of Emotion. *In* Michael LEWIS, Jeannette M. HAVILAND-JONES et Lisa FELDMAN BARRETT, éditeurs : *Handbook of Emotions*, pages 3–16. The Guilford Press, New York, 3rd édition, 2008.

[27] Baruch SPINOZA : *L'éthique*, volume 235 de *Folio Essais*. Gallimard, 1994. Trad. Roland Callois du texte de 1677.

[28] W. WUNDT : *Outlines of psychology*. Wilhelm Englemann, Leipzig, 1907. translation of *Gundriss der Psychologie*, Leipzig: Wilhelm Englemann, 1905.

[29] R.B. ZAJONC : Feeling and Thinking – Preferences Need No Inferences. *American psychologist*, 35(2):151–175, février 1980.

[30] Robert B. ZAJONC : On the Primacy of Affect. *American Psychologist*, 39(2):117–123, 1984.

[31] Robert B. ZAJONC : Feeling and Thinking; closing the Debate over the Independence of Affect. *In* J.P. FORGAS, éditeur : *Feeling and Thinking: the Role of Affect in Social Cognition*, pages 31–59. Cambridge University Press, 2000.