Université de Montréal

Les émotions comme opérateurs des engagements: Une métaphore pour les structures de contrôle dans les systèmes multi-agents

par

Michel Aubé

Département de Didactique

Faculté des sciences de l'éducation

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiae Doctor (Ph.D.)
en sciences de l'éducation, option didactique

avril 1997

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Les émotions comme opérateurs des engagements: Une métaphore pour les structures de contrôle dans les systèmes multi-agents

présentée par:

Michel Aubé

a été évaluée par un jury composé des personnes suivantes:

Gisèle Lemoyne — président-rapporteur
Alain Berthiaume — directeur de recherche
Alain d'Anglejan — codirectrice
Jesus Vasquez-Abad — membre du jury
Keith Oatley — examinateur externe

Thèse acceptée le: 22. 10. 1997

Sommaire

La thèse se présente sous la forme d'une série de quatre articles posant différents jalons d'un modèle théorique et fonctionnel des émotions, comme fondement pour les structures de contrôle dans les systèmes multi-agents, en intelligence artificielle distribuée (IAD). Bien que d'abord intéressé par la question de la motivation scolaire, il nous avait cependant semblé qu'une percée significative en ce domaine requérait une approche théorique fondamentale. Le premier article situe donc le problème de recherche dans le contexte de l'éducation, tout en faisant ressortir la nécessité de disposer de modèles théoriques robustes pour aborder le domaine de la motivation et des émotions. Le texte rappelle les principales contributions de l'intelligence artificielle (IA) pour la compréhension des mécanismes d'apprentissage et de construction des savoirs, et fait ressortir l'utilité des spécifications ainsi obtenues pour la conception de stratégies pédagogiques appropriées. L'article propose ensuite d'appliquer l'approche de l'IA pour la conceptualisation des phénomènes motivationnels et émotionnels, et donne un aperçu du genre de modélisation qui pourrait en résulter, ainsi que des applications qui pourraient en découler, afin de rendre plus efficaces les interventions pédagogiques destinées à rehausser la motivation scolaire.

Le second article effectue une recension approfondie en psychologie cognitive des émotions. Sa fonction essentielle, dans le cadre de la thèse, est de repérer les concepts fondamentaux pour la formulation du modèle théorique et d'accumuler un large répertoire d'exemples, aptes à supporter ou à confronter la théorisation proposée. Nous avons notamment tiré profit des principales controverses en psychologie cognitive des émotions comme révélateur des concepts-clés et des articulations théoriques les plus significatives. De cette méta-analyse a émergé l'hypothèse que les structures émotionnelles constituaient une couche particulière de contrôle, entre celle qui assure la gestion des besoins et celle qui supporte le fonctionnement rationnel. La recension a également permis d'identifier et de retenir, comme composantes de cette couche de contrôle, un petit nombre de mécanismes de base apparaissant comme responsables de la très grande majorité des comportements émotionnels, et d'articuler avec précision les conditions de leur déclenchement.

Le troisième article expose le modèle "computationnel" proprement dit, et notamment le lien fondamental que nous établissons entre motivation, gestion des ressources et engagements. Ces derniers y sont expressément conceptualisés comme un type particulier de ressources reposant essentiellement sur la réciprocité et la collaboration entre les agents. Les émotions sont alors présentées comme des *opérateurs des engagements* dont la fonction est d'assurer la gestion et la régulation des variations significatives qui peuvent se produire à l'intérieur de ce stock de ressources particulières, essentielles pour rendre possible et maintenir de façon adaptative la coopération comme fondement des espèces sociales. L'article est d'ailleurs précédé d'une section approfondissant les phénomènes de coopération et de réciprocité, ainsi que les conditions favorisant leur émergence.

Le dernier article explore, dans une perspective évolutionniste, les différentes contraintes qui président à l'apparition (dans l'évolution) ou à la conception (en ingénierie) d'agents biologiques ou informatiques véritablement doués d'autonomie. La capacité pour un organisme de déterminer ses propres motivations est présentée comme la caractéristique première de l'autonomie, et la gestion des ressources est abordée comme le principe essentiel derrière tout mécanisme d'ordre motivationnel. Les engagements étant envisagés dans notre modèle comme un type particulier de ressources, le texte réexamine à cette lumière le concept tel qu'utilisé en intelligence artificielle distribuée. L'article se termine par une énumération des spécifications qui découlent du modèle, en vue d'une éventuelle implantation dans un environnement informatique à base d'acteurs.

La conclusion générale rassemble les contributions originales de la recherche, notamment en ce qui concerne le rôle primordial joué par le concept d'engagement dans la compréhension des phénomènes émotionnels. Ce construit hypothétique, tel que défini et utilisé dans la thèse, offre en effet une valeur descriptive et explicative exceptionnelle. Il permet d'établir une distinction plus claire que jamais entre les besoins et les émotions, il rend compte de la nature intrinsèquement sociale des émotions, et il offre un cadre cohérent pour la répartition taxinomique des différentes catégories d'émotions. Au profit de la recherche en IAD, il offre une base pour l'explicitation des contraintes qui président à la coopération entre les agents, et permet de formuler des spécifications précises pour une implantation de structures de contrôle remplissant les fonctions principales des mécanismes émotionnels chez les vivants. Finalement, dans le contexte de la motivation scolaire, il offre un champ nouveau d'exploration pour mieux comprendre les bases intrinsèques et interactives sur lesquelles reposent la persistance et l'engagement des élèves.

Table des matières

Liste des tableaux

Liste des figures

Remerciements

L'un des postulats fondamentaux de la présente recherche, est que les connaissances en général, et les contributions scientifiques en particulier, ne se construisent jamais en vase clos, mais sont plutôt le résultat d'efforts multiples, distribués sur un vaste réseau de personnes. Par ailleurs, l'un des résultats significatifs de ce travail, est que les engagements qui relient entre eux les agents, constituent l'une des ressources essentielles dont est tissé ce réseau, et que les émotions - comme par exemple la gratitude qui s'exprime dans ces remerciements - ont surtout pour fonction de régénérer ces ressources, et d'en assurer la protection et la circulation.

En tout premier lieu, je souhaite exprimer une reconnaissance chaleureuse aux deux personnes qui ont assuré la direction de cette recherche. Alain Senteni a fait preuve d'une audace singulière en acceptant de superviser un sujet qui, en prétendant que les émotions pouvaient contribuer à la conception et à l'implantation de systèmes informatiques plus performants, apparaissait pour le moins périlleux dans la communauté "rationaliste" de l'intelligence artificielle. Je lui dois en particulier de m'avoir appris, de façon efficace et quasi initiatique, à pénétrer les réseaux internationaux de diffusion et de publication scientifiques. Lorsque Monsieur Senteni a quitté le pays pour oeuvrer comme professeur invité à l'étranger, Madame Alison d'Anglejan a généreusement proposé à la Faculté des études supérieures d'agir en co-supervision jusqu'à la fin du projet. Sa formation en psychologie, son sens exceptionnel de la rigueur scientifique et ses connaissances approfondies de la langue anglaise - dans laquelle ont été rédigés chacun des articles - ont contribué largement au progrès et à la qualité du produit final. Je dois également les remercier tous les deux de l'énorme patience dont ils ont fait preuve, après mon retour au travail, devant ma lenteur et mon obsession à finaliser soigneusement la rédaction finale.

Divers collègues ont également contribué de leurs commentaires et de leurs suggestions sur l'un ou l'autre des articles qui composent cette thèse: Gilles Kirouac de l'Université Laval, Claude Lamontagne de l'Université d'Ottawa, Jacques Tardif et Rolland Viau de l'Université de Sherbrooke, Danielle Bracke, étudiante au doctorat à l'Université de

Montréal, et Ian Wright, étudiant au doctorat à l'Université de Birmingham, en Grande-Bretagne. Qu'ils en soient tous vivement remerciés. Je dois également saluer le soutien généreux de mon employeur, l'Université de Sherbrooke, qui m'a accordé un congé rare et précieux de deux années entières consacrées à ces études doctorales. Je dois également à mes collègues du Département de pédagogie de la Faculté d'éducation d'avoir appuyé ma demande, accepté mon remplacement temporaire et rendu ainsi possible l'exercice d'un tel privilège.

Dans les guides de rédaction et de présentation des mémoires et des thèses, on invite généralement à ce que cette section de remerciements soit empreinte de sobriété, qu'elle se confine aux contributions intellectuelles ou matérielles, et qu'elle reste discrète sur la question des contributions affectives ou privées. Il me semble cependant que le sujet de la présente recherche autorise une reconnaissance plus substantielle. Je pourrais en effet difficilement prétendre essayer de comprendre et de contribuer quoi que ce soit de significatif à la complexe question de la motivation scolaire, en ignorant les racines profondes de ma propre curiosité intellectuelle, qui m'a guidé avec un plaisir soutenu dans l'exploration de problèmes complexes, traversant - et transgressant même - les frontières de disciplines multiples. Je dois certainement à ces pédagogues dans l'âme, mes parents Joseph Aubé et Rollande Poulin, d'avoir entretenu constamment sous mes yeux ce modèle de curiosité, d'émerveillement renouvelé pour les êtres et les choses, et cette ouverture sur le monde qui me fascinent encore chez eux, après presque quatre-vingt ans de vie. Je leur dois aussi et surtout une affection indéfectible, capable néanmoins de fermeté, mais suffisamment empreinte de sécurité pour accueillir aussi la confrontation. Oui, certainement de grands spécialistes de la gestion des engagements!

Une telle recherche s'appuie évidemment sur bien plus que les cinq années de travail qui ont mené à la rédaction finale. Car chercher à comprendre la nature et le fonctionnement des émotions, c'est finalement se pencher avec ardeur sur l'une des dimensions les plus sensibles de notre vie. Dans ce contexte, probablement plus que pour tout autre sujet, le climat de la vie affective quotidienne prend une importance exceptionnelle. J'exprime donc ma tendresse et ma gratitude la plus vive à ma compagne de tous les jours, Danièle Bracke, avec qui j'ai pu éprouver et explorer à loisir, de l'intérieur comme avec recul, toute la richesse et la complexité de nos ressources émotionnelles. Que cette réflexion puisse offrir également à d'autres la même transparence qui nous a éclairés tous les deux.

Bien que n'ayant pas moi-même d'enfants, mes relations étroites et soutenues avec mes neveux et nièces, en particulier Pieter, Sophie, Bernard et Joanne, ont constitué un oasis émotionnel fécond, où l'affection et la réflexion ont toujours circulé allègrement. Il n'est pas évident de nos jours de pouvoir entretenir, avec des jeunes de 13 à 21 ans, le plus souvent à leur propre demande, des discussions élaborées sur la théorie de l'évolution, sur l'instinct et le comportement animal, sur le quotient intellectuel et la nature de l'intelligence, sur le sexisme et les rapports entre hommes-femmes... ou sur le fonctionnement des émotions. L'exemple qui ouvre la thèse illustre par ailleurs l'intensité de l'attachement qui me relie à eux, ainsi que la contribution que ces relations précieuses ont réellement apportée à la réflexion rapportée ici.

Même si l'intention pourra paraître étrange à certains, les gens qui me connaissent bien comprendront que je peux difficilement terminer cette section sans formuler également une pensée pour Brigand, notre vieux Bouvier de Flandres, et l'affectueux compagnon de tous les jours. Comme dans l'exemple de Lorenz relaté au dernier chapitre de la thèse, l'interaction quotidienne avec cette brave bête a généré plus d'une réflexion sur la nature et sur la fonction des structures émotionnelles. Il m'a fait réaliser notamment à quel point sa propre survie dépend comme ressource essentielle de l'engagement que nous avons à son égard, que ses réactions émotionnelles constituent un moyen vital par lequel ces engagements sont chaque jour réactivés, et qu'elles s'avèrent un véhicule extrêmement efficace de l'expression des intentions. Sur la question de la compréhension de l'évolution des structures émotionnelles, le vieux serviteur est parfois devenu le maître.

Outre la présence de Brigand, j'ai aussi la chance d'habiter un environnement dont la richesse naturelle et faunique est exceptionnelle. L'observation attentive et le recensement quasi quotidien des oiseaux et autres bestioles qui peuplent cet environnement immédiat ont également contribué en générant une mine de renseignements et d'enseignements appréciables sur le comportement animal, directement transférables au sujet même de la recherche. D'où sans doute les efforts farouches et les combats répétés pour en assurer la protection et en préserver l'intégrité. Comme j'ai essayé de le faire comprendre tout au long de ce travail, la connaissance résulte fondamentalement d'engagements.

Et elle engage alors à son tour. Pour la survie...

À Dany,

toi qui sais si bien
faire éclore et laisser croître
les fleurs et les émotions
dans le même jardin

Chapitre 1.

Introduction générale

1.1. Introduction: deux anecdotes...

Un soir de juillet 1989, je déambulais dans les rues du Vieux Montréal, avec ma compagne Danièle, ma nièce Sophie, âgée de 11 ans, et mon neveu Pieter, âgé de 13 ans. À une intersection, j'eus l'imprudence de traverser un boulevard à quatre voies, tenant ma nièce par la main, et circulant au travers des voitures immobilisées aux feux rouges. Au moment d'atteindre le trottoir, j'entendis derrière moi le bruit des embrayages, compris que les feux avaient dû virer au vert, et fus pris d'un terrible pressentiment. Je pivotai aussitôt, aperçus ma compagne seule sur le trottoir d'en face, et Pieter qui courait vers nous à vive allure, zigzaguant entre les voitures qui s'apprêtaient à démarrer. Dans le même éclair, j'aperçus, dans l'allée de gauche, celle qui était la plus près de moi, une voiture s'éjectant du tunnel de l'autoroute à pleine vitesse, dans une voie libre et devant un feu vert. Le cri d'horreur qui surgit alors de ma poitrine, probablement le plus violent que j'aie jamais hurlé de toute ma vie, fut si aigu et si intense que les conducteurs et les passagers des voitures à l'arrêt me jetèrent des regards effarés. Mais il eut aussi pour effet que Pieter figea de stupeur... et que la voiture fila à quelques centimètres devant lui, et à quelques millisecondes seulement de le happer de plein fouet!

L'événement fut si fortement imprégné dans mon esprit que, plusieurs jours durant, je pouvais revoir en pensée, avec une précision et une vivacité extrêmes, le choc d'une collision qui n'avait pourtant pas eu lieu. *Aujourd'hui encore, je n'arrive toujours pas à imaginer quoi que ce soit d'autre que ce cri, qui eût pu, dans les circonstances, éviter le pire et prévenir le drame.* Je n'aurais certes pas eu le temps de bondir pour retenir l'enfant, et je ne peux concevoir aucune phrase utile à proférer ni aucune stratégie rationnelle qui auraient pu, dans une situation aussi critique et aussi désespérément urgente, avoir un impact aussi salutaire. Bien sûr, il n'était pas inévitable que cette stratégie réussisse, et

j'aurais tout aussi bien pu crier une fraction de seconde trop tard, avec le résultat que Pieter se pétrifie juste devant le bolide. Il me semble néanmoins évident que, dans cette circonstance, nos structures émotionnelles respectives, *à mon neveu aussi bien qu'à moi,* ont rempli l'une des fonctions essentielles qui ont présidé à leur apparition et à leur sélection à travers l'évolution. Car l'efficacité de mon intervention n'a pu reposer que sur l'heureuse combinaison de ma réaction automatique à proférer un cri d'une telle intensité, mais également de la prédisposition de mon neveu (comme des autres témoins d'ailleurs!) à réagir à un tel cri avec une frayeur suffisante pour freiner sa course.

L'événement illustre comment les émotions peuvent, dans certaines circonstances, mettre à la disposition des individus des stratégies de résolution de problèmes extrêmement puissantes, et souvent plus efficaces pour leur survie que celles rendues ultérieurement possibles par le développement de l'intelligence et de la rationalité. Même si, dans notre société instrumentale et technologique, les réactions émotionnelles sont souvent décriées au profit de la rationalité, même si l'on encourage inconsidérément leur contrôle sinon leur mise au rancart, aussi bien dans la vie professionnelle que dans la vie de l'école, il reste que peu de gens apprécieraient vraiment un individu totalement dénué d'émotions. Bien que nous nous targuions d'être l'espèce douée des facultés intellectuelles les plus complexes et les plus élaborées, le fait est qu'on ne connaît pas non plus d'espèce plus émotionnelle que la nôtre, et que l'évolution de nos fonctions cognitives les plus évoluées semble s'être réalisée *étrangement de pair* avec la complexification croissante des capacités émotionnelles de nos ancêtres mammifères et primates (Damasio, 1994a, 1994b; Kolb et Taylor, 1990; LeDoux, 1989; MacLean, 1993; MacLean et Guyot, 1990).

La nature des relations qui existent entre nos capacités cognitives et nos capacités émotionnelles reste cependant fort ténébreuse et, dans le contexte d'éducation qui nous préoccupe au premier chef, il n'est pas du tout évident que l'intégration éventuelle de ces capacités offre des possibilités tellement prometteuses ou innovatrices pour l'amélioration des stratégies pédagogiques et didactiques. L'anecdote suivante illustre cependant l'impact qu'une meilleure compréhension des interactions entre les domaines de l'émotion et de la cognition pourrait entraîner sur les capacités d'apprentissage des apprenantes et des apprenants.

À la Faculté d'éducation de l'Université de Sherbrooke, j'ai enseigné pendant plusieurs années les rudiments de la programmation en Logo, à une clientèle étudiante se destinant à

l'utilisation pédagogique de l'ordinateur en éducation. En début de session, après avoir expliqué le plan de cours et les exigences d'évaluation, je démarrais chaque fois en regroupant les étudiantes et les étudiants par équipes de deux devant l'ordinateur, et en leur faisant exécuter d'emblée quelques exercices simples à l'aide des primitives de base du langage. Mon idée était d'enseigner ce langage un peu comme une langue naturelle, et d'amener très rapidement mon auditoire à bredouiller quelques "phrases informatiques" dès que les premiers mots ont été acquis. Je déambulais alors dans la salle de classe, commentant les essais et les erreurs de chacun ou chacune, avant de récapituler plus formellement pour l'ensemble du groupe la matière couverte jusque là.

À l'une de ces séances, j'aperçus une équipe qui semblait hésiter devant les exercices à effectuer, et je constatai que la jeune femme du groupe semblait tendue, en état de panique, avec des larmes au coin des yeux. D'un ton que je souhaitais le plus neutre possible, je m'informai comment se déroulaient leurs exercices d'exploration. L'étudiante débita aussitôt qu'elle ne comprenait rien à tout ce charabia, qu'elle n'avait d'ailleurs jamais été douée pour les mathématiques ou les sciences, et qu'elle allait certainement devoir abandonner le cours. Je pensai alors adopter mon ton le plus rassurant, un peu paternaliste même, du genre: "Mais non voyons, ça ira, vous allez voir, c'est plus facile que vous ne le croyez..." Il me vint cependant une autre idée, qui devait s'avérer particulièrement féconde. Je me tournai plutôt vers le jeune homme, et lui demandai si tout allait de son côté, et s'il comprenait bien, pour sa part, les consignes et les exercices. Il se troubla un peu, bredouilla qu'il ne savait pas encore vraiment, qu'il pensait faire d'abord quelques essais, et qu'il verrait bien ensuite. L'étudiante interloquée répliqua: "Comment cela, toi non plus? Ah bon, je ne suis peut-être pas si bête après tout, on va essayer pour voir..."

De ne plus se percevoir comme déficiente, à ses propres yeux comme par rapport au reste du groupe, lui donna un tel regain de confiance qu'elle prit même goût et curiosité au reste de la séance, et y gagna de plus en plus de plaisir - et de compétence - de semaine en semaine. Elle termina finalement la session parmi le peloton de tête, et appliqua même sur un emploi de monitrice pour un camp d'été en informatique offert aux jeunes de la région, emploi qu'elle obtint d'ailleurs sans difficulté. Entre la jeune femme qui se préparait à abandonner carrément un certificat, et celle qui appliquait seulement quelques mois plus tard pour un travail requérant justement des habiletés en programmation informatique, et même une aisance à l'expliquer à d'autres, il y a certes tout un monde!

Or ce ne sont pas tellement des stratégies d'ordre pédagogique ou didactique, au registre du rationnel, qui sont parvenu à susciter un tel "bond mental", mais une toute petite intervention, juste au moment opportun, *au registre de l'émotionnel*. Certes, les réactions d'ordre émotionnel n'ont pas toujours un impact aussi spectaculaire, ni aussi clairement bénéfique. Toutefois, les deux exemples présentés ci-haut révèlent que nos structures émotionnelles renferment une *puissance* d'opération et de transformation des conduites souvent insoupçonnée. Non seulement peuvent-elles, comme dans le premier exemple, contribuer carrément à sauver des vies humaines, mais aussi peuvent-elles, au plan de l'apprentissage, de la formation et des choix de carrière, déverrouiller, comme l'illustre le second exemple, des barrières qui, d'un point de vue strictement cognitif, auraient pu apparaître littéralement insurmontables.

1.2. Le problème

1.2.1. Vers un modèle "computationnel" des émotions

Une utilisation efficace des ressources émotionnelles à des fins d'enseignement et d'apprentissage, et même plus généralement comme outils de résolution de problème, requiert cependant un modèle suffisamment sophistiqué des processus émotionnels, ainsi que de leur articulation avec les processus étudiés ailleurs en sciences cognitives, tels la métacognition, les stratégies de résolution de problème, les différents modes de représentation des connaissances, les processus qui interviennent dans le raisonnement analogique et le transfert des connaissances... Les concepts théoriques élaborés dans ces divers domaines de recherche ont en effet déjà permis une explicitation satisfaisante (et utilisable) des mécanismes mentaux qui interviennent dans l'enseignement et l'apprentissage. Ils se sont alors traduits, notamment dans le domaine de *l'enseignement stratégique* (Tardif, 1992, 1995), par un découpage et un réaménagement plus adéquats du curriculum scolaire, et par la conception de procédés plus efficaces au profit de l'éducation et de la pratique pédagogique.

La contribution des travaux poursuivis en Intelligence Artificielle (IA) s'est avérée radicale et décisive à cet égard. Il n'est pas exagéré en effet d'affirmer que cette discipline a joué un rôle de catalyseur au sein des sciences cognitives (Gardner, 1985). Cela s'est produit

parce qu'elle requérait d'office, pour la résolution des problèmes complexes abordés (par exemple l'automatisation du traitement de la langue naturelle), l'interaction étroite de plusieurs champs disciplinaires autrefois étanches. Une autre raison importante consiste cependant dans le fait qu'elle offrait, par l'intermédiaire des langages informatiques utilisés, une solide base de formalisation des phénomènes de pensée et de comportement, dont la rigueur n'avait désormais rien à envier aux modélisations qui avaient contribué à donner leurs lettres de noblesse aux sciences physiques (Simon, 1995).

Le succès de l'approche de l'IA repose en bonne partie sur l'attitude épistémologique à la fois rationaliste et constructiviste que cette discipline suscite. L'idée fondamentale en est d'étudier les phénomènes mentaux, non pas tellement par une observation empirique des comportements, mais en essayant plutôt de les simuler et de les reproduire. On y retrouve en fait l'esprit de la perspective piagétienne, selon laquelle l'intelligence se développe par une reconstruction active du monde extérieur, effectuée par le sujet, qui conceptualise le réel en agençant progressivement les représentations mentales des transformations qu'il a lui-même effectuées sur son environnement (Piaget, 1936, 1947). La démarche de l'IA constitue essentiellement une approche d'ingénierie. Seymour Papert (1973) rappelait, dans une de ses conférences, que les principes de l'aéronautique avaient été élaborés, moins en observant systématiquement le vol des oiseaux, qu'en essayant activement de le reproduire, à l'aide de dispositifs comme des planeurs ou des avions, capables de réaliser une performance comparable. D'une façon analogue, l'IA essaie de spécifier, en essayant de les reproduire, les divers processus sous-jacents aux comportements intelligents.

À titre d'exemple, même pour un problème aussi trivial que le jeu de tic-tac-toe, un programmeur novice devra, au préalable, réfléchir à la façon dont lui-même jouerait ses coups. Il devra représenter soigneusement chacune de ses stratégies, ainsi que la façon dont il sait reconnaître et représenter chaque situation de jeu. La première version d'un tel programme sera généralement battue à plates coutures par le premier joueur humain venu. C'est qu'à moins de disposer déjà d'une analyse sophistiquée du jeu, le cogniticien en herbe aura vraisemblablement oublié de spécifier certaines conditions ou stratégies plus subtiles, qui sont moins facilement identifiables à l'introspection, ce qui relancera la programmation vers une version plus performante. Ce va-et-vient extrêmement fécond opère une espèce de grossissement, un peu à la façon d'un microscope projeté sur les structures mentales, facilitant l'identification et l'explicitation des processus requis pour l'accomplissement d'une tâche donnée. Par les ratés d'une simulation ou par la singularité

de certains comportements recréés, il met en évidence les spécifications imprécises, il souligne les carences, il retourne contre elles-mêmes les théorisations vagues, incomplètes ou inadéquates.

Diverses conceptualisations fondamentales, aussi bien pour le développement de la cognition que pour l'éducation, sont issues de l'application de cette métaphore établie entre le cerveau et l'ordinateur. C'est par exemple le cas des conceptions actuelles de la mémoire, en termes de mémoire sensorielle, de mémoire de travail à capacité limitée, de mémoire à long terme, conçue comme un vaste réseau sémantique, où la recherche d'information s'effectue par propagation d'activation (Anderson et Bower, 1973). C'est également le cas dans la caractérisation des différents modes de représentation, en termes de connaissances déclaratives, procédurales ou conditionnelles, ou sous le format plus structuré des "frames", des scripts ou des schémas (Anderson, 1980). Autant d'outils théoriques qui se prêtent désormais à une exploitation pragmatique dans le contexte pédagogique de l'école et des autres milieux de formation (Gagné, Yekovitch et Yekovitch, 1993; Paquette, Aubin et Crevier, 1994; Tardif, 1992, 1995).

Pour diverses raisons, il nous a semblé qu'une perspective analogue pourrait s'avérer également féconde afin d'opérationaliser la myriade des concepts qui recouvrent les phénomènes affectifs. Tout d'abord, les processus émotionnels, ayant leur siège dans les structures du système limbique et des lobes frontaux et temporaux (Davidson, 1992; Kolb et Taylor, 1990; LeDoux, 1987, 1989, 1993; MacLean, 1993; MacLean et Guyot, 1990), appartiennent également au domaine mental, et leurs interactions avec les processus cognitifs, tels qu'en témoignent les exemples présentés dans la section précédente, suggèrent qu'on se retrouve dans un ordre de complexité comparable. Deuxièmement, l'émotion constitue, tout comme l'intelligence, un construit hypothétique recouvrant des phénomènes dont les effets ne sont perceptibles qu'indirectement (Kirouac, 1989, 1993). Or, l'approche de simulation et d'ingénierie utilisée en IA a déjà fait ses preuves pour l'explicitation des processus mentaux, en abordant des phénomènes apparemment aussi peu descriptibles que l'intuition, l'inspiration, la perspicacité, le processus de découverte (Kaplan et Simon, 1990; Kulkarni et Simon, 1988; Simon, 1995). En troisième lieu, l'évolution des recherches dans le domaine de l'IA rend de plus en plus envisageable, *en même temps que théoriquement avantageuse,* l'intégration des dimensions affectives dans la modélisation des processus mentaux. C'est notamment le cas à travers la préoccupation toujours centrale d'arriver à représenter le sens commun (McCarthy, 1990; Minsky,

1985), à travers le développement de la robotique et de la vie artificielle (Brooks, 1991b; Maes, 1991a; Steels et Brooks, 1995), et à travers les recherches sur les systèmes multi-agents en intelligence artificielle distribuée (Bond et Gasser, 1988; Demazeau et Müller, 1990, 1991; Gasser et Huhns, 1989; Huhns, 1987).

Plusieurs auteurs en sciences cognitives et en IA on en effet souligné l'importance, voire la nécessité, pour arriver à capter dans un système artificiel les habiletés de raisonnement qui correspondent au simple "bon sens", de le doter de mécanismes d'ordre motivationnel ou émotionnel (Minsky, 1985; Norman, 1980; Simon, 1987; Sloman, 1987, 1992, 1995; Sloman et Croucher, 1981). L'un des concepts clés à cet égard est celui *d'intentionnalité*. Intimement reliée à la génération et à la gestion d'une structure de buts, cette notion apparaît effectivement, dans la plupart des théories cognitives des émotions, comme l'une des caractéristiques essentielles des structures émotionnelles (Frijda, 1986; Johnson-Laird, 1988; Oatley et Johnson-Laird, 1987; Oatley, 1992; Ortony, Clore et Collins, 1988; Sloman, 1987).

Les progrès de la robotique, pour leur part, ont en quelque sorte contribué à "donner un corps" aux systèmes artificiels, en assurant un câblage efficace entre les capteurs sensoriels et les dispositifs de locomotion. Pour certains auteurs (Brooks, 1991a, 1991b), la compréhension et la maîtrise de la sensori-motricité apparaissent même comme un prérequis incontournable à la conceptualisation et à la réalisation de systèmes artificiels véritablement intelligents, une vision qui n'est pas sans rappeler la position de Piaget sur la genèse de l'intelligence à partir des tout premiers schèmes réflexes (Piaget, 1936, 1947, 1967). Cet objectif a même entraîné l'apparition d'un nouveau domaine de recherche au sein de l'IA, la vie artificielle (Steels et Brooks, 1995). Le chercheur y est cependant vite conduit à questionner s'il suffit à un organisme, vivant ou artificiel, de disposer de mécanismes de perception et de locomotion, pour être automatiquement doué d'autonomie et adapté à son environnement (McFarland, 1995; Steels, 1995). La question de la sensori-motricité renvoie en effet à celles de la motivation, des besoins et des émotions: pour un organisme donné, qu'est-il *nécessaire et vital* de chercher à percevoir et à reconnaître (par exemple, des sources de nourriture, un parent ou un petit, un prédateur potentiel, etc.), et vers où faut-il *choisir* de se déplacer? C'est ce qui explique que ces préoccupations prennent une place de plus en plus importante dans ce domaine de recherche (Maes, Mataric, Meyer, Pollack, et Wilson, 1996). Réciproquement, l'importance que revêt le corps physique dans la phénoménologie et dans l'expression des

émotions, *et probablement jusque dans leur représentation* (Damasio, 1994a; voir en particulier le chapitre 8: "The somatic marker hypothesis") devient désormais accessible à la modélisation, notamment par le biais des diverses matérialisations ("embodiments") qui constituent les objets essentiels de la robotique.

Finalement, l'apparition relativement récente d'ordinateurs à calcul parallèle a entraîné une autre reconceptualisation significative dans la modélisation des processus mentaux. L'une de ces transformations concerne l'émergence de l'intelligence artificielle distribuée (IAD). Apparue il y a une quinzaine d'années à peine, cette nouvelle discipline relève d'une "métaphore" à caractère plutôt *sociologique* (Gasser, 1991; Minsky, 1985) que *psychologique*. Alors que l'IA "classique" s'efforçait d'élaborer des systèmes capables de représenter des connaissances et de résoudre des problèmes, en prenant essentiellement comme modèle l'individu, le paradigme de l'IAD repose plutôt sur le constat que les problèmes complexes du monde réel sont généralement abordés et solutionnés, non par un individu qui y travaille isolément, *mais par une "équipe de spécialistes" qui collaborent entre eux à leur résolution* (Kornfield et Hewitt, 1981; Lenat, 1975). L'intelligence artificielle distribuée cherche donc à incorporer en IA cette dimension particulièrement efficace du fonctionnement humain, en *distribuant* une tâche donnée en sous-processus autonomes (ou agents) qui opéreront en parallèle, contribuant collectivement à la résolution du problème initial. Outre la question d'efficacité, il est intéressant de souligner que cette approche est également porteuse d'une réflexion épistémologique sur la nature de la connaissance envisagée comme *transaction entre sujets connaissants*, une vision qui n'est sans doute pas si éloignée de celles qu'avaient par exemple Mead (1934), Vygotsky (1978) ou Bandura (1986) à propos de l'apprentissage.

Or, bien que l'idée d'une communauté de processeurs agissant de concert dans la résolution d'une tâche complexe s'avère une métaphore séduisante, elle n'en soulève pas moins son propre contingent de problèmes, notamment celui d'assurer la coordination de ce "travail d'équipe" et de déterminer les modes de communication et de négociation entre les différents partenaires. Dans une revue générale des différentes recherches réalisées en IAD, Bond et Gasser (1988, p. 10) identifient les problèmes de fond qui confrontent les chercheurs dans ce domaine:

- comment formuler et décomposer les problèmes pour en assurer une distribution efficace entre les différents agents?
- quel protocole de communication adopter; quoi communiquer, et quand le faire?

- comment assurer la cohérence dans le comportement des agents, comment réduire le risque d'effets globaux, possiblement néfastes, résultant de décisions prises sur une base locale?
- comment doter les agents de la capacité de se représenter adéquatement les connaissances et les comportements des autres agents, de façon à assurer une bonne coordination de leurs actions respectives?
- comment concilier les points de vue divergents et résoudre les conflits qui surgissent au sein d'une communauté d'agents cherchant à coordonner leurs efforts?

Quelle est donc la caractéristique commune à ces différents problèmes? Elle est en tout cas suffisamment abstraite, car elle ne dépend d'aucune situation particulière ou spécifique. En fait, ce n'est pas tellement du *contenu* de l'information qu'il est question ici, mais de *contrôle*, au sens informatique du terme: généralement, il s'agit en effet de déterminer, par les mécanismes les plus simples et les plus économiques possibles, comment circulera l'information. Il n'est cependant pas trop surprenant de constater que les difficultés qui constituent les entraves les plus sérieuses à la conception et à l'élaboration de systèmes d'IAD concernent justement la communication, le contrôle et la répartition de l'information entre les agents. Car, d'aussi loin que l'on remonte, on peut considérer que l'une des lignes directrices de l'évolution de la science informatique en général, et des langages d'intelligence artificielle en particulier, a précisément consisté dans la recherche et la spécification de *structures de contrôles* adéquates assurant la gestion et la coordination des diverses tâches que les programmes ont pour fonction de résoudre (Winston, 1984; voir en particulier le chapitre 5: "Control metaphors").

En face de telles difficultés, la stratégie générale de l'IA, comme on l'a vu plus haut, consiste à rechercher une métaphore, le plus souvent à caractère *anthropomorphique*, qui permette de générer des "insights" suffisamment éclairants pour contourner l'impasse. On se regarde agir, on réfléchit à la façon dont les humains résolvent des problèmes d'une nature ou d'une complexité analogue, on essaie de caractériser la solution un peu plus formellement, on l'implante, puis on la révise, en prenant là encore le comportement humain comme référence. En rapport avec les problèmes identifiés plus haut pour l'IAD, il est donc pertinent de se demander quels sont, chez les humains (ou plus généralement chez les animaux), les processus psychiques dont la fonction semble se rapprocher le plus de celle qui est justement remplie par les structures de contrôle dans les systèmes informatiques.

Or, en psychologie, c'est généralement aux processus motivationnels que l'on attribue cette propriété (Cofer et Appley, 1964), c'est-à-dire à *l'ensemble des facteurs et des dispositions qui poussent un organisme à agir et qui déterminent chez lui ses préférences et son choix entre divers comportements possibles.* Lindsay et Norman (1972; voir en particulier le chapitre 17: "Motivation"), dans leur introduction à la psychologie en termes de traitement de l'information, abordent d'ailleurs explicitement les systèmes motivationnels comme des *systèmes de contrôle* du comportement. Il y a donc de bonnes raisons de croire que les structures émotionnelles pourraient offrir, au moins à titre de métaphore, un modèle à partir duquel conceptualiser les spécifications des structures de contrôle requises en IAD. Réciproquement, il est bien possible que les réflexions déjà accumulées en informatique sur les structures de contrôle et sur les systèmes d'opération puissent contribuer à éclairer en retour les mécanismes qui opèrent chez les organismes vivants, au chapitre de la motivation et des émotions.

Bref, que l'on regarde du côté de la représentation de l'intentionnalité et du sens commun, que l'on se préoccupe de l'amélioration des stratégies adaptatives inscrites dans les automates réalisés dans les domaines de la robotique ou de la vie artificielle, ou que l'on cherche à formaliser des structures de contrôle adéquates pour assurer la coordination, la communication et la coopération dans les systèmes multi-agents, on voit chaque fois surgir, dans chacun de ces domaines de pointe de l'IA, l'importance d'intégrer à l'approche une réflexion approfondie des mécanismes affectifs et motivationnels. D'une part, l'approche d'ingénierie de l'IA semble offrir des possibilités intéressantes pour une conceptualisation et une opérationalisation renouvelées des processus émotionnels. D'autre part, une meilleure compréhension des mécanismes sous-jacents au déclenchement et à l'expression des émotions pourrait bien, en retour, contribuer à déverrouiller certains problèmes de fond concernant les structures de contrôle, l'autonomie et l'adaptabilité dans les systèmes multi-agents.

1.2.2. Quelques distinctions: affect, motivation, émotion

Nous avons jusqu'ici utilisé de façon quasi interchangeable les trois termes d'affectif, de motivationnel et d'émotionnel. Bien que ces termes entretiennent effectivement des relations très étroites, il importe d'établir au préalable quelques nuances. Le premier, celui d'affect, est probablement le plus général: "On désigne habituellement par processus affectifs tous les états impliquant des sensations de plaisir-déplaisir ou encore liés au

registre agréable-désagréable" (Kirouac, 1993, p. 44). C'est donc dire que ce vocable peut qualifier assez généralement une foule d'états tant physiologiques que psychiques, incluant aussi bien le plaisir, la douleur et les impressions sensorielles, que les attitudes, les émotions, ou les sentiments.

Le terme de motivation constitue également un concept général, même s'il a été parfois utilisé, dans la littérature, dans un sens plus ou moins équivalent à celui d'émotions. Plusieurs auteurs se sont attachés toutefois, avec des succès mitigés, à distinguer les deux termes. Ainsi, une distinction qui a prévalu jusqu'à la fin des années 1940 (Young, 1943, 1949), attribuait un caractère nettement *dysfonctionnel* et perturbateur aux émotions, alors qu'était plutôt associé au concept de motivation un caractère *adaptatif* et organisateur des conduites. Cette conception sera cependant vigoureusement remise en question par Leeper (1948; voir aussi Duffy, 1948; et la réplique de Young, 1949), qui proposera une vision plus adaptative des émotions, vues comme une catégorie particulière de motivations apparues chez les animaux plus évolués. De plus en plus d'auteurs s'entendent désormais (Toates, 1986, 1988; Plutchik, 1980, 1984) pour considérer la motivation comme le terme générique, tout en préservant son aspect fonctionnel dans la détermination des buts ("goal-orientedness") de l'organisme.

Affect et motivation constituent donc deux termes englobants: dans le premier, l'accent est placé sur la détermination du caractère agréable ou désagréable d'un objet ou d'un événement, alors que dans le second, l'accent est placé plutôt sur ce qui détermine les buts et oriente les conduites. Le concept de motivation est ainsi défini par Valleyrand et Thill (1993) comme "le construit hypothétique utilisé afin de décrire les forces internes et/ou externes produisant le déclenchement, la direction, l'intensité et la persistance du comportement" (p. 18), une définition qui rejoint d'assez près celle de Cofer et Appley (1964) présentée dans la section précédente. Ainsi entendu, cette notion recouvre aussi bien divers états physiologiques comme la faim, la soif ou le désir sexuel, que des états psychiques tels la peur, l'anxiété, la colère, la joie ou l'attachement parental. Le terme de *besoins* sera utilisé pour désigner la première de ces sous-catégories, et celui *d'émotions* sera plutôt réservé à la seconde, la distinction entre les deux types de comportements reposant sur les critères suivants (Plutchik, 1980, pp. 362-363; 1984, p. 214):

- les émotions sont habituellement déclenchées par des stimulations extérieures, alors que les besoins résultent de changements internes dans certains états physiologiques;

- les émotions surgissent en présence d'objets ou d'événements déclencheurs, alors que les besoins se manifestent plutôt en l'absence de certaines stimulations à caractère homéostatique;
- les émotions ne dépendent pas d'objets spécifiques (une variété de situations peuvent susciter la peur), alors que les besoins dépendent d'objets particuliers, comme l'eau ou la nourriture;
- les émotions sont consécutives à la perception et à l'évaluation d'un objet ou d'un événement, alors que les besoins précèdent au contraire la recherche d'un objet nécessaire pour leur satisfaction;
- les émotions dépendent d'événements qui peuvent se produire sur une base aléatoire, alors que les besoins présentent plutôt un caractère rythmique.

Les émotions constituent donc un type d'état particulier. Plus précisément, selon un consensus relatif qui semble se dégager parmi plusieurs théoriciens actuels (Frijda, 1986; Izard, 1993; Kirouac, 1989, 1993; Oatley, 1992; Plutchik, 1984), elles peuvent être définies comme *un état mental hypothétique, résultant de l'évaluation d'un objet ou d'un événement significatif pour l'organisme, et qui se traduit alors par un ensemble complexe de réactions*. Ces diverses manifestations incluent: une "coloration" phénoménologique particulière, une expression typique (faciale, vocale, posturale) ayant valeur de signal, une réaction du système nerveux autonome, ainsi que l'activation d'un répertoire spécifique d'actions.

Il s'agit en effet d'un construit hypothétique, puisqu'une partie seulement du processus émotionnel est observable, et que la plupart des événements mentaux sous-jacents doivent être inférés. L'état psychique caractérisant la réaction émotionnelle semble tout d'abord requérir une évaluation ("appraisal") de l'événement déclencheur, sans lequel celui-ci ne peut prendre de signification pour l'organisme: un animal doit par exemple être perçu comme menaçant ou dangereux pour pouvoir évoquer la peur. Il existe une relative unanimité dans la littérature à l'effet que certaines "classes de situations" entraînent universellement des réactions émotionnelles analogues (Scherer et al., 1986; Lutz, 1988): la tristesse semble ainsi partout reliée à une perte (d'un être cher ou d'un objet valorisé); la joie est fréquemment associée à la présence d'êtres proches ou à la réception de ressources; la colère est souvent évoquée quand un obstacle entrave la réalisation d'un but convoité, etc. Il est important de préciser cependant que ces évaluations ne sont pas toujours nécessairement conscientes, et que la plupart des organismes semblent même disposer de patrons innés de réactions à certaines situations typiques. L'organisme doit également entretenir déjà certaines dispositions potentielles favorables ou défavorables

("concerns") à l'endroit de l'événement, pour que l'avènement de celui-ci puisse produire un effet significatif.

Chaque émotion se caractérise en outre par une façon particulière de ressentir: être triste est phénoménologiquement très différent d'avoir peur ou d'être en colère, et il serait extrêmement difficile (voire impossible prétendent Johnson-Laird et Oatley, 1989, p. 90) de communiquer la "qualité" de cette expérience à quiconque ne l'a jamais éprouvée. Une émotion se traduit aussi par une configuration de mimiques, de gestes, de tons de voix, de postures, qui prennent valeur de signal et servent à communiquer avec les organismes avoisinants (pas nécessairement de la même espèce). Cette caractéristique a été largement documentée dans le cas de l'expression faciale, dans une grande variété de cultures (Ekman, 1973; Ekman, Friesen et Ellsworth, 1982b), chez de très jeunes enfants, même aveugles de naissance (Charlesworth et Kreutzer, 1973; Ekman et Oster, 1982), et jusque chez certains primates (Chevalier-Skolnikoff, 1973; Redican, 1982). Les réactions émotionnelles sont également accompagnées de perturbations significatives au niveau du système nerveux autonome: accélération ou décélération des rythmes cardiaque ou respiratoire, modification de la transpiration, vasoconstriction ou vasodilatation, etc. Finalement, le déclenchement d'une émotion active automatiquement un répertoire d'actions généralement en rapport avec la situation évocatrice: ainsi la peur rendra disponible une variété de réactions comme de crier, de fuir, de se figer, d'attaquer, etc.

Une autre caractéristique intriguante des structures émotionnelles est qu'elles semblent le plus souvent intervenir dans le contexte de relations interpersonnelles, ou du moins entre individus, si l'on parle plutôt d'espèces animales autres qu'humaines. Bien que ce soit parfois moins évident dans le cas de la peur (Averill, 1980; Wallbott and Scherer, 1986; Scherer and Tannenbaum, 1986), dans la plupart des autres réactions émotionnelles cependant, la dimension sociale et interactive apparaît nettement déterminante. En effet, rien ne déclenche plus efficacement la joie que le rassemblement de parents ou d'amis, ou l'échange de ressources entre individus (Scherer et al., 1986; Lutz, 1988). Rien ne déclenche la colère comme la transgression de normes ou de conventions, surtout si elles sont le fait d'individus auxquels nous sommes intimement reliés (Averill, 1979, 1983). Les comportements de détresse ou de tristesse n'existent apparemment pas chez les espèces qui ne s'occupent pas de leurs petits, et qui ne manifestent pas de comportements maternels, de coopération ou de protection (MacLean, 1993; MacLean et Guyot, 1990). La culpabilité ou la honte requièrent par ailleurs, dans leur définition même, le regard d'un

autre, en particulier d'un proche, que notre propre comportement risque de décevoir, et par rapport auquel notre image personnelle risque d'être entachée ou fragilisée (Baumeister, Stillwell et Heatherton, 1994). Même dans le cas de la peur, l'existence même de signaux d'alarme suppose une certaine interaction sociale relative à la protection mutuelle, sans quoi ces signaux serviraient surtout à informer les prédateurs sur la localisation des proies, un comportement qui s'avérerait alors assez peu adaptatif.

Paul Ekman (1984, 1992b), qui s'est consacré principalement à l'étude de l'expression faciale des émotions, propose de rajouter trois composantes à celles déjà énumérées. Tout d'abord, les réactions émotionnelles sont susceptibles d'un déclenchement extrêmement rapide, en fait de l'ordre de quelques millisecondes, si bien que la conscience même de leur déclenchement nous échappera souvent. Elles sont ensuite d'une durée relativement courte, de l'ordre de quelques secondes seulement. D'après Ekman, cette caractéristique permet, selon la persistance de l'état qui est subi, de distinguer les émotions d'autres états affectifs comme les *sentiments*, les *humeurs* ("moods"), les *traits* de personnalité et jusqu'aux *désordres* émotionnels. C'est donc cette variable qui distinguerait, selon que la durée est plus ou moins longue, que l'on éprouve simplement de la tristesse, ou que l'on est affligé par un deuil, que l'on se sent d'humeur maussade, que l'on est de tempérament pessimiste, ou que l'on souffre carrément de neurasthénie. Ekman rajoute finalement le caractère involontaire des réactions émotionnelles, qui semblent s'imposer d'elles-mêmes, et opérer en dépit de notre volonté. Or ces trois caractéristiques confortent assez bien l'idée, que nous avions formulée plus haut, d'exploiter les processus émotionnels comme prototypes de structures de contrôle pour les systèmes multi-agents. Combinées à l'idée de l'évaluation d'événements qui puisse s'effectuer sur une base sommaire et automatique, à l'idée de signaux de communication universels entre les différents agents et à l'idée d'un répertoire préprogrammé de réactions spécifiques aux situations d'urgence, ces caractéristiques suggèrent déjà quelques spécifications fonctionnelles pour le design de structures de contrôle en IAD.

1.2.3. Les émotions comme structures de contrôle

Cette idée d'envisager les processus émotionnels comme structure de contrôle n'est pourtant pas neuve. Il y a une trentaine d'années déjà, Herbert Simon (1967) postulait qu'un organisme adaptatif devait nécessairement disposer de *systèmes d'interruption* permettant de rediriger l'attention et le traitement vers les situations d'urgence qui

pouvaient toujours se manifester dans un environnement ouvert. Dans le contexte d'un processeur essentiellement sériel, cela supposait qu'une partie du temps de traitement soit régulièrement affecté à la surveillance de certains signaux d'urgence. Les réflexions de Simon s'avéraient peut-être trop innovatrices pour l'époque, puisqu'elles ne donnèrent suite à aucun développement significatif avant le début des années 1980. Depuis lors, cependant, de plus en plus de théoriciens et de chercheurs (Minsky, 1985; Oatley, 1992; Oatley et Johnson-Laird, 1987, 1996; Sloman, 1987; Sloman et Croucher, 1981; Frijda, 1986; Frijda et Swagerman, 1987) vont reprendre ce filon pour tenter d'articuler les relations incontournables qui existent entre émotion et cognition.

Minsky (1985; voir section 3.6, "Pain and pleasure simplified", p. 37), par exemple, attribue aux émotions une fonction de gestion par le fait qu'elles imposent une simplification radicale dans la complexité des choix auxquels se retrouve habituellement confronté l'organisme. Ainsi, l'animal qui a peur n'a plus à choisir entre l'objectif de se nourrir, de jouer ou d'explorer son environnement: il est concentré sur le danger, et toutes ses ressources, aussi bien physiologiques que cognitives sont mobilisées à cette fin. Une autre contribution intéressante de cet auteur à la compréhension des rapports entre émotion et cognition concerne ce qu'il appelle "l'apprentissage-par-attachement". Minsky (1985; voir section 17.2, "Attachment-learning", p. 175) postule que la tâche de se construire une identité et de formuler les valeurs et les idéaux qui structureront ses comportements constitue une tâche beaucoup trop complexe pour un individu seul. La solution qui aurait été élaborée progressivement à travers l'évolution consisterait dans une espèce de "verrouillage" plus ou moins inné attachant l'enfant ou le jeune animal (on parlera alors plutôt d'*imprinting*) à un petit nombre d'adultes spécifiquement déterminés dont il emprunterait de vastes fragments de comportements tout faits et déjà pratiquement opérationnels dans l'environnement physique - et surtout *social* - où ce nouvel individu aurait à évoluer.

Certaines recherches récentes en neurophysiologie (Rolls, 1990) ainsi qu'en éthologie humaine (Trevarthen, 1984; voir aussi à ce sujet: Izard, 1984; Harris, 1989), suggèrent en effet que l'enfant est déjà équipé, dès la naissance, pour ce "verrouillage", par sa capacité de reconnaître la forme des visages et les sons de la voix humaine. En outre, les réactions émotionnelles de joie ou de détresse, qui se traduisent par les rires, les sourires, les cris et les pleurs, ainsi que par des expressions faciales typiques, assureraient justement le démarrage de la communication entre la mère et son enfant. Bien sûr, au début, il ne

semble pas que l'enfant interprète grand chose des interventions de sa mère. Mais il y réagit et, comme on peut le percevoir à partir d'enregistrements magnétoscopiques, c'est *elle* qui donne du sens aux réactions de son enfant, en les "ponctuant" de ses propres réactions émotionnelles. Ainsi, progressivement, l'enfant semble *apprendre le sens de ses propres interventions* à partir de la façon dont sa mère elle-même les interprète. Pour un observateur extérieur, les premières réactions de l'enfant peuvent sembler déjà chargées de sens. *Mais c'est l'interaction émotionnelle de la mère avec son enfant qui confère en réalité cette signification,* et la qualité de cette interaction sera vraisemblablement déterminante pour le développement ultérieur de l'intelligence, du langage et des outils de communication.

Dans une telle perspective, les structures émotionnelles constituent une assise fondamentale sur laquelle élaborer par la suite des structures cognitives plus complexes. Oatley et Johnson-Laird (1987, 1996; voir aussi Oatley, 1992) proposent même que les émotions jouent un rôle essentiel dans la planification "rationnelle" d'un organisme et la poursuite des buts qu'il se fixe. Ces chercheurs ont en effet élaboré une théorie cognitive selon laquelle les émotions surgissent lorsque l'évaluation de la réussite d'un plan poursuivi par un individu, est modifiée par un événement quelconque. Selon eux, les émotions apparaissent typiquement *lorsqu'un organisme est confronté à la réalisation d'objectifs multiples, dans un environnement complexe et difficilement prévisible, où l'information et les ressources sont rares et incomplètes, et où l'interaction (coopération ou conflit) entre plusieurs organismes apparaît inévitable* (Oatley, 1992, p. 31-36). Les émotions se seraient justement développées, à travers l'évolution, pour permettre à l'organisme de détecter les situations où surgissent des événements qui imposent des modifications soudaines dans la structure des buts. Elles assureraient une prise en charge adaptative de ces situations problématiques en émettant certains signaux spécifiques dont la fonction serait d'activer un répertoire de réponses stéréotypées, mais le plus souvent appropriées.

Un aspect intéressant de cette théorie, c'est que *les mêmes mécanismes de base* servent aussi bien pour la gestion *interne* des comportements d'un individu, que pour la coordination *externe* des membres d'un groupe: une réaction émotionnelle de peur aura ainsi pour effet de mobiliser et de contrôler différents sous-processus (sensoriels, moteurs, affectifs, cognitifs) au sein d'un même individu, mais son expression alertera en même temps les autres membres de la communauté, dont le comportement en subira alors

également les effets en termes de coordination et d'interaction. L'exemple présenté au tout début illustre d'ailleurs de façon remarquable comment les structures émotionnelles peuvent agir comme structure de contrôle dans un système multi-agents en permettant d'interrompre ou de modifier drastiquement la séquence d'exécution des comportements internes déjà enclenchés chez l'un des agents, à partir d'un signal émis par un autre agent, opérant lui-même sur ses propres buts: le cri exprimant ma terreur a ainsi pu réussir à modifier la course effrénée de Pieter. Il est essentiel de comprendre ici qu'un signal qui n'aurait pas disposé de ce *statut* de structure de contrôle, avec une capacité d'interruption et une priorité extrêmement élevée, n'aurait jamais pu produire un tel effet.

En bref, la recherche menée dans le cadre de la présente thèse vise à expliciter la fonction adaptative des structures émotionnelles en exploitant la métaphore cerveau-ordinateur. À l'instar des recherches d'intelligence artificielle destinées à élucider le fonctionnement des processus cognitifs, elle tentera de proposer un modèle plausible de l'articulation et du fonctionnement des structures émotionnelles. Nous chercherons donc à déterminer certaines des contraintes ("requirements") qui auraient pu, dans le contexte de l'évolution, *requérir*, comme solution adaptative, l'opération de mécanismes correspondant aux émotions. Nous tâcherons ensuite de formuler un certain nombre de spécifications permettant éventuellement l'implantation de tels mécanismes sous la forme de structures de contrôle efficaces pour la coordination et la gestion de systèmes multi-agents, peuplés d'agents informatiques doués chacun d'une relative autonomie.

1.3. Considérations méthodologiques

Il s'agit donc évidemment d'une recherche de type théorique - ou "spéculative", selon la terminologie de Van der Maren (1993). Bien que cet auteur distingue trois formes de l'analyse spéculative - conceptuelle, critique ou inférentielle - notre recherche devra forcément puiser à chacune d'entre elles. L'analyse conceptuelle "a pour objectif de dégager le sens et les possibilités d'application d'un concept" (p. 6.4). L'analyse critique "a pour fin d'évaluer un ensemble d'énoncés théoriques afin de mettre en évidence ses lacunes, ses contradictions, ses paradoxes, ses conditions, ses présupposés, ses implications et ses conséquences" (p. 6.8). Finalement, l'analyse inférentielle "a pour objectif le développement ou l'extension de théories. [...] Le développement d'une

nouvelle théorie dans un domaine donné s'obtient par le transfert d'une théorie d'un autre domaine à la suite de la perception d'une analogie entre les domaines" (p. 6.10).

Selon cette typologie, l'approche que nous avons retenue relève surtout de l'analyse *inférentielle*, puisque nous tenterons d'appliquer au domaine de l'IAD une analyse des processus émotionnels issue des sciences cognitives, et d'appliquer en retour, pour la conceptualisation des structures émotionnelles, des notions empruntées à la science informatique. Plusieurs des concepts ainsi utilisés proviennent surtout du contexte des langages programmation à base d'acteurs (ou d'objets concurrents), concernant les mécanismes d'interruption, le contrôle de la priorité des processus et la prise en charge des exceptions (Agha, 1986; Briot, 1994; Dony, 1990). D'autres notions concernent la gestion des problèmes de communication, de coopération et d'organisation rencontrés dans les systèmes multi-agents (Bond et Gasser, 1988; Bouron, 1992). Cependant, nous devrons aussi mener une analyse *conceptuelle* sur quelques concepts clés utilisés pour notre modèle, comme ceux d'émotions de base ("basic emotions"), d'appropriation ("appraisal"), ou d'engagement ("commitment"), et nous devrons par ailleurs effectuer une analyse *critique* de certaines théories fondamentales en psychologie des émotions ("Appraisal theories" vs "Basic emotions theories") auxquelles le modèle proposé visera à apporter certaines nuances et certains correctifs.

Dans cette section, nous exposerons donc, dans un premier temps, comment nous avons choisi de procéder pour la sélection des concepts et pour la construction du modèle, afin d'assurer à notre démarche toute la rigueur requise. Nous aborderons ensuite la question de la validation d'un modèle "computationnel" comme celui que nous proposons ici.

1.3.1. Le choix des concepts et la construction du modèle

Nous avons mentionné déjà notre intention d'exploiter l'approche constructiviste de l'IA afin d'expliciter les mécanismes sous-jacents aux comportements émotionnels. Nous nous sommes alors demandé à quoi ces mécanismes pourraient bien correspondre dans le contexte d'un système complexe de traitement de l'information tel qu'en produit la science informatique. L'hypothèse fonctionnaliste à laquelle nous souscrivons est que l'évolution, envisagée comme une espèce de *"designer"* aveugle (Dawkins, 1986), mais patient et sélectif, n'a vraisemblablement dû laisser persister à long terme que des solutions adaptatives. Or pour être efficaces, celles-ci doivent nécessairement prendre en compte les

limites de traitement (notamment en rapidité et en capacité de stockage) inhérentes à n'importe quelle machine physique, qu'elle soit mécanique, électronique... ou biologique.

L'adoption d'une telle hypothèse éclaire en fait le sens que nous accordons à une expression comme "émotion artificielle". Car ce n'est pas à l'aspect phénoménologique des émotions, au fait qu'elles suscitent telle ou telle façon de ressentir, que nous nous attacherons ici. Le biologiste ou le médecin qui s'intéressent à la douleur, ne sont pas, eux non plus, *d'abord* intéressés à la phénoménologie de la souffrance ou de la détresse qui en résulte (même si cela peut également les préoccuper), mais plutôt au dysfonctionnement interne qu'un tel signal révèle, et aux mécanismes sous-jacents par lesquels la douleur sert ainsi à préserver et à protéger l'organisme blessé. Le psychologue de la motivation qui étudie les besoins n'est pas non plus tellement préoccupé par la qualité du plaisir ou du déplaisir ressenti par l'animal qui a faim ou soif, mais plutôt par la nature et le détail des dispositifs homéostatiques qui interviennent dans la régulation de ces fonctions.

D'une façon analogue, lorsque l'on parle de coeur artificiel, de rein artificiel, ou de membre artificiel, ce n'est pas tellement de l'apparence de ces organes qu'il est question, du fait qu'ils soient en métal ou en plastique, ou qu'ils opèrent de façon mécanique ou électronique, mais plutôt des *fonctions* vitales qu'ils remplissent, que l'on essaie alors de capter par un dispositif physique quelconque, généralement à des fins de survie. C'est probablement le sens le plus fécond qu'il faille également accorder au terme d'intelligence artificielle (Bobrow et Hayes, 1985; Guha et Lenat, 1990; Minsky, 1985), et c'est le sens que nous accordons pour notre part à l'entreprise d'élucider le fonctionnement des émotions en adoptant l'approche, les concepts et les instruments développés en IA.

En cherchant ainsi comment il serait possible de reproduire de semblables fonctions, nous tâcherons donc de rendre plus transparents les mécanismes constitutifs des comportements émotionnels qui en font présumément des instruments particulièrement efficaces d'adaptation. Guidé par l'intuition de Simon (1967) d'envisager les émotions comme système d'interruption en face de situations imprévues, mais parfois vitales pour l'organisme, il nous a semblé significatif d'explorer la question des structures de contrôle. Mais avant de pouvoir préjuger d'un tel rôle, il nous a fallu aller sonder systématiquement la littérature concernant les modèles théoriques et les recherches empiriques dans le domaine de la psychologie - et parfois de la neurophysiologie - des émotions. Comme le domaine est assez vaste, allant de la psychologie animale, de l'éthologie et de la

primatologie, jusqu'à la psycholinguistique et l'anthropologie culturelle, en passant par la psychologie expérimentale et la psychologie du développement, il nous a semblé qu'il fallait systématiser la cueillette et imposer quelques lignes directrices.

Le dispositif méthodologique que nous avons adopté à cet égard s'apparente à la méta-analyse, et a consisté à relever les *controverses* majeures qui, dans les dernières décennies, ont animé la réflexion théorique, la recherche empirique et les publications scientifiques en psychologie cognitive des émotions. Nous avons pensé utiliser ces débats de façon heuristique, pour jeter de vigoureux coups de sonde au coeur des théories les plus significatives, et faire par là ressortir les forces en même temps que les faiblesses des concepts les plus fondamentaux. Bien que ce domaine ait été passablement agité, notamment par des discussions relatives à la rétroaction faciale ("facial feedback hypothesis"), à l'opposition entre les approches dimensionnelles ou catégorielles dans la définition et la description des émotions, au déterminisme biopsychologique, au constructivisme social, et à l'hypothèse de l'universalité des expressions émotionnelles, nous avons surtout retenu deux débats qui ont respectivement occupé l'avant-scène au cours des deux dernières décennies. Lancé à la fin des années 1970, le débat entre Zajonc et Lazarus (Lazarus, 1981, 1982, 1984; Zajonc, 1980, 1984a, 1984b) concernant la dépendance relative entre émotion et cognition, a probablement mobilisé l'essentiel des discussions durant les trois premiers quarts de la décennie suivante. Zajonc soutenait que les processus émotionnels étaient d'une nature intrinsèquement différente de celle des processus cognitifs, reposant même sur des structures neurophysiologiques spécifiques. Lazarus, de son côté, y voyait plutôt une forme particulière de cognition, et argumentait qu'aucune réaction ne pouvait être déclenchée sans analyse cognitive préalable.

À la fin des années 1980, les tenants de la théorie de l'appropriation ("appraisal") ont lancé à leur tour une attaque vigoureuse sur la question des émotions de base, envisagées comme processus primitifs du fonctionnement émotionnel et de son expression, une querelle qui a mobilisé les figures les plus marquantes du domaine, et qui a perduré jusqu'au milieu de la décennie 1990 (Camras, 1992; Davidson, 1992; Ekman, 1992a, 1992b; Izard, 1992; Johnson-Laird et Oatley, 1992, 1996; Ortony et Turner, 1990; Panksepp, 1992; Stein et Oatley, 1992; Stein et Trabasso, 1992; Turner et Ortony, 1992; Wierzbicka, 1992). En un certain sens, il s'agissait d'une réactivation partielle du débat précédent, puisque le concept d'"appraisal" recouvrait justement l'idée d'une évaluation cognitive prérequise au déclenchement de toute réaction émotionnelle. Alors que leurs

opposants supposaient l'existence de mécanismes spécifiques - vraisemblablement "câblés" au niveau neurologique - pour un petit nombre d'émotions de base, les tenants de la théorie de l'appropriation postulaient que les unités constitutives des réactions émotionnelles consistaient plutôt dans les différents "appraisals" servant à évaluer chaque situation significative. Selon ces théoriciens, un même type d'"appraisal" - comme par exemple la perception d'un obstacle ou d'une entrave à un but désiré - pourrait agir comme composante d'émotions fort différentes, telles la peur, la colère ou la tristesse.

Nous avons donc mené une analyse approfondie des arguments invoqués de part et d'autre de chaque controverse, afin de cibler, pour notre modèle, les concepts qui apparaissaient incontournables dans la description des structures sous-jacentes aux comportements émotionnels. Ces débats s'avéraient d'autant plus importants pour notre élaboration théorique, qu'ils abordaient notamment la question de l'universalité des réactions émotionnelles, par delà les cultures et même parfois par delà les espèces, en interrogeant les conditions de déclenchement spécifiques à chaque émotion, en même temps que leurs modalités caractéristiques d'expression. Cela nous a permis de spécifier avec beaucoup plus de certitude ce qui pouvait être considéré ou non comme comportement émotionnel, et d'en mieux saisir la dynamique, en rendant explicites les contextes particuliers pour lesquels ces comportements semblaient avoir progressivement émergé comme solutions adaptatives. Nous avons donc utilisé ces affrontements épistémologiques et conceptuels comme instrument de sélection des éléments de base de notre modèle, et nous nous sommes appuyé sur les recherches empiriques relatées à l'appui de ces confrontations pour contraindre et valider l'assemblage de ces concepts dans notre propre élaboration théorique.

Bien que nous ayons exploité de façon moins systématique les controverses existant dans le domaine de l'intelligence artificielle, nous avons quand même recherché là aussi ce que certains débats récents opposant les tenants de l'IAD (Bond et Gasser, 1988; Gasser, 1991; Hewitt, 1985, 1991; Kirsh, 1991b) ou ceux de la Vie artificielle (Brooks, 1991a, 1991b; Steels et Brooks, 1995) à ceux de l'IA plus "classique", pouvaient révéler au niveau des problèmes de fond à résoudre (Bobrow et Hayes, 1985; Kirsh, 1991a), et au sujet des concepts vraiment fondamentaux qui pouvaient résister à de telles confrontations. Nous avons en effet rappelé, dans une section précédente, que la majorité des problèmes de fond rencontrés en IAD, afin d'assurer la coordination d'une communauté d'agents, consistaient en problèmes de contrôle (Bond et Gasser, 1988, p.

10). Par ailleurs, dans une série de textes provocateurs, Brooks (1991a, 1991b, 1995), l'un des principaux tenants de la Vie artificielle, a questionné vivement le concept de représentation qui est à la base de la plupart des recherches en IA "classique", plaidant pour une approche "réactive", selon laquelle le comportement intelligent émergerait plutôt d'un câblage sensori-moteur efficace, greffant intimement l'organisme à son environnement, *sans recours à une structure interne de représentations* ("...use the world as its own model", Brooks, 1991b, p. 140). Or, dans une analyse critique de l'approche de Brooks, Kirsh fait ressortir que, du point de vue de cette théorie également:

> [...] *the hardest problems of intelligent action are related to the control issues* involved in coordinating the various behavioural abilities so that the world itself [...] will be sufficient to decide which activity layer has its moment in the sun. In short, the theme of this alternative theory is that *representation is exchanged for control.* (Kirsh, 1991b, p. 168; les italiques sont de nous)

Bref, aussi bien au coeur de la problématique de l'IAD que de celle de la Vie artificielle, on retrouve cette prédominance des structures de contrôle comme concept fondamental et incontournable. Même au sein de l'IA plus "classique", des débats comme ceux qui porté, durant les années 1970, sur la distinction entre les modes de représentations déclaratif ou procédural, reposaient en fait sur des questions de contrôle (Winograd, 1975). D'un côté, les tenants des représentations déclaratives plaidaient en faveur d'une dissociation entre le contenu des connaissances et les opérations de contrôle effectuées sur ces connaissances, une dissociation qui a d'ailleurs été à la base même de la conceptualisation des langages de programmation logique, comme Prolog. Les tenants des représentations procédurales argumentaient de leur côté que la contextualisation des connaissances requérait que les conditions de leur utilisation fassent partie intégrante des représentations elles-mêmes.

1.3.2. La question de la validation du modèle

La validation d'un modèle théorique sans vérification empirique ne va pas de soi et soulève un problème méthodologique de taille. Dans le contexte de l'IA, c'est l'implantation informatique qui joue habituellement ce rôle. L'exécution du système de programmes réalisé offre à l'observation une simulation du comportement qui a été ainsi représenté, et permet d'évaluer la plausibilité du modèle en même temps que ses failles. Ce n'est cependant pas toujours le cas, et l'un des exemples notoires est probablement la thèse de doctorat de Charniak (1972), qui a eu le mérite d'introduire en intelligence artificielle et en science cognitive rien de moins que le concept de "Démon". Charniak

cherchait à modéliser la façon dont de jeunes enfants parviennent à comprendre des histoires apparemment aussi simples que la suivante:

> Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a top. "Don't do that" said Penny. "Jack has a top. He will make you take it back." (Charniak, 1972, p. 32)

L'une des difficultés, pour une éventuelle implantation informatique, consiste à déterminer comment l'enfant (et par la suite le programme) peut arriver à rattacher au bon référent le pronom "it" à la fin de l'histoire. Bien sûr, il s'agit d'une toupie, mais de laquelle? La règle générale est que le référent se trouve être le plus récent substantif dont il a été question dans la phrase ou le paragraphe. Mais très certainement, Jack ne fera pas retourner au magasin la toupie qu'il possède déjà, et c'est donc plutôt celle qui lui sera offerte qui devra être retournée. Or comment les enfants peuvent-ils déchiffrer aussi facilement un scénario d'une pareille complexité?

Selon Minsky (1985, p. 259), Charniak passa plus d'un an à réfléchir sur cette seule histoire, et plus particulièrement *sur les contraintes qu'elle imposait pour la conception d'un système d'interprétation approprié*. La solution élégante qu'il proposa, c'est que, lorsque notre système cognitif est confronté à des événements significatifs - comme la fête d'enfants dont il est question ici - il génère dans la mémoire de travail des espèces de "sentinelles" dont la fonction est de surveiller la suite de l'histoire afin de déceler toute autre information pertinente qui pourrait se rapporter au thème de l'événement en cours. L'auteur qualifia de "démons" ces structures conditionnelles dynamiques. Sans toutefois se traduire par aucune implantation informatique concrète, sa thèse porta sur diverses questions relatives aux conditions d'activation de ces démons, à la quantité et la complexité de démons requis pour une histoire donnée, aux interactions possibles entre plusieurs démons, etc.

Même si ce chercheur avait initialement envisagé une implantation, la complexité du problème abordé requérait un sérieux approfondissement conceptuel avant qu'une traduction dans un système informatique quelconque puisse être entreprise. Ce qui reste significatif cependant, c'est que l'approche d'ingénierie continua néanmoins d'opérer à travers la recherche d'une conceptualisation suffisamment précise pour permettre "éventuellement" une réalisation informatique. Or, dans notre ambition de déchiffrer le fonctionnement des structures émotionnelles, il n'est sans doute pas surprenant de réaliser

que nous avons été confronté à une surestimation comparable. En effet, il nous avait semblé au départ qu'une implantation dans un environnement informatique à base d'acteurs s'avérait un horizon atteignable dans le cadre de la recherche doctorale. Mais nous avions sous-estimé l'ampleur du champ conceptuel à explorer, et la complexité des choix théoriques à effectuer.

Par ailleurs, le banc d'essai que nous avions initialement envisagé pour l'implantation, celui des *agendas distribués,* s'est révélé, au fur et à mesure de nos progrès, de moins en moins adéquat pour tester et supporter les conceptualisations que nous avions élaborées. Le défi à relever dans ce cas aurait consisté dans la gestion d'un agenda partagé entre plusieurs agents obligés de se rencontrer régulièrement, et forcés de trouver pour cela des périodes de temps communes, ainsi que des lieux disponibles. Le problème est familier dans la plupart des institutions et, en dehors d'une solution "imposée" à l'avance (comme les horaires de cours et de spectacles, les départs de trains et d'avions, etc.) ou du simple va-et-vient d'essais et d'erreurs, il n'y a toujours pas de solution optimale connue. En outre, il nous avait semblé au départ assez naturel d'imaginer les dimensions "émotionnelles" qui pouvaient être impliquées dans la résolution de ce type de problème, que ce soit en termes d'urgence, de priorité, d'insatisfaction, de conflit, d'anxiété, etc. Or, au fur et à mesure de nos lectures au sujet des antécédents qui semblent, quasi universellement et quasi systématiquement, déclencher les mêmes réactions émotionnelles, il nous est progressivement apparu qu'il s'agissait justement là de situations qui échappent généralement à la prévision que permet la tenue d'un agenda. En d'autres termes, le recours à un agenda vise précisément à réduire les risques de "mauvaises surprises", et réciproquement, les émotions interviennent pour aider l'individu à *prendre en charge ce que son agenda habituel ne lui a pas permis de prévoir,* qu'il s'agisse d'inconvénients à éviter ou d'opportunités dont il puisse tirer parti. L'exemple relaté au début de la thèse est particulièrement éloquent à cet égard.

Mais s'il n'y a pas d'implantation informatique dont on pourrait observer et évaluer le comportement, comment alors procéder pour valider le modèle théorique élaboré? Il nous a semblé qu'une alternative valable pouvait consister à soumettre le problème abordé, les objectifs visés ainsi que la solution proposée à un petit nombre d'experts en IA. Dans ce cas cependant, le choix des experts est critique, puisqu'il doit garantir leur neutralité en même temps que leur compétence à évaluer la pertinence et la généralité du modèle. Or il existe dans la communauté scientifique un mécanisme bien rodé et communément reconnu

qui offre de pareilles garanties: la procédure d'arbitrage des publications scientifiques avec comité de lecture. Les évaluateurs y sont anonymes, spécialistes du domaine et généralement fort critiques des productions soumises. Nous avons donc pensé que de traverser une pareille sélection, *surtout si le modèle est confronté à plus d'une école de pensée*, pourrait constituer une forme satisfaisante de validation par des experts. C'est d'ailleurs là l'une des principales raisons qui nous a fait retenir le format de thèse par articles.

Outre un article substantiel portant sur les principales controverses en psychologie des émotions (soumis à *Psychological Bulletin* et reproduit au chapitre 3), trois textes exposant l'élaboration du modèle et quelques-unes des spécifications requises pour son implantation, ont ainsi été acheminés pour publication à des organismes scientifiques reconnus dans le domaine de l'intelligence artificielle. Or *les trois textes ont été retenus*, et deux d'entre eux (rassemblés au chapitre 4) ont même déjà fait l'objet de publication durant l'année 1996 (Aubé et Senteni, 1996a, 1996b). Quant au troisième (reproduit au chapitre 2), il devrait apparaître dans l'un des prochains numéros du *Journal of Artificial Intelligence in Education* où il avait été proposé (voir la lettre d'acceptation, Annexe A). En fait, le premier de ces trois articles avait même été soumis simultanément à deux congrès différents dont les dates d'échéance coïncidaient, l'un aux États-Unis (CIKM'95) et l'autre en Europe de l'Ouest (MAAMAW'96). C'était le premier texte que nous soumettions sur le sujet de notre thèse, et il nous semblait qu'en raison de la singularité du sujet et du fait que nous ne disposions encore pas d'implantation informatique, les chances d'acceptation n'étaient peut-être pas tellement élevées. Or, malgré une sélection rigoureuse dans chaque cas (environ un tiers seulement des propositions ont été retenues), *l'article fut accepté pour présentation - et pour publication - dans chacune des conférences* (Aubé et Senteni, 1995, 1996a).

Nous n'avons évidemment retenu qu'une des deux versions pour la thèse, (troisième section du chapitre 4, la publication de l'autre version étant reproduite à l'Annexe B), mais il est important de comprendre ici que le contenu présenté a néanmoins su convaincre et rallier des experts de deux communautés passablement différentes en IA. La "version américaine" fut présentée au *CIKM Workshop on Intelligent Information Agents*, organisé par Tim Finin et James Mayfield, dans le cadre de la *Fourth International Conference on Information and Knowledge Management (CIKM'95)* tenue à Baltimore les 1er et 2 décembre 1995. Elle avait été sélectionnée avec treize autres présentations retenues sur une

cinquantaine de textes soumis. Cet atelier rassemblait des chercheurs travaillant sur la conception d'agents de service informatiques, et visant notamment le réseau Internet et le World Wide Web comme créneaux privilégiés d'utilisation. Tim Finin se trouve être l'un des directeurs du projet KQML (Knowledge Query and Manipulation Language) destiné à élaborer, sur la base de la théorie des actes de langage, les protocoles standards d'échange et de communication sur les réseaux informatiques. C'est donc avec intérêt qu'on a retrouvé sur Internet (http://www.smli.com/research/tcl/lists/AGENTS/0710.html - reproduit aussi à la fin de l'Annexe B), quelques mois après la conférence CIKM, un court message de Tim Finin suggérant la lecture de notre texte en réponse à un chercheur japonais qui s'informait des possibilités d'exploitation des structures émotionnelles pour la conception d'agents intelligents.

Par ailleurs, la "version européenne" (reproduite au chapitre 4) fut présentée au *7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'96)*, tenu à Eindhoven en Hollande du 22 au 25 janvier 1996, et rassemblant surtout des chercheurs européens et japonais dans le domaine de l'IAD et des systèmes multi-agents. Elle fait partie des 17 présentations retenues sur un total de 51 soumissions, et est désormais publiée chez Springer dans la collection *Lecture Notes in Artificial Intelligence No 1038* (Aubé et Senteni, 1996a). Dans la préface du volume, les éditeurs Walter Van de Velde et John W. Perram présentent notre texte de la façon suivante:

> The paper by Michel Aubé and Alain Senteni, *Emotions as Commitments Operators : A Foundation for Control Structure in Multi-Agents Systems,* is an ontological analysis of control based on notions of resources and commitment. If successful such an analysis may have profound technical impact on the MAS field, just like Newell and Simon's analysis of human problem solving determined - for better or for worse - the technical direction for 20 years of research on rule-based reasoning and weak problem solving methods. (Van de Velde et Perram, 1996, p. VIII; la section de la préface contenant ce passage est également disponible sur le Web à l'adresse: http://arti.vub.ac.be/~walter/papers/maamaw/node2.html)

Or Walter Van de Velde (http://arti.vub.ac.be/~walter/CV/section3_3.html) est co-directeur du département d'IA de l'Université Libre de Bruxelles, aviseur scientifique auprès de la Communauté Scientifique Européenne, membre des jurys de sélection de plusieurs organismes scientifiques en IA (IJCAI, IEEE Expert, Robotics and Autonomous Systems), et conseiller critique pour les projets soumis dans le cadre d'ESPRIT, le projet conjoint de la Communauté européenne en IA. La comparaison, pour le moins élogieuse, qui est faite avec les travaux de Newell et Simon, deux pionniers émérites en IA, atteste

vraisemblablement du fait que les propositions contenues dans notre modèle ne présentent pas d'incohérence majeure avec les concepts existant en IAD, et offrent même suffisamment de nouveauté pour intéresser les chercheurs de pointe dans le domaine.

Nous avons mentionné par ailleurs qu'un autre de nos articles (reproduit au chapitre 2) avait été accepté pour publication dans le *Journal of Artificial Intelligence in Education*. Il s'agit dans ce cas d'une communauté de chercheurs qui oeuvrent principalement en éducation, et qui sont spécifiquement impliqués dans le développement d'outils et d'applications pour l'apprentissage et l'enseignement, à partir des concepts et des résultats issus de la recherche en IA. Compte tenu de l'analyse que nous y faisons des besoins et des possibilités d'appliquer les ressources de l'IA à l'étude et à la modélisation de la motivation et des émotions, l'acceptation de notre article nous conforte une fois de plus de la pertinence et du succès de notre entreprise.

Finalement, nous avons également soumis un texte pour publication à la *Fourth International Conference on Simulation of Adaptive Behavior (SAB'96): From Animals to Animats 4*, tenue à Cape Cod du 9 au 13 septembre 1996, sous la présidence de Pattie Maes. Il s'agit là d'une conférence rassemblant principalement des chercheurs dans le domaine de la robotique, de la Vie artificielle et de l'écologie du comportement ("Behavioral Ecology"). Notre intérêt à y présenter les résultats de nos travaux reposait sur notre souci de garantir à nos positions théoriques la plus grande généralité possible, en situant notre modèle dans une perspective évolutionniste, de sorte qu'il apparaisse également applicable aux émotions animales. Là aussi, la sélection fut passablement sévère, puisque 66 seulement des 150 textes soumis furent retenus pour publication aux MIT Press (Maes, Mataric, Meyer, Pollack, et Wilson, 1996), chaque article ayant fait l'objet d'une double ronde de lecture (voir à ce sujet la préface des Actes à l'Annexe C). Par ailleurs, seulement 35 des 66 textes publiés furent également retenus pour présentation, sur la base de la représentativité et de l'intérêt des positions exposées pour les chercheurs des différentes disciplines rassemblées à cette conférence. Notre article (Aubé et Senteni, 1996b, également reproduit à la quatrième section du chapitre 4) fut sélectionné à la fois pour publication dans les actes édités aux MIT Press, et pour présentation dans le cadre de la conférence elle-même.

En bref, notre modélisation a su pénétrer, et même avec brio, quatre communautés relativement différentes dans le domaine de l'IA, allant de celle intéressée aux applications

éducatives de l'IA, à celle portant sur la communication et les actes de langage, à celle concernant les systèmes multi-agents, et même à celle portant plus spécifiquement sur la robotique et la Vie artificielle. Il nous semble donc pouvoir affirmer que notre élaboration théorique a connu une validation satisfaisante par une large communauté d'experts provenant d'écoles de pensée relativement diversifiées.

1.4. L'organisation et le contenu de la thèse

Outre la présente introduction et une section de conclusion exposant les résultats et les contributions de la recherche, la thèse comprend trois chapitres rassemblant quatre articles ayant déjà été soumis pour publication. Étant donné les créneaux de diffusion visés, tous ont été rédigés en anglais. Il est sans doute important de souligner également ici que le format d'une thèse par articles résulte forcément dans une moins grande étanchéité entre les parties que celui d'un format plus traditionnel. En effet, comme chaque article visait des communautés de lecteurs potentiellement différentes, n'ayant pas nécessairement eu accès aux autres publications, il s'est parfois avéré nécessaire, pour des fins de clarté et de cohérence, de reprendre ou de résumer certaines notions présentées dans les autres textes. Ainsi, bien que chaque article aborde et approfondisse des aspects différents et complémentaires du modèle, le lecteur qui, dans le contexte de la thèse, en fera une lecture suivie et cumulative pourra en retirer à certains moments une vague impression de répétition. Mentionnons finalement que, pour les fins de la thèse, chaque article a été édité afin de rendre les références compatibles avec celle des autres articles, et afin d'ajuster la numérotation des sections en conformité avec la structure de la thèse telle qu'annoncée dans la table des matières. Outre ces ajustements mineurs, chaque article a été reproduit fidèlement, tel qu'acheminé aux organismes de publication.

L'article rapporté dans le chapitre suivant (chapitre 2) situe d'abord le problème de recherche dans le contexte de l'éducation, en faisant ressortir l'extrême nécessité de disposer de modèles théoriques robustes pour aborder le domaine de la motivation et des émotions, ainsi que leurs relations avec le fonctionnement cognitif proprement dit. Le texte rappelle les principales contributions de l'IA pour la compréhension des mécanismes d'apprentissage et de construction des savoirs, notamment au sujet des différents modes de représentation des connaissances. Il fait alors ressortir l'importance de la précision des spécifications obtenues en IA pour la conception ultérieure de stratégies pédagogiques et

didactiques efficaces. L'article plaide ensuite en faveur d'appliquer l'approche et les méthodes de l'IA pour la conceptualisation des phénomènes motivationnels et émotionnels, et donne un aperçu du genre de modélisation qui pourrait en résulter. Il termine en évoquant quelques-unes des applications qui pourraient découler d'un tel modèle afin de rendre plus efficaces les interventions pédagogiques destinées à rehausser la motivation scolaire.

Le troisième chapitre effectue une immersion radicale en psychologie cognitive des émotions. Sa fonction essentielle, dans le cadre de la thèse, était d'identifier et de recueillir une bonne panoplie de concepts fondamentaux, de sélectionner les fragments qui semblaient les plus utiles parmi les principales théories existantes, et d'accumuler du même coup un large répertoire d'exemples, en vue de l'élaboration de notre modèle théorique. Comme nous l'avons mentionné plus haut, nous avons tiré profit des principales controverses en psychologie des émotions comme révélateur des concepts-clés et des articulations théoriques les plus significatives. Cette longue incursion nous a amené à formuler et à consolider l'hypothèse que les structures émotionnelles constituaient une couche particulière de contrôle, entre celle qui assure la gestion des besoins et celle qui supporte le fonctionnement rationnel. Elle nous a également permis d'identifier et de retenir, comme composantes de cette couche de contrôle, un petit nombre de mécanismes de base apparaissant comme responsables de la très grande majorité des comportements émotionnels, et d'articuler avec précision les conditions de leur déclenchement.

Le quatrième chapitre rassemble les deux articles exposant le modèle "computationnel" proprement dit. Il nous a cependant paru important de faire précéder ces deux contributions d'une section portant sur les concepts de coopération et de réciprocité, et reliant ces notions à celle d'engagement ("commitment") qui est développée dans les articles. Le concept d'engagement est en effet de plus en plus fréquemment utilisé en IAD pour rendre compte des phénomènes d'organisation, de coordination et de coopération entre les agents (Bouron, 1992; Gasser, 1991; Winograd et Flores, 1986), et il a progressivement pris un rôle crucial dans notre modèle du fonctionnement émotionnel. Comme nous avons fréquemment référé à ce concept, il nous a paru utile, pour en faciliter la compréhension, d'expliciter au préalable l'éclairage qu'il permet d'apporter au délicat problème de la coopération. Il n'est sans doute pas exagéré d'affirmer ici que la définition particulière que nous proposons de la notion d'engagement dans ce chapitre constitue l'un des concepts-clés et l'une des contributions majeures de notre recherche.

Vient ensuite le texte que nous avons soumis à la conférence de MAAMAW'96, et qui a été publié chez Springer dans la collection *Lecture Notes in Artificial Intelligence* (Aubé et Senteni, 1996a). On y trouve l'articulation de base du modèle, et notamment le lien fondamental que nous établissons entre motivation, gestion des ressources et engagements. Ces derniers y sont expressément conceptualisés comme un type particulier de ressources reposant essentiellement sur la collaboration entre les agents. Les émotions sont alors présentées comme des *opérateurs des engagements* dont la fonction est d'assurer la gestion et la régulation des variations significatives qui peuvent se produire à l'intérieur de ce stock de ressources particulières, mais essentielles pour rendre possible et maintenir de façon adaptative la coopération chez les espèces sociales.

La dernière section du chapitre quatre consiste dans le texte qui a été présenté à SAB'96 et publié chez MIT Press (Aubé et Senteni, 1996b). Au confluent de la Vie artificielle et de l'Écologie du comportement, cet article explore, dans une perspective évolutionniste, les différentes contraintes qui président à l'apparition (dans le contexte de l'évolution) ou à la conception (par l'ingénieur cogniticien) d'agents biologiques ou informatiques véritablement doués d'autonomie. La gestion des ressources y est encore une fois présentée comme le principe essentiel derrière tout mécanisme d'ordre motivationnel, et la relation entre les deux "couches" de contrôle que constituent les besoins et les émotions y est clairement exposée. Le texte fait ensuite une présentation sommaire du concept d'engagement tel qu'utilisé en IAD, et formule un certain nombre de réserves critiques à propos des principales interprétations qui en ont été faites jusqu'ici. L'article se termine par une énumération de quelques-unes des spécifications qui découlent du modèle en vue d'une éventuelle implantation dans un environnement informatique à base d'acteurs.

La conclusion générale de la thèse (chapitre 5) vise finalement à rassembler les contributions originales de la recherche, notamment en ce qui concerne le rôle primordial joué par le concept d'engagement pour la compréhension des phénomènes motivationnels. Nous tâcherons en effet de montrer que ce construit hypothétique, tel que nous l'avons défini et utilisé, présente une valeur descriptive et explicative exceptionnelle. Il permet d'établir une distinction plus claire que jamais entre les besoins et les émotions, il aide à mieux comprendre et formuler le type de contraintes qui ont pu contribuer à l'apparition des structures émotionnelles comme instruments inédits d'adaptation dans l'évolution des mammifères et des primates, il rend compte de la nature intrinsèquement sociale des

émotions, et il offre une cadre cohérent pour la répartition taxinomique des différentes catégories d'émotions. Au profit de la recherche en IAD, il offre une base pour l'explicitation des contraintes qui président à la coopération entre les agents, et permet de formuler des spécifications précises pour une implantation éventuelle de structures de contrôle remplissant les fonctions principales des mécanismes émotionnels chez les vivants. Finalement, dans le contexte de la motivation scolaire, il offre un champ nouveau d'exploration pour mieux comprendre les bases intrinsèques et interactives sur lesquelles reposent la persistance et l'engagement des élèves. Nous terminons la conclusion générale en dégageant quelques-uns des avantages et certaines des limites de l'approche adoptée pour la réalisation de cette recherche.

Chapitre 2.

La nécessité d'une telle recherche pour l'éducation

2.1. Présentation de l'article

De prime abord, il est loin d'être évident que l'élaboration d'un modèle "computationnel" des structures émotionnelles et de leur fonctionnement, puisse offrir une contribution significative au domaine de l'éducation. D'une part, il y a lieu de se demander si l'approche de modélisation utilisée en intelligence artificielle est appropriée au genre de problèmes qui sont rencontrés dans le contexte de l'école et des pratiques d'enseignement. D'autre part, il s'agit d'examiner quelle compréhension des phénomènes émotionnels est effectivement requise et pertinente pour le milieu de l'éducation, et quels outils d'interprétation et d'intervention pourraient être tirés d'une conceptualisation adéquate. Le rôle du premier article est donc de préciser le sens de cette démarche et proposer un certains nombre d'arguments convaincants à cet égard.

Nous commençons par rappeler l'importance pour l'éducation de plusieurs contributions provenant des sciences cognitives, et nous situons le rôle de catalyseur que l'intelligence artificielle a joué dans les progrès que la plupart de ces disciplines ont connus dans les deux dernières décennies. Nous énumérons alors quelques-uns des concepts qui, bien que provenant clairement de l'IA, ont pénétré avec bonheur le domaine des pratiques pédagogiques et didactiques. Les illustrations les plus évidentes concernent la description et l'analyse des différents modes de représentation des connaissances, par exemple sous les formes déclaratives, procédurales ou conditionnelles. Mais il est surtout important de mentionner que ces concepts autorisent désormais des traductions avantageuses, en termes de stratégies novatrices d'enseignement et de support à l'apprentissage. En effet, si ces notions ont déjà commencé à influencer les milieux de l'école et de la formation des maîtres, c'est parce que l'approche constructiviste qui avait initialement présidé à leur

conception, afin de doter des systèmes informatiques de pensée et de raisonnement, s'applique tout aussi naturellement à l'objectif de favoriser chez les enfants le développement de ces mêmes capacités. Or, bien d'autres concepts sont également issus de cette entreprise, offrant des supports accrus pour le développement cognitif, tels une meilleure compréhension des différents systèmes de mémoire, l'enrichissement du répertoire de stratégies générales de résolution de problèmes, l'exploitation dynamique des erreurs ("bugs") comme leviers d'apprentissage, etc.

Mais il est justement frappant de constater que les contributions de l'IA à l'éducation restent surtout cantonnées au plan cognitif, alors que la question de la motivation scolaire est pratiquement ignorée. Il s'agit pourtant d'un domaine qui soulève une foule de problèmes, *probablement même parmi les plus complexes et les plus déterminants de toute la carrière scolaire d'un enfant.* Or, comme pour la représentation des connaissances, la résolution de problème ou la métacognition, il est souhaitable que là aussi les interventions puissent s'appuyer sur les modèles théoriques les plus fiables et les plus solides possibles. Nous proposons donc que la même approche constructiviste d'ingénierie soit appliquée à la conceptualisation et à la simulation des processus responsables de la motivation et des émotions. Comme stratégie pour mieux comprendre leur raison d'être et leur fonctionnement, nous nous demandons pour quelles raisons, ou dans quel contexte, un système informatique complexe pourrait tirer profit - et même requérir carrément l'existence pour son adaptation - de mécanismes comme ceux qui sont responsables du déclenchement et de l'expression des besoins ou des émotions. Quelques tentatives déjà amorcées en ce sens en intelligence artificielle sont également énumérées.

Cette réflexion nous amène à examiner d'un peu plus près les principaux modèles existants dans le domaine de la psychologie scolaire et de l'éducation, afin de décrire les composantes de la motivation scolaire et d'en expliquer la dynamique. Nous tâchons ensuite de faire ressortir les limites de ces approches, notamment le fait que ces modèles cherchent surtout à composer, en un schéma cohérent, les différentes variables qui ont été identifiées expérimentalement comme effectives, telles la perception des attentes du système scolaire, la conception que l'élève a de l'intelligence, la signification qu'il attribue à la tâche, la contrôlabilité qu'il croit en avoir, etc. Mais ces théories n'expliquent pas d'où ces variables tirent leur importance, ils n'expliquent pas d'où viennent les buts mêmes que se fixe un sujet, bref ils ne démontent pas sous nos yeux la mécanique de la motivation. Nous croyons que cette limite vient notamment du fait que la motivation est plus ou moins

prise pour acquise, puisque les humains, comme la plupart des autres vivants, semblent déjà avoir d'inscrit en eux le dynamisme qui les pousse à agir, et à choisir entre plusieurs opportunités d'action. Or cette présupposition n'est pas acceptable pour le chercheur en IA ou en Vie artificielle qui souhaiterait reproduire ou simuler un organisme, même doué d'une autonomie minimale. Il lui faudra en effet trouver comment inscrire, de façon précise et opérationnelle, ce dynamisme de base à l'intérieur de l'automate ainsi créé, et donc se demander au préalable quelle est la fonction de cette propriété nouvelle, en vue d'une meilleure adaptation de l'organisme à son environnement. Dans la suite de l'article, nous exposons comment un tel défi pourrait éventuellement être relevé, quels concepts additionnels semblent requis, et nous donnons un aperçu sommaire du modèle qui sera développé au cours de la thèse, comme contribution théorique à cet égard.

Nous terminons l'article en soulignant quelques-unes des conséquences espérées d'une telle approche pour l'éducation. Nous examinons notamment comment certains éclaircissements au sujet du concept *d'engagement*, qui est perçu par la plupart des chercheurs comme l'un des principaux indicateurs de la motivation scolaire, pourraient suggérer des pistes d'interventions intéressantes. Il s'agit d'un concept qui prend également de plus en plus d'importance dans le domaine de l'intelligence artificielle distribuée, pour rendre compte des comportements d'organisation, de coopération et de gestion des conflits. Bien qu'il existe certaines nuances dans les définitions qui sont données de ce concept dans chacune des disciplines, il nous a semblé qu'une interprétation à caractère plus sociologique (en terme de contrat implicite) que psychologique (en terme de persistance), mettant l'accent sur l'interaction entre les acteurs en jeu, pouvait offrir une compréhension plus approfondie de la mécanique motivationnelle, et suggérer du coup, de meilleures stratégies d'intervention. L'article a été soumis au *Journal of Artificial Intelligence in Education* en décembre 1996, et son acceptation pour publication a été confirmée par écrit (voir Annexe A).

2.2. Towards computational models of motivation: A much needed foundation for social sciences and education[1]

Abstract. Most basic concepts about knowledge representation, such as declarative or procedural knowledge, come from research in artificial intelligence, and have percolated from there through other cognitive sciences to finally reach the domain of educational practice. Therein they are beginning to have a profound impact upon curriculum design, didactics, teaching strategies and teacher education. Meanwhile, the domain of student motivation is calling for increasing attention, but existing models remain essentially descriptive: they relate the variables that are known to matter empirically, yet they offer little explanatory power about what activates willful behavior and goal orientation. The paper advocates the need for basic computational models of motivation as new fundamental tools for education, and suggests using the design-based approach of artificial intelligence to lay down such models. Motivation includes physiological needs as well as emotional reactions, but our proposal for a model will only be concerned with emotions. It is argued that motivation has to do with resource management, and that the particular resources that concern emotions are the ones that are obtained from others, through their having been committed to provide them. Thus viewed, emotions are essentially of a social nature, and have to do with regulating and managing commitments that bind individuals together in action, communication and cooperation. Preliminary specifications for a computational model of emotions are formulated, and consequences for current theories of emotions and of education are formulated in conclusion.

Keywords. Artificial intelligence, Education, Motivation, Emotions, Commitments

Acknowledgements: I am grateful to Alison d'Anglejan, Jacques Tardif and Rolland Viau for their careful and critical reading of the manuscript, and their helpful comments.

2.2.1. Introduction

Within the last thirty years, Artificial Intelligence (AI) has been one of the research areas that has most contributed to the understanding of cognitive phenomena and to the development of related disciplines (Gardner, 1985; Johnson-Laird, 1988; Lindsay and Norman, 1972; Norman, 1981; Pylyshyn, 1984). This field first had a dramatic influence upon linguistics, notably through Chomsky's work on transformational grammars and automata theory (Chomsky, 1963, 1965). Then it gradually spread to cognitive psychology, especially through the work of Newell and Simon (1963, 1972) on human problem-solving, and through many influential doctoral theses at the MIT AI laboratory (Minsky, 1968; Winograd, 1970; Winston, 1975). These new trends in research quickly found their way through most of the significative domains of interest within the area of

---

[1] Texte soumis, en décembre 1996, pour publication dans le *Journal of Artificial Intelligence in Education*. La lettre attestant de son acceptation est reproduite à l'Annexe A.

information and knowledge processing: visual and auditory perception; locomotion, prehension and action; learning, memory and attention; deduction, induction and problem solving; analogical reasoning and imagery; communication, speech processing and natural language (Johnson-Laird, 1988; Lindsay and Norman, 1972; Neisser, 1967)...

2.2.2. AI and cognitive science as foundations for education

Such beneficial effects inevitably were to reach the domain of education as well. The first applications bore on computer assisted instruction (CAI), due to the strong influence skinnerian behaviorism still had on the field (Kurland and Kurland, 1987). But subsequently a new trend of educational research emerged that was more "epistemologically grounded", notably with the work of Alan Kay on the Dynabook at the Xerox Palo Alto Research Center, which led to the design of the Smalltalk object-oriented language (Goldberg and Robson, 1983) and to related teaching applications (Goldberg and Kay, 1977; Ryan, 1991), and with the work of Seymour Papert on computational geometry at MIT (Abelson and diSessa, 1980; Minsky and Papert, 1969), which led to the development of the Logo programming environment (Feurzeig et al., 1969; Papert, 1980). The latter project induced a large diversity of applied studies during the following years, centered mainly on teaching and learning mathematics and science (Carver, 1987; Clements, 1987; Pea and Kurland, 1987; Swan, 1989).

Yet the most profound and most enduring impact of AI upon education probably came from the new concepts developed for characterizing thought processes which filtered mainly through cognitive psychology (Anderson, 1980; Gagné, Yekovitch and Yekovitch, 1993; Tardif, 1992, 1995). Such an influential role should not be too surprising though, considering that one of the basic goals of most educational systems is to foster knowledge acquisition and transfer, and the fact that adequate tools for representing such processes are still scarce. It is indeed from AI research that stem some of the basic concepts that are currently used to describe and characterize knowledge representation, in declarative or procedural formats (Hewitt, 1971; Winograd, 1970, 1975), or in terms of more compact data structures such as frames (Minsky, 1975, 1985), scripts (Abelson, 1981; Schank and Abelson, 1977) or schemata (Rumelhart, 1975, 1981).

Such tools emerged from the concrete task of having computer programs represent various situations in ways that would make it possible to transform them and manipulate them dynamically. One good illustration of how such concepts emerged and of the profound impact they soon had upon educational practice, comes from the concept of *procedure*, a basic notion within the Logo programming language. This concept is more or less akin to the concept of a computer program, yet a program that has its own name so as to be invoked in a modular fashion (Minsky, 1970). It could thus be used within an instruction line just as another primitive of the language. The powerful programming technique of recursion actually amounts to the special case wherein a procedure makes a call to itself as it would for any other instruction.

So far, this is mere computer science. But the concept of procedure also came to be used in AI and simulation as a way of capturing thought processes themselves, through offering a format different from the usual propositional (or declarative) representation of knowledge. Indeed, one might well consider that a procedure for drawing a regular polygon, amounts to a kind of dynamic definition of such an object, an approach that could be used as well for representing a large variety of other concepts in a readily usable form. Those definitions are formal ones, since they are stated in a computer language, and they can thus be used for theorem proving. Turtle Geometry (Abelson and diSessa, 1980) actually offers one good example of such formal reasoning, from problems of traditional Euclidian geometry, to tridimensional and spherical geometry, to the field of topology, and even up to capturing - and proving - some of the geometrical properties specified within the theory of general relativity itself. But at the same time, procedures could be used as the embodiment of direct and concrete representations of various goals and actions to be executed. Winograd (1970), in his doctoral dissertation on the design of a computer system for understanding natural language, is probably the researcher who went the farthest in implementing such a procedural model. All pieces of knowledge within his system were actually represented as procedures: rules of syntax whose execution would parse the word strings into noun groups or verb groups, rules of semantics for assigning meaning to those groups and updating the data base accordingly, as well as the problem-solving strategies for translating the verbal statements into concrete actions upon the simulated microworld of blocks. Even the dictionary definitions were implemented as fragments of procedures, with blanks to be filled when assembled together with other fragments.

Procedural representations thus stepped on stage. Such a remarkable experience with procedures led a few researchers from the AI community (Hewitt, 1971; Minsky, 1970; Winograd, 1970) to envision this concept as a possible basis for representing certain types of knowledge within the brain itself. The following example will illustrate the idea at stake: when asked for a definition of unfamiliar concepts, such as the spiral, most people first have to sketch in the air with their fingers an instance of this geometrical curve, before they can state anything more formal. Such typical behavior very much suggests that the definition for the corresponding concept might be contained within the set of actions necessary to produce the required instance: the appropriate procedure first has to be run, and then, step by step, the instructions themselves can be recovered, and the concept fully described. A similar phenomenon occurs when someone has to spell an unfamiliar word: the person typically has to write down various versions of the word, as if the orthographical information were not attached to the word itself but rather were contained within the procedures for writing or for reading it.

Attractive as it might seem, this new paradigm was not established so easily, and its venue raised an important controversy within the AI community (Winograd, 1975). The main opponents were defendents of declarative representations, an alternate paradigm whose origin could be traced at least back to Aristotle's propositional logic. This older view held that most knowledge could be rigorously represented as propositions, whose truth values could be derived with the tools of Boolean algebra. Predicate calculus offered a still more powerful treatment of such representations, by permitting the inclusion and management of variables and quantifiers within propositions. Yet the controversy between both forms of knowledge had the beneficial and heuristic effect of stimulating further conceptual elaborations for representation, in the direction of integrating information from both formats into more structured packages such as frames, scripts and schemata (Minsky, 1975; Rumelhart, 1975, 1981; Schank and Abelson, 1977). Indeed, the two modes made sense and proved useful in different circumstances and with different kinds of problems, they both had their merits and their shortcomings, and it seemed plausible that the brain had evolved ways to get the best from both of them.

Resolving the controversy by introducing new structured ways for representing and manipulating information led to the design of yet more powerful programming languages, providing the user with a larger variety of tools embodying different types or categories of knowledge (Bobrow and Winograd, 1977, 1979). This progress towards the specification

of representational and thinking processes quickly spread in turn to related fields in cognitive science, such as linguistics and psychology (Abelson, 1981; Anderson, 1980; Bower, Black and Turner, 1979). And as could be expected, it is precisely around those concepts that recent approaches in education (Gagné, Yekovitch and Yekovitch, 1993; Paquette, Aubin and Crevier, 1994; Tardif, 1992, 1995) have built their own working model of the learner's cognitive structure and founded their general principles underlying efficient learning and teaching strategies. Distinguishing between such recognizable formats as declarative, procedural or contextual representations enabled practitioners from education to partition the curriculum accordingly, and to design appropriate sets of activities through which students could best acquire them and put them into pratice.

Declarative representations, for instance, would correspond to such pieces of knowledge as facts, rules, laws or principles, and they would be mainly acquired through listening, reading and memorizing. The teacher could facilitate their learning though, by using two types of strategies. The first one, elaboration, consists in providing the learner with many access routes for the new information to be stored. The teacher could make use of striking anecdotes to render the facts more salient, or give a variety of examples and then ask the students to personalize them by generating their own. Organization, the second strategy, has to do with repackaging the information into a smaller number of units and then stressing the hierarchical relations between them. The whole network of information could thus be more easily retrieved and manipulated through invoking the top unit as the entry point. Procedural representations, on the other hand, typically correspond to sequences of actions that capture the "know-how" aspect of learned skills. These are mainly acquired through apprenticeship, by careful observation and imitation of an appropriate model, followed by thorough practice. Here again, the teacher can support their acquisition by resorting to some well-defined strategies. The first one, proceduralization, involves making explicit, in terms of simple productions, every single action the student should go through in performing the task, while the second, composition, explains how to link and merge these productions into a coherent sequence of steps whose execution will enact the skill.

Similar strategies have been devised as well for conditional knowledge, a particular form of procedural representations for the recognition of contextual constraints, that some authors prefer to distinguish from the standard procedural format (Pintrich and Garcia, 1994; Tardif, 1992). The main conclusion of all this work, is that cognitive science,

especially through some of the powerful ideas issued from AI theorizing, could thus be seen as providing a kind of foundation for educational research, most essentially in the domain of knowledge representation. It offered new theories for testing (Anderson, 1983; Newell, 1990; Newell and Simon, 1972), new principles for the design of tutorials (Anderson et al., 1987), a new rationale for the conception of effective teaching strategies (Paquette, Aubin and Crevier, 1994; Tardif, 1992). Through all such endeavors, the cognitive science enterprise thus created new conceptual tools for the understanding of complex and enduring problems about knowledge, learning and education, such as: the simulation and modelling of phenomena like intuition, insight and discovery (Kaplan and Simon, 1990; Kulkarni and Simon, 1988; Simon, 1995); the characterization of expertise versus novice behavior (Anders Ericsson and Smith, 1991; Chi, Glaser and Farr, 1988; Hoffman, 1992); the heuristic function of errors in learning (Tardif, 1995; Winston, 1970); the transfer of learning (Butterfield and Nelson, 1989; Cormier and Hagman, 1987; Detterman and Sternberg, 1993; Singley and Anderson, 1989; Winston, 1978); cooperation and collaborative problem solving (Winograd and Flores, 1986), and analogical reasoning (Gentner and Stevens, 1983; Holyoak and Thagard, 1995; Vosniadou and Ortony, 1989).

2.2.3. Motivation in education: in need of basic models and concepts

But as is often the case in such circumstances, the sudden popularity of cognitive science has also brought its own drawbacks. One of these has been a rapid increase in the gap between cognitive and motivational aspects of learning and education, a weakness that had been stressed from the very beginning by some of the insiders (Brown et al., 1983; Neisser, 1963; Norman, 1981; Simon, 1967), all of whom have advocated a more systemic and a better integrated view of cognition and affect in behavior. As a matter of fact, this split in interests has perhaps even been partly intentional, as Gardner (1985) reports in the history of the cognitive science movement, while he was stating the basic characteristics of the new approach:

> The third feature of cognitive science is the deliberative decision to de-emphasize certain factors which may be important for cognitive functioning, but whose inclusion at this point would unnecessarily complicate the cognitive-science enterprise. These factors include the influence of affective factors or emotions, the contribution of historical and cultural factors, and the role of the background context in which particular actions or thoughts occur. (Gardner, 1985, p. 7)

Within the last ten years though, there has been a slight reversal in this trend. For one thing, there has been an increasing preoccupation with cultural factors and background context, which grew into a domain of research that is now being called "situated learning" or situated cognition (Brown, Collins and Duguid, 1989; Greeno, Smith and Moore, 1993; Rogoff and Lave, 1984). On the other hand, there also has been a revival of interest around motivational factors in education (Dweck, 1986; Lepper, 1988; Pintrich, Brown and Weinstein, 1994; Tardif, 1992; Viau, 1994). And this is good news, since therein lie some of the most dramatic problems educators are confronted with in their daily practice: what is it that makes some learners invest so much into their learning activities, while others find so little interest in them - or even get so stressed by them - that they eventually drop out of the educational system? This problem is at the very core of the whole enterprise, since it questions whether there is such a thing as "teaching methods". These only seem to work with interested students - who might perhaps even succeed without them, or even without a teacher - but they abruptly hit a rough barrier with all those who are not already willing to invest much in their classroom lives.

A most interesting aspect of current studies on student motivation, is that they have been quite influenced themselves by the "cognitivist revolution", which gave "inside variables" a scientifically valid epistemological status. These studies thus stand at a fair distance from the original behaviorist trend of earlier psychological work upon the subject of motivation, that excluded from consideration factors and variables that could not be directly observed. This change of approach is probably best witnessed by the fact that the main interest has shifted from extrinsic causes of motivation to intrinsic ones: namely how the individual's goals and intentions get affected by the way the subject construes and interprets his environment, as well as his own role and status (Deci and Ryan, 1985; Lepper, 1983; Pintrich and De Groot, 1990; Pintrich and Garcia, 1994; Schunk, 1991; Schunk and Meece, 1992). In the cognitivist perspective, student motivation is defined in terms of the individual's commitment and persistence in his choices of plans and actions (Tardif, 1992, 1995; Viau, 1994). It is a dynamic state that is much affected by the subject's inner world: his memory of past successes or failures, his conception of learning and intelligence as innate or as constructive processes, his perception of the schooling system as truly fostering learning or as merely performing assessment, his estimation of the task as more or less meaningful, as more or less achievable, and as being more or less under his control.

Yet the bulk of research on student motivation remains mostly empirical, and essentially practice-oriented. This might seem a mandatory focus, considering the urgent and dramatic aspect of the problem: there is indeed a myriad of teachers (and students!) gasping daily on the front lines, waiting for adequate tools and remedies to help meet their demands. But on the other hand, we lack some basic models for understanding what is really happening when someone gets "wounded" on the motivational "battlefield", what are the basic underlying processes involved, how affective matters relate to the cognitive processes, and what could best be done to help, if something could actually be attempted. Of course, there already exist a few useful models (Brophy, 1983; Tardif, 1992; Viau, 1994). But they are all typically characterized by their efforts to organize into some structure variables that have been shown empirically to affect student motivation, including some of the inner conceptions of the student. These have to do, as stated earlier, with the goals that are attributed to the schooling system, with the conception of intelligence as fixed or as modifiable, with the perceived meaningfulness of the task, with its controllability, with the efforts estimated as required for mastering it... This is already quite helpful, yet one might wonder whether it would be possible to delve further so as to reach the kind of theorizing that led, on the cognitive side, to such concepts as declarative, procedural or conditional representations, and later on to the more complex data structures of frames, scripts and schemata. The actual models in motivation do point to some of the factors upon which the teacher may act so as to foster in students greater persistence, interest in and commitment to school work, or so as to reduce the unnecessary stress and burden of certain learning situations. But they do not explain at all what it is that makes an individual like or dislike something, where the willingness - or the resistance - comes from. They state that such and such a factor will affect motivation more or less indirectly, but they offer no indication as to why such factors should have become important, they do not say much about what activates willful behavior or goal-orientation, they do not really get at the mechanics of motivation.

The question that is raised then is whether such an explanation could be rigorously offered? In other words, is it reasonable to ask for a workable computational model of motivation? As far back as 1967, in a paper entitled *Motivational and Emotional Controls of Cognition*, Herbert Simon, one of the founders of the AI field, was probably the first to seriously address such a question, and to outline as well a tentative answer:

> This paper will attempt to show how motivational and emotional controls over cognition can be incorporated into an information-processing system, so that thinking

> will take place in "intimate association with emotions and feelings", and will serve a "multiplicity of motives at the same time". (Simon, 1967, p. 30)
> The theory explains how a basically serial information processor endowed with multiple needs behaves adaptatively and survives in an environment that presents unpredictable threats and opportunities. (Simon, 1967, p. 39)

Simon's model rested upon two basic mechanisms, a goal-terminating mechanism dealing with current goals, one at a time, and an interruption mechanism that could set aside ongoing programs when real-time needs with high priority were encountered. Simon used the term motivation "to designate that which controls attention at any given time" (1967, p. 34), and he associated the operation of the interruption device with what is usually called emotions when the corresponding behavior occurs in humans. Undoubtedly, this might appear to be oversimplified, but it nevertheless captured one important aspect of emotional behavior, one that had already been singled out in Psychology by advocates of the so-called conflict theories of emotion (Angier, 1927; Mandler, 1964). Be it as it may, nobody dared follow Simon on such a perilous path for about ten more years, that is until the work of Aaron Sloman and his team at the University of Birmingham (Beaudoin, 1994; Beaudoin and Sloman, 1993; Sloman, 1987, 1992; Sloman and Croucher, 1981). Recently, in a humble mood, Sloman (1995) stated that he considered most of his work on the subject as "but footnotes to Simon's seminal paper". However, on more careful examination, his endeavour appears quite elaborate and fairly ambitious:

> For a full account of [emotions and attitudes] we require a theory of how mental states are generated and controlled [...] - a theory about the mechanisms of the mind. The theory should explain how internal representations are built up, stored, compared, and used to make inferences, formulate plans or control actions [...] Emotions are analysed as states in which powerful motives respond to relevant beliefs by triggering mechanisms required by resource-limited intelligent systems. (Sloman, 1987, p. 218)

Due to the central role attention plays within the theory, together with the basic requirement of a context-sensitive filter for controlling which motive could be attended to at any given time, and hence be allocated due resources for completion, the hypothesized model has been called the "Attention Filter Penetration Theory (AFP)". Now, such a rigorous attempt offers an excellent illustration of the kind of approach and the level of analysis that now seems mandatory within the domain of motivation research. Affect has indeed to be deeply integrated with cognition, but it has to be specified in terms of the primitive processes upon which it rests, and the basic layer at which it operates.

One contention of the present paper is that AI methods and concepts provide the proper foundation for such theorizing. Over the last 30 years, the development of computer hardware and software to represent behaviors endowed with thinking and intelligence has been highly rewarding. This has been reflected not only in the design of new hardware or software per se, but also in the generation of new concepts for the rigorous understanding of knowledge representation, and in a deepened conceptualization of such processes as memory, language and expert problem-solving. Would it not be reasonable then to assume that comparable pay-offs could emerge from the grinding of affective and motivational processes through the very same mill?

2.2.4. Towards a computational model of emotions

The aim of the present article is thus to advocate the relevance of building workable computational models of basic motivational processes, using the AI approach, to serve the interests of educational theory and practice. But since the whole domain of motivation is perhaps too extensive for the purpose envisioned here, we will first define some more realistic parameters. For one thing, the field of motivation typically encompasses all kinds of factors and dispositions that lead an organism to act, and which would hence determine its choices and preferences among various alternative behaviors (Cofer and Appley, 1964). But this includes different types of incentives such as thirst, hunger, fatigue or sexual drive (Toates, 1986), as well as the whole set of emotional reactions evidenced in anxiety, joy, shame, guilt, attachment, anger or sadness.

There is a fairly large consensus that the whole field should at least be split into the broad categories of needs and emotions, and a few criteria for distinguishing them have been proposed in the psychological literature (Plutchik, 1984, p. 214). For instance, emotions are typically triggered by external stimuli, while needs emerge from internal variations within physiological variables. Moreover, emotions usually arise as a consequence of the presence of certain objects or events, while needs, on the contrary, rather manifest themselves when there is a lack or absence of stimuli of homeostatic value. Antecedents of emotions - such as threat, goal thwarting, danger or loss - are relatively abstract and general, while needs are dependent upon fairly specific substances, such as food or water. Most emotional reactions follow from the perception and the appraisal of an event, while needs often precede the search for their appropriate incentive. Events that trigger emotions could occur on a random basis, while needs appear to be of a cyclic nature.

Although we feel compelled to establish relationships among the various forms of motivational processes that must be integrated within a functional system, it nevertheless seems wise to limit the analysis initially to emotional behavior. On heuristic grounds, it is likely that the motivational layer responsible for the regulation of basic physiological needs has much less direct impact on learning or cognition than does the (evolutionary more recent) layer which appears responsible for the management processes involved in pride and contentment, trust and confidence, stress and anxiety, persistence and commitment, self-image and attachment figures... Yet it is far from obvious that processes such as those governing emotions, even properly modelled, could offer conceptual tools sufficiently powerful to capture some of the basic problems that confront learning and education. In AI terms, this amounts to asking whether making provision for emotions within a complex problem-solving program could increase the overall performance and efficiency of the system. In the preceding section, we recalled Sloman's stance that, for a full understanding of attitudes and emotions, one would need a more general theory of cognition, planning, problem-solving and other mental processes. But what about the converse view that, perhaps, a computational theory of emotions might prove necessary for a full account of learning and intelligent behavior? Interestingly enough, in the *The Society of Mind*, Minsky (1985) goes so far as to suggest that:

> The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions. I suspect that once we give machines the ability to alter their own abilities we'll have to provide them with all sorts of complex checks and balances. (Minsky, 1985, p. 163)

The author begs the question of the purpose emotions serve, and raises the possibility that they might play a much greater role than envisioned so far in the management and regulation of thought processes. If knowledge and concepts are to interact in an efficient and articulated manner within the educated mind, they have to be weighted some way, according to some structured system of values. Otherwise, every piece of knowledge is going to be of equal importance, with the result that most decision-making procedures would go awry and no inner conflict could ever be settled. But then, where is this weighting scheme to come from, and how should it be designed? Similar interrogations about the relation between cognition and emotion are being heard more and more frequently within many disciplines inside cognitive science, and it so happens that they all seem related as well to the questions of situated cognition and of the socio-cultural factors

which condition and constrain contextual learning. We will examine briefly a few examples taken from such diverse fields as neurosciences, evolutionary biology and distributed AI, that might throw new light upon this problem. In all cases, *the very fabrication of knowledge as well as its use seem to emerge as an adaptive way of handling fundamental problems raised through social interactions*, such as communication, cooperation and conflict.

2.2.4.1. Neuroscience

In a recent book entitled *Descarte's Error: Emotion, Reason and the Human Brain*, the neurologist Antonio R. Damasio (1994a; see also Damasio, 1994b) presents the results of twenty years of research with patients suffering from prefrontal lobe damage. As a prototypical example, he recalls a famous case from the last century, Phineas Gage, a railroad foreman who, as the result of an explosion, had a one meter iron rod traverse his skull and forehead. The strange thing is that such a staggering wound did not seem to affect his knowledge, memory capacity, reasoning skills or learning abilities. Yet, to his neighbours, he became an entirely different man, no longer showing respect for social conventions, with little concern for reward or punishment, unable to keep commitments, make decisions, anticipate the future and plan accordingly. His whole personality was destroyed, he lost friends, wife, employment one after the other, and ended life miserably. The neurologist then compares Gage's fate to that of a modern patient, Elliot, who had gone through brain surgery to have a large deadly tumor removed from his prefrontal lobe. Extensive testing of his intellectual capabilities has revealed no specific impairment, although in real-life decision making, he showed totally defective. Damasio thus concludes that the deficits and alterations in this patient's behavior are similar to those found in Gage, and reflects that "the cold-bloodedness of Elliot's reasoning prevented him from assigning different values to different options, and made his decision-making landscape hopelessly flat." (Damasio, 1994a, p. 51)

It so happens that the frontal lobes (especially the orbital region) have more reciprocal connections with the limbic structures responsible for emotional processing than any other cortical region, and are thus considered as the neocortical representation of the limbic system (Davidson, 1984, 1992; LeDoux, 1987; MacLean, 1993; Rolls, 1990). Interestingly enough, the prefrontal cortex is also the most recent acquisition of the primates in evolution. This area is responsible for willful behavior resulting from an

ability to anticipate the future, plan a course of action and reach decisions accordingly. It is also critical in developing appropriate social behavior and interaction. Monkeys with orbital frontal lesions become socially withdrawn and even fail to re-establish the close relationships they had entertained with their kins before the surgery (Kolb and Taylor, 1990). Such individuals often get attacked and injured by their former friends and relatives (Kling, 1986). One plausible hypothesis might be that, as in Gage's and Elliot's cases, they evidence some kind of *social blindness* with no ability to decode the current signals that are used to regulate interaction among peers. In computational terms, this suggests that perhaps our most complex and recent brain development makes use of the older emotional structures buried within the limbic system as some kind of operating system. If this were indeed the case, not much cognition - at least not very complex cognition - could by-pass this underlying emotional control structure.

2.2.4.2. The "social intelligence hypothesis"

The evolutionary biologist R. I. M. Dunbar (1993) has recently proposed a fascinating theory about the coevolution of the human brain, social behavior and language. He took the ratio between the volume of the neocortex and the volume of the rest of the brain as a basic measure of neocortical size. He then compared this index for more than thirty different nonhuman primate genera, and found that it covaried nicely with the mean group size in each species (accounting for 76% of the variance): animals organized in larger group life also bore larger neocortical ratios. When Dunbar applied his equation to human neocortical size, the predicted group size was of 147,8 individuals which, at first sight, seems to largely underestimate our typically overcrowded demography. In nonhuman primates, though, group size does not refer to the whole population, but to the usual stable cohesive band. Dunbar thus scanned through the ethnographic literature for any mention of historical or contemporary human hunter-gatherer societies and found a mean size for the typical band, clan or village of... 148,4 inhabitants. He also found that the estimates of the size of Neolithic villages in Mesopotomia were of similar magnitude. He then wondered about typical organized groups, such as professional armies, and found that the company, the smallest independent unit in most armies, invariably contained, from the Roman times up to the Second World War, between 100 and 200 men! Even academic communities rarely consist of more than 200 researchers, a point from which there is usually a split into subdisciplines. The author also reports experiments which

estimated the total acquaintances network size of their subjects at a mean number of 134 persons.

This finding fits well with the "social intelligence hypothesis" (Cheney, Seyfarth and Smuts, 1986; Jolly, 1966; Small, 1990) according to which primate intelligence evolved mainly to solve social (and political!) problems, and was only later extended to solve problems outside the domain of affiliation, dominance and social relationships. For behavior to remain adaptive within the social environment, additional memory space was required in order to register the many significant acquaintances, as well as the rapidly growing number of possible relationships among them: kinship, friendships, alliances, hierarchies... Not only was there a requirement for additional memory, but new interactive skills and strategies also had to be mastered. In nonhuman primates, this acquisition is facilitated through the frequent activity of mutual grooming. But grooming takes time, and within the early hominid tribes, there hardly could have been enough grooming time to meet the demands of the increased group size. This leads the author to suggest that the "language device" might well have developed to serve in the acquisition of the complex social knowledge and intricacies concerning a whole network of individuals. By way of verification, the author analysed the topics of conversation in naturally formed groups. He found that from 65% to 75% of the information exchanged actually dealt with personal relationships, personal experiences or planning future social activities, a kind of information somewhat similar to that which grooming enables individuals to obtain within primate groups. From what has been said before, it seems likely that the very first linguistic signals would have emerged and been differentiated from the more primitive emotional signals already available. As a matter of fact, expressive speech acts that communicate emotional reactions, such as interjections, are thought to be the oldest forms of words in any languages; they are also the first ones to be acquired in chilhood, the last words to be lost in severe aphasia, before communication resorts to mere gestures, and they are the first ones to reappear during the recovery phase.

2.2.4.3. Distributed artificial intelligence

Another related line of thought comes from distributed artificial intelligence (DAI), a new trend in research that emerged from "mainstream AI" some fifteen to twenty years ago, with a different epistemological stance based on a sociological rather than on a psychological metaphor about the mind (Gasser, 1991; Minsky, 1985). Whereas classical

AI had based its goal of designing intelligent systems on the postulate that the human individual could be taken as the basic model, the new paradigm adopted the view that most real world problems were typically worked upon, not by an isolated individual, but by a whole team of specialists interacting together, sharing or opposing partial views, sometimes cooperating and sometimes conflicting, finding their way towards a collective solution (Kornfield and Hewitt, 1981; Lenat, 1975). DAI is thus trying to incorporate rigorously into AI theorizing and practice this most efficient human strategy of distributing a given task among as many processing "agents" as needed, which could then operate in parallel and so contribute to the solution. Interestingly enough, such a view bears much resemblance to social-cognitive theories of learning and education such as Vygotsky's (1978) or Bandura's (1986), wherein knowledge is seen as an emerging by-product of the active transactions among agents with different skills and various sophistication levels.

Although this idea of a community of processors acting together to solve a complex problem might seem a seductive metaphor, it nevertheless generates its own share of difficulties, namely how to facilitate teamwork coordination, how to insure reliable communication, how to negociate cooperation and resolve conflicting views. As a matter of fact, in their critical review of DAI research, Bond and Gasser (1988) point to the following problems as the most basic ones for the whole field:

> How to [...] decompose, and allocate problems [...] among a group of intelligent agents
> How to enable agents to communicate and interact [...]
> How to ensure that agents act coherently [...] avoiding harmful interactions
> How to enable individual agents to represent [... the] knowledge of other agents [...]
> How to recognize and reconcile disparate viewpoints and conflicting intentions among a collection of agents trying to coordinate their actions [...] (Bond and Gasser, 1988, p. 10)

2.2.4.4. Control versus content

Now, what is the common characteristic of all these problems? In any case, it is surely relatively abstract since it does not seem to depend on any particular situation. In fact, *it is not information content that is involved here, but essentially control,* in the precise sense in which this term is currently used in computer science. Indeed, one has to specify, as simply and as efficiently as possible, what kind of events are to affect information flow, and what should ensue: that is, who is communicating with whom and when; which protocol is to be used so as to insure a coherent and common interpretation; how the

knowledge of other agents is to affect the interaction, how cooperation will be promoted, how conflicts will be resolved... *In other words, all of the problems listed above basically have to do with the design of an appropriate control structure, and more precisely with distributed control.* Clearly, specifying a proper control scheme has always provided a path for the emergence and the evolution of new fundamental concepts in computer science (Winston, 1984). The declarative/procedural controversy raised in the early seventies is a good illustration of this. The real question was whether control structure should be merged with data structure, or whether it should be kept separate and operate upon data from the outside. And we have seen how profitable such concepts have proved for cognitive psychology and for education. It now seems more and more obvious that collective interaction lies at the core of complex knowledge processing and problem solving. But this calls for the design of simple yet powerful distributed control structures.

Our contention is that the time might be ripe for the conceptualization of a new control scheme based upon the metaphor of emotional processing, and that the new concepts thus generated will have a serious impact on many of the cognitive and social sciences, including education. The next section spells out some of the specifications that seem mandatory, if the modelling of emotions is to remain properly grounded in the more primitive layer of physiological needs, yet serve as a coherent foundation for the modelling of more complex interactions, including the ones that should secure the acquisition of knowledge and of higher cognitive processes. One basic tenet is that adaptive goal-oriented behavior has to do with motivation, and that motivation has in turn to do with resource management. Another important argument is that the concept of commitment acts as a critical resource and plays a fundamental role in regulating the complex transactions evidenced in higher social animals, at least within individuals of species that manifest nurturing and cooperating behavior, and that have, for this purpose, evolved emotional control structures.

2.2.5. Sketch of a model: emotions as commitment operators

If one is to design artificial systems that are to evidence motivational or emotional capacities, what requirements are these beings likely to be faced with, and what kind of ontology would be apt to meet such requirements? Oatley (1992, p. 31-36) suggests that emotions might have developed as designed solutions for individuals of species faced with the challenge of "augmented planning". This problematic augmentation is threefold:

more complex species have to deal with multiple goals to be handled in parallel, which increases the risks of conflicts; yet these organisms could only count on imperfect knowledge and limited resources; moreover these individuals generally have to interact with many other agents, which calls for efficient mechanisms insuring communication, coordination and cooperation. Hewitt (1985, 1991) likewise specifies the essential requirement for adaptive agents of having to cope with "open systems", wherein various unforeseen events could happen at any time and so hinder current planning. In such cases, provisions for solutions have to be generated quickly enough, often in spite of many interacting and conflicting agents. Hewitt also stresses that agents have but local knowledge and bounded scope, so they are typically limited to "arm's length" relationships.

On the other hand, these natural agents are to be "evolvable" creatures, submitted to the "subsumption principle" (Brooks, 1991). This means that they should not have been faced with too abrupt changes or global reorganization in their design, such that later acquisitions should "subsume" the older ones. Their real-time reactive capabilities thus have to result from the operation of successive layers of control, an argument in favor of the functional additivity of those interconnected layers, such that, at each level, the design proves adaptive in itself, but that every successive layer adds new functionalities without disrupting the preceding ones. For our purpose, the subsumption principle points to the idea that emotional structures likely developed on top of simpler existing motivational structures, and evolved so as to enhance and extend further the existing functionalities of the older layer.

An additionnal requirement for agents immersed in dynamically changing open systems, is that they should display a fair amount of autonomy (McFarland, 1995; Steels, 1995). But mere "executive autonomy" (Castelfranchi, 1994) whereby one is only free to execute whatever command comes from outside, is clearly not sufficient here. Real autonomous agents should be self-controlling, they should have the capability for self-determination of goals. This means that, more often than not, they will behave in ways that will be primarily beneficial to themselves, with the consequence that they cannot always be assumed to be benevolent, nor even sincere and trustworthy. In pursuit of their own goals, they might also have to influence other agents of their kind. This implies that, by virtue of sharing a common ontology, such agents should themselves be liable to

dependence and influencing. So the stakes are fairly high, and the challenge of conceptualizing agents that are to meet such requirements is no easy task.

A basic premise of our work is that the underpinning of motivation provides one essential line of attack for this design problem. In specifying "a principled theory of agency and autonomy", Luck and d'Inverno (1995) defined autonomous agents as progressive refinements within a hierarchy of computational entities. They first defined *objects* as generic entities with attributes and capabilities, pretty much along the lines these concepts have been defined in object-oriented languages (Goldberg and Robson, 1983; Stefik and Bobrow, 1986). Then *agents* are specified as forming the subclass of objects that have goals. Finally, *autonomous agents* are conceptualized as agents with motivations. Indeed, the concepts of autonomy and motivation are intimately related: if agents are ever to become autonomous, it will only be through getting control over their motivations, by having ways of setting their own goals. But what is motivation about in the first place, where do goals ultimately come from? Our claim is that *motivation basically has to do with managing resources*, a task which is fundamental for the agents, and which has to become their ultimate goal.

It is thus necessary to specify what is to be counted as a resource. In the agent context, we define resources more or less as energy is defined in physics: whatever is required for the production of work. It could be physical energy like heat or electricity or food, but it could also be other things: time, money, affiliation, power, knowledge... For instance, various forms of affiliation, such as kinship or friendship, typically refer to other agents that one could count on in distressful situations, in order to get access to a needed resource. Likewise, power over other agents is a way of harnessing their resources to benefit one's own, and could as such be counted as a proper resource. For educational concerns, it is a welcome fact that knowledge itself is also to be considered as a very precious kind of resource. To illustrate the point, one merely has to imagine that a robot has to get to a given location, an effort which should cost it some amount of physical energy. But if the agent learns about a shorter route than usual to that location, this specific knowledge is clearly worth some economy in fuel, it could hence be counted as a resource, and even be quantitatively measured in terms of the energy saved.

Now, it seems useful to refine the concept and to distinguish at least two different kinds of resources: first-order and second-order. First-order resources are ones that an agent can

access directly by himself. Second-order resources, on the other hand, are those that an agent gets only indirectly, by way of other agents. For instance, in nurturing species like birds or mammals, newborns can get access to food only through having their parents provide the resources for them. Hence, the basic distinction between first-order and second-order resources, is that the latter rest essentially upon a mediating agent. More specifically, it is not the agent per se that is the resource, but whatever makes that one can count on him: namely, *the commitment is the resource!* It should be clear though that commitments are not just "potential" resources. Some food out of reach, that one could try to access, could be seen as merely potential, but the definition of second-order resources proposed here relies upon some other agent, that one counts on. It is intrinsically bounded with multi-agent societies (MAS). Yet, this does not mean that commitments have to be formulated in a conscious or reflective way: attachment structures, for instance, do involve commitments between caretakers and infants that are presumably triggered through the activation of wired-in releasers and partially built-in structures (Lewis and Rosenblum, 1974; Reite and Field, 1985).

Being specifically defined in terms of resources, commitments thus appear as the foundation of "sociality" (Castelfranchi, 1990), the ability for autonomous agents to enter into social relations: by regulating exchange, they enable the agents to use one another as a way to access additional resources that would otherwise remain out of reach, or too costly to obtain. Within complex species that pursue multiple goals, and who absolutely require as such the cooperation of other agents, commitments might even become the most important of all resources! One important corollary is that MAS thus require an effective control mechanism to regulate and control commitment fluctuations. Hence, it is likely that, in social species like ours, evolutionary forces would have favored and selected for precisely those organisms that would have become equipped with powerful mechanisms to insure control over commitments. This fits rather nicely with the "social intelligence hypothesis" presented in the preceding section!

Among animals, *needs*, such as hunger or thirst, are the primary motivational processes (Toates, 1986) that insure management and regulation of first-order resources, be it available food or water, security or reproduction. In higher (social, or at least nurturing) animals, *emotions*, such as anger or guilt, are the secondary motivational processes that insure management and regulation of second-order resources, namely commitments (Aubé and Senteni, 1995, 1996a, 1996b). Very much like needs, emotional processes are

powerful *control structures* that act so as to protect agents and societies of agents from running out of resources. In terms of the subsumption principle, they could be conceived as an additional layer of design that developed on top of the underneath layer of needs, using much of its control power, but taking over whenever the access to resources requires the cooperation of other agents. Presumably, one should also think of an even further layer of design, exploiting the communicative and the intentional capabilities offered by the emotional layer, and responsible for reasoning, long-term planning and complex problem solving strategies. Operating much like needs, the emotional control processes are triggered by significant variations in the flow of resources, specifically second-order ones, for instance when there is a threat of breach in commitments, or an opportunity for establishing new ones. Certain events are likely to have an impact upon an agent's stock of resources. When those variations cross a critical threshold, needs are called in to regulate them. In more complex animals, other agents can supply the missing resources, provided they are committed to behave so. Emotions are designed in to regulate the operation of this additional layer.

Of course, *such a stance does restrict emotions to the social domain.* But it so happens that universal antecedents of emotions such as joy, sadness, anger, shame or guilt... typically have to do with social relations, and even that they occur most often within close relationships (Baumeister, Stillwell and Heatherton, 1994; Ekman, 1994; Lazarus, 1991; Scherer, 1988). Herbert Simon, in his paper on motivation, had already noted that: "situations involving interaction with other human beings are characteristically more heavily laden with emotions than any other situations" (Simon, 1967, p. 37). Oatley and Johnson-Laird (1987), in presenting their "communicative theory of emotions", likewise pointed that:

> Most emotions of interest to humans occur in the course of our relations with others. (Oatley and Johnson-Laird, 1987, p. 41)
> Many adult plans are mutual: [...] Such plans are among the most important that we make: marriage, parenthood, employment, friendships etc. Many of our more intense and problematic emotions concern plans where mutuality has been sought for, set up, or assumed. (Oatley and Johnson-Laird, 1987, p. 43)

These considerations would lead us to exclude from our list of emotions, for instance, certain fears such as that of humans reacting to a sudden noise. Such reactions are thought to rest upon a more primitive layer for handling needs (security being involved here). Our basic criterion for counting a motivational process as an emotion, is that it should provide

for some communicative means in the service of interagent binding and commitment (such as alarm or distress calls), even if the corresponding message is not actually sent in some particular situation. There is much experimental data from the psychophysiology of emotions that provide suggestive evidence for different origins behind the repertoire of human fears (Öhman, 1986) and for the fact that some of these behaviors rest upon much simpler and evolutionary older structures (Ekman, Friesen and Simons, 1985). It is our contention that reserving "emotions" for the control structures that are responsible for commitments management will do much to help disambiguate the term itself and specify the possible design of these structures.

Along with this social characterization of emotions, another prerequisite is that all agents belonging to the same community have been built according to the same ontology, so they cannot escape from responding to the emotional acts emitted by other members. The expression of anger, for instance, only makes sense if the recipient is constrained, through some built-in design, to react to it significantly. Otherwise, it would likely have already been dropped from the behavior repertoire, through natural selection. Hence, one should envision, for every category of emotions, some built-in detector that would be triggered if the appropriate reaction is mandatory, and some list of preferred actions towards the target agent, from which the system should choose. A more detailed account of the computational specifications of the model, has been presented elsewhere (Aubé and Senteni, 1996a, 1996b), specifying how significant variations in the "stock" of commitments would act as signals to trigger the various emotions, and which set of actions are thereby selected so as to regain control over the perturbation.

2.2.6. Conclusion

The present paper has argued for the relevance of computational design-based theories of motivation and emotions as a much needed foundation for cognitive and social sciences, including education. We first recalled how insightful AI models have been in generating powerful concepts about problem solving and knowledge representation, which in turn fostered the conceptualization of new efficient teaching strategies. Although empirical verification will always remain mandatory to sort out theories, careful abstract reflection and modelling still offer a powerful source of new ideas. Just consider, for instance, that the concept of constructivism, which surely lies at the core of cognitive science, and of most modern approaches in education, did not actually come from empirical research, but

has progressively emerged from the boisterous debates between the nativists and the empiricists that marked a few centuries of philosophical thinking (Boring, 1942). Secondly, we contrasted the current research on student motivation with that which bears on cognitive processes, and suggested that a similar approach might also prove fruitful in generating useful concepts. Our contention is that the time is ripe for facing the challenge of addressing the issue of motivation by trying to design autonomous artificial agents that would have to care for themselves and about others, if they are to adapt in complex environments, mainly populated by other agents of similar design. In consonance with the "social intelligence hypothesis" as well as with the basic postulates of distributed AI, we are even tempted to suggest that, for knowledge to be securely acquired by members of a given community, and efficiently communicated to other members in order to achieve coordination and cooperation, an educational system has to take into account and rely upon an adequate control structure of the kind we proposed.

We especially want to stress here the importance of the concept of commitment. It has been noted that student motivation has often been defined in the literature in terms of the individual's commitment to and persistence in his choices of plans and actions (Tardif, 1992, 1995; Viau, 1994). Now that particular meaning of "commitment" refers to the individualistic definition of the term, as it has been typically used in the psychological literature (Kiesler, 1971), and connotes the idea of persisting in one's own intention (Cohen and Levesque, 1990a). We have tried to show that adopting a more interactive sense of commitment, as a kind of contract - whether implicit or explicit - between two or more agents, might take us much further toward deciphering motivated behavior. We nevertheles contrast our view with the position of symbolic interactionist sociologists who see "commitment as the overall organization of an agent's participation in many settings simultaneously" (Gasser, 1991, p. 113). These theorists conceive it as an emergent phenomenon, resulting from joint courses of action. "Commitment from the social perspective is grounded in the action of many agents' activities taken together - it is not a matter of individual choice" (Gasser, 1991, p. 114). Our way of integrating the individualistic and the emergent stances is through the notion of second-order resources. Individuals in species that have evolved "sociality", the ability for autonomous agents to enter into social relations (Castelfranchi, 1990), depend for their survival upon the resources they can get from the others. Although they are bound to persist in their individualistic interests and intentions, the scheme of "sociality" may still work adaptively through repeated interactions, provided these individuals could resort to a powerful

regulating mechanism insuring control over each other's commitments. This is where the emotional signals and functions fit in.

Consequences of such a design for human behavior are manifold. For instance, in social psychology, it has been known for about fifty years that intragroup communication promotes cooperation in social dilemmas (Orbell, van de Kragt and Dawes, 1988), but the reason why this works has evaded explanation until recently. Social dilemmas are situations where individuals have to choose between actions that will benefit the group as a whole, and actions that will benefit them individually (Dawes, 1980). One example is to accept higher gasoline prices, or to use public transportation more frequently, in order to reduce overall pollution. What is it that happens during group discussion that so enhances the individuals' will to cooperate? The best explanation so far is that through talking, "group members made and honored commitments to cooperate" (Kerr and Kaufman-Gilliland, 1994, p. 513). This works even if the choice is taken anonymously, even if the individual does not fully realize why he has changed his behavior, provided he did take an active part in the discussion. Recent applications of speech act theory (Austin, 1962; Searle, 1969, 1979) point in the very same direction. Statements phrased in one's language do not merely assert propositions that are true or false. They essentially convey the emitter's intentions, and can only work because of the commitments (Winograd and Flores, 1986) each speaker makes, through the use of the linguistic conventions, to recognize the other's intentions. Even mere assertions commit the emitter to his sincerity about the truth of the statements made, and call in turn upon the receptor to admit the truth thus professed. Work on cooperation in game theory (Axelrod, 1984; Axelrod and Hamilton, 1981; Frank, 1988) as well as on the evolution of reciprocal altruism in animal biology (Trivers, 1971) also stresses the utility of some automatic control mechanism pertaining to the emotions for instigating the will to cooperate and for regulating ongoing behavior.

Returning to education, our analysis suggests that the proper sense of commitment underlying student motivation is very likely the interactive one. If such is the case, it might not be sufficient simply to try to act upon the students' memories or beliefs about past successes or failures, about intelligence as innate or as constructive processes, about the schooling system as truly fostering learning or as merely performing assessment, about the current task as being more or less meaningful, as being more or less achievable, or as being more or less under their control. From our standpoint, the interventions

proposed in current theories do not seem totally adequate, since they do not "interactively" commit the students towards their teacher, towards their friends or parents, or even towards society as a whole, nor do they commit the teacher, or the other actors, towards the student. Our model advocates a more intensive use of the underlying control structure of emotions to this end: anger when someone else is transgressing standards, guilt when self is defecting, joy when new alliances are created, sorrow when old alliances are lost... This does not mean mere manipulation though: we truly believe that *these are the forces that we inherited through evolution so as to harness the difficulties of getting along and learning to cooperate in a world of egoism and selfishness.* Just as a better understanding of one's own cognitive processes might foster learning through offering recourse to metacognitive strategies, in a similar way, a better understanding of the processes underlying motivation and social conduct might serve to invoque "meta-emotive" strategies, of the kind that Salovey and Mayer (1989) have termed "emotional intelligence":

> a set of skills hypothesized to contribute to the accurate appraisal and expression of emotion in oneself and others, the effective regulation of emotion in self and others, and the use of feelings to motivate, plan, and achieve in one's life. (Salovey and Mayer, 1989, p. 185)

We also think that the kind of computational analysis linking goal-oriented behavior to resource management that we have proposed here, might eventually help uncover the mechanics of such powerful means of acquisition as imitation (Dautenhahn, 1995; Meltzoff and Gopnik, 1993; Mitchell, 1987), social referencing (Klinnert et al., 1983; Trevarthen, 1984) and attachment learning (Minsky, 1985), all tools that prove so efficient in getting the child to acquire relatively quickly extremely complex skills, such as language, humour, social conventions and a whole system of values.

Chapitre 3.

Les fondements psychologiques du modèle

3.1. Présentation de l'article

Il n'est évidemment pas possible d'envisager la conception d'un modèle plausible des structures émotionnelles et de leur fonctionnement sans disposer de solides connaissances sur les données empiriques déjà accumulées à ce sujet, et sur les diverses propositions théoriques qui ont cherché à mettre un peu d'ordre dans ces observations. Un examen approfondi de la littérature contemporaine en psychologie cognitive des émotions s'est donc avéré incontournable pour l'atteinte de notre objectif. L'intention qui a présidé à la conception et à la rédaction de l'article faisant l'objet du présent chapitre, était justement d'assurer à notre modélisation des fondements psychologiques solides, appuyés sur les données empiriques recueillies dans le domaine de la psychologie - et parfois aussi de la biologie et de la sociologie - des émotions. C'est d'ailleurs l'article le plus long et le plus largement documenté de ceux produits dans le cadre de cette thèse, et la revue à laquelle nous l'avons acheminé, *Psychological Bulletin*, accepte effectivement des articles de fond de cette ampleur, surtout lorsqu'ils proposent un éclairage nouveau sur l'un ou l'autre des débats majeurs qui habitent la discipline et captent l'attention des chercheurs.

L'étendue et la diversité des recherches sur la question émotions sont cependant tellement vastes, qu'il nous a fallu opérer avec méthode pour repérer et sélectionner les résultats les plus significatifs. La stratégie que nous avons utilisée à cette fin a été d'effectuer une méta-analyse à partir des principales controverses qui ont traversé l'étude des émotions dans les deux dernières décennies, et de les exploiter comme révélateur des concepts et des résultats les plus significatifs. La lecture préalable d'ouvrages de synthèse récents (Frijda, 1986; Lazarus, 1991a; Kirouac, 1989; Oatley, 1992), de manuels de base (Lewis et Haviland, 1993; Plutchik et Kellerman, 1980, 1983, 1986) d'actes de congrès ou

d'autres recueils thématiques similaires en psychologie des émotions (Clark et Fiske, 1982; Hamilton, Bower et Frijda, 1988; Harré, 1986; Izard, Kagan et Zajonc, 1983; Scherer et Ekman, 1984; Shweder et LeVine, 1984; Stein, Leventhal et Trabasso, 1990), nous a ainsi amené à identifier les débats qui semblaient les plus fondamentaux. Nous avons complété ce tour d'horizon par un examen systématique des articles publiés dans deux des revues les plus récentes et les plus importantes de ce domaine, *Motivation and Emotion* et *Cognition and Emotion* (cette dernière n'étant effectivement éditée que depuis 1987). Cette recension nous a permis de repérer par la suite d'autres sources pertinentes pour approfondir notre analyse et notre compréhension des enjeux - souvent d'ordre épistémologique - qui se profilent derrière ces débats théoriques.

Outre l'introduction et la conclusion, l'article qui suit comporte trois sections plus substantielles, les deux premières abordant chacune des deux principales controverses qui ont occupé l'avant-scène en psychologie des émotions dans les quinze dernières années, et la troisième proposant une piste de rapprochement, permettant d'atténuer l'écart théorique entre les positions divergentes, et conduisant directement à l'élaboration de notre modèle. La première de ces trois parties traite de la controverse qui a opposé principalement Zajonc et Lazarus, (Clark et Fiske, 1982; Izard, Kagan et Zajonc, 1983; Lazarus, 1981, 1982, 1984; Leventhal et Scherer, 1987; Zajonc, 1980, 1984a, 1984b; Scherer et Ekman, 1984) durant la décennie 1980, sur l'autonomie ou la dépendance relative entre les processus émotionnels et les processus cognitifs. La position de Zajonc soutenait que les émotions opéraient de façon indépendante du fonctionnement rationnel ("Preferences need no inferences", Zajonc, 1980), et que leur activation reposait même sur des structures neuro-physiologiques spécifiques. Lazarus répliquait que, sans évaluation cognitive permettant de déterminer la signification d'un stimulus, n'importe quel événement pourrait déclencher n'importe quelle émotion à n'importe quel moment, ce qui enlèverait du même coup aux émotions tout leur pouvoir adaptatif.

Nous tâchons de montrer, qu'au fur et à mesure de son déroulement, ce débat s'est surtout enlisé dans des querelles de définitions, chacun redéfinissant à son tour de façon beaucoup trop extensive les deux termes du conflit, *émotion* et *cognition*. Nous terminons cette section en proposant d'adopter une approche fonctionnaliste, inspirée notamment de l'IA et de la Vie artificielle, et visant plutôt à spécifier en termes de *design* ce que chaque terme peut recouvrir. Nous envisageons ainsi les structures émotionnelles comme une couche de contrôle particulière, apparue entre celle, plus primitive, qui gère les besoins et

les instincts, et celle, beaucoup plus récente dans l'évolution, qui sous-tend l'analyse rationnelle.

La section suivante de l'article analyse la controverse qui a opposé les théoriciens des "émotions de base" aux théoriciens de l'appraisal (Camras, 1992; Davidson, 1992; Ekman, 1992a, 1992b; Ellsworth, 1991; Izard, 1992; Johnson-Laird et Oatley, 1992; Ortony et Turner, 1990; Panksepp, 1992; Stein et Oatley, 1992; Stein et Trabasso, 1992; Turner and Ortony, 1992; Wierzbicka, 1992). Tout aussi notoire que le précédent, ce débat a été lancé dans la communauté des chercheurs en psychologie des émotions par un article particulièrement provocateur, *What's basic about basic emotions?*, publié dans *Psychological Review* par Ortony et Turner (1990), deux adeptes de la théorie de l'appropriation. Selon les théoriciens des "émotions de base", il existerait un petit nombre de structures émotionnelles - probablement moins d'une douzaine - reposant sur l'opération de circuits neurophysiologiques sous-jacents, et suffisant à rendre compte de la diversité en même temps que de l'universalité des émotions. Leurs opposants, les théoriciens de l'appropriation, postulent plutôt que les différentes évaluations cognitives ("appraisals") effectuées en présence d'un événement - comme sa nouveauté, ou sa contrôlabilité - constituent plutôt les véritables unités de base de l'expérience émotionnelle.

Nous tâchons de montrer qu'il y a une ambiguïté fondamentale dans l'usage qui est fait du concept d'appraisal, et que les deux théories sont peut-être moins distantes l'une de l'autre qu'il n'y paraît, contribuant chacune de façon significative, mais complémentaire, à la résolution du puzzle. Ainsi, l'identification des différents appraisals permet de spécifier avec précision les circonstances et les conditions de déclenchement de chaque structure émotionnelle, dont la théorie des émotions de base permet de décrire, sous la forme de *scripts*, le fonctionnement et les modalités d'expression.

Nous profitons également de cette analyse pour repérer les candidats qui, à travers l'ensemble de la littérature que nous avons parcourue, apparaissent les plus plausibles comme mécanismes de base de la "couche de contrôle" émotionnelle. Nous tâchons ensuite de formuler, à partir des deux classes de théories examinées, une structure de déclenchement cohérente, qui puisse en même temps servir de principe taxinomique pour la catégorisation et l'identification de chaque structure émotionnelle. Dans la perspective d'élaborer un modèle computationnel des émotions, une telle opérationalisation était en effet incontournable, puisqu'il faut spécifier avec précision et exactitude le type de signal

dont l'avènement doit entraîner une interruption ou une modification du traitement en cours.

Nous terminons l'article en réaffirmant, à partir de la multiplicité des données empiriques considérées, le caractère essentiellement social et interactif de cette structure de contrôle. Cette perspective nous amène à invoquer le concept d'engagement comme principe explicatif, ce qui en fait un concept-clé pour la compréhension des structures émotionnelles, et pour la conceptualisation de structures de contrôle efficaces pour les systèmes multi-agents. Cette notion sera cependant abordée plus en profondeur au chapitre suivant.

3.2. Design for emotions: A computational stance integrating basic emotions and components of appraisal[2]

Abstract. Ever since the psychology of emotions has emerged from the larger domain of motivation, it has been populated by a host of controversies. It is argued that these debates offer a promising path along which to probe the critical phenomena that have to be represented and explained in a consistent theory. Two of these are analysed more thoroughly here, with the aim of specifying a computational design for emotions. To resolve the famous controversy between Lazarus and Zajonc, it was proposed to adopt a design-based stance borrowed from Artificial Life research, in considering emotional processes as forming a specific layer, with specific functional characteristics. As a solution to the debate opposing basic emotions and appraisal theorists, it is proposed that the appraisal view has focused its research on the elicitation side of emotions, while the basic emotions stance has rather been speculating about the content of the emotional experience, through the execution of script-like encapsulated processes. The comparison of both positions brings to light their complementary contributions, and offers by the same token a workable compromise to the debate opposing the biopsychological and the social constructionist views of emotions. It is thus proposed that emotions are motivational processes whose basic function is to insure the management and regulation of second-order resources, or commitments, that bind individuals into multiagent cooperating societies. Consequences of such a view for a computational implementation are examined in the conclusion.

Acknowledgements: I am grateful to Alison d'Anglejan, Gilles Kirouac and Claude Lamontagne for their careful and critical reading of the manuscript, and their helpful comments.

3.2.1. Introduction

Since the mid-seventies when the psychology of emotions emerged from the larger domain of motivation, as a specific and well delineated field of studies (Kirouac, 1989, 1993; Strongman, 1987), it seems that it has constantly been plagued with heated and passionate controversies. To recall but a few, suffice it to mention:

- the famous *debate between Lazarus and Zajong* (Lazarus, 1981, 1982, 1984, 1991b; Zajonc, 1980, 1984a, 1984b; Zajonc and Markus, 1984), in the *American Psychologist* about the independency of cognitive and emotional processes;
- the *facial feedback hypothesis* attributable to Darwin (1872/1965), Tomkins (1962, 1963) and Izard (1971), which got reactivated by Laird (1974), was vigorously attacked by Tourangeau and Ellsworth (1979) and engendered various comments and replies (Buck, 1980; Ellsworth and Tourangeau, 1981; Hager and Ekman, 1981; Izard, 1981; Laird, 1984; Matsumoto, 1987; Tomkins, 1981; Winton, 1986);

---

[2] Texte soumis, en décembre 1996, pour publication dans *Psychological Bulletin.*

- the *biopsychological* (Gray, 1982, 1990; LeDoux, 1989, 1993; Panksepp, 1982, 1986, 1989, 1993; Rolls, 1990) *versus the social constructionist view* of emotions (Averill, 1980; Harré, 1986; Kemper, 1987; Levy, 1984a, 1984b; Lutz, 1986, 1988; Rosaldo, 1984; Solomon, 1984);
- the *dimensional versus categorical* characterization of emotions (Clore and Ortony, 1991; Etcoff and Magee, 1992; Fehr, Russell and Ward, 1982; Fehr and Russell, 1984; Fromme and O'Brien, 1982; Russell, 1979, 1980, 1991a; Russell and Bullock, 1986; Russell, Lewicka and Niit, 1989; Schlosberg, 1941, 1952, 1954; Shaver, Schwartz, Kirson and O'Connor, 1987);
- the *basic emotions controversy* raised by Ortony and Turner (1990) that led to passionate replies in the *Psychological Review* and to a whole issue of *Cognition and Emotion* (Camras, 1992; Davidson, 1992; Ekman, 1992a, 1992b; Izard, 1992; Johnson-Laird and Oatley, 1992; Panksepp, 1992; Stein and Oatley, 1992; Stein and Trabasso, 1992; Turner and Ortony, 1992; Wierzbicka, 1992);
- and more recently the *innateness-universality thesis* challenged by James A. Russell (Ekman, O'Sullivan and Matsumoto, 1991; Ekman, 1994; Izard; 1994; Russell, 1989, 1991b, 1991c, 1994).

Fortunately, such intellectual tournaments, far from weakening a growing field of interest, make for scientific vigor by stimulating further theoretical refinements as well as redirecting more carefully controlled data gathering. Indeed, such practice is quite consonant with Popper's (1979) and Kuhn's (1970) views of the evolution of scientific knowledge through refuting well formulated conjectures and establishing new paradigms. Comprehensive theories of emotions that would include most relevant aspects - from neurology, to behavior, to interactions with learning, cognitive and sociological processes, to development and pathology - are still fairly scarce. Frijda (1986, 1988), Izard (1971, 1977, 1984), Mandler (1975, 1984), Oatley and Johnson-Laird (1987, 1996; Johnson-Laird, 1988; Oatley, 1992), Plutchik (1980, 1984), Scherer (1984, 1986, 1988a, 1988b, 1992, 1993), and Tomkins (1962, 1963, 1984) would count among the daring few. But how to choose among such theoretical accounts, and how to come up with even better theories? That is precisely where the heuristic value of intellectual debates such as the ones mentioned comes in. They act as powerful lenses to scrutinize theories, they bring flaws into sharp focus, they call for the clarification of unspecified details or fuzzy conceptualizations, they enlarge the scope and help foresee the limits of a given perspective. And since different theories often share common facets, one single, well stated controversy could generate a fair amount of empirical research and foster many theoretical reformulations.

The present article will actually draw on two of these controversies in order to re-evaluate the current paradigms and to examine one that seems to be progressively emerging from

the massive data accumulated and from the reciprocal sharpening of the various arguments. The first proposal we are to consider is whether emotion and cognition result from the operation of distinct and separate systems, or whether emotion should more or less be subsumed under the broader concept of cognition. The second one questions whether any specific status should be given to some restricted subset of discrete emotions, or whether all of them could be explained more parsimoniously by specifying the proper appraisals or attributions that lie behind each emotional experience.

We have chosen these proposals because they are quite intricately related to each other, yet are also relevant to the issues underlying many other controversies in the field. If it becomes obvious, for example, that affect does require some cognitive appraisal (first question), then any proposition for the design of a discrete emotion (second question) would have to make provision for such an evaluative device. On the other hand, the quest for specific neurological substrates of emotions, which characterizes the biopsychological standpoint, together with the hypothesis of facial efference being determinant in the aetiology and the phenomenology of emotions, would argue in favor of independent processes for emotions. Furthermore, it is fairly easy to understand that the discrete emotions approach is essentially categorical, that one of the main criteria for basic emotions is precisely the innateness-universality thesis, that another criterion rests on the existence of a specific neurological circuitry which runs counter to a strict social constructionist point of view, and that a convincing argument in favor of the facial feedback hypothesis would appear congruent with the basic emotions stance. Our aim then is to re-examine carefully the various arguments behind these two leading controversies, to scrutinize the core concepts in each - namely what is specifically meant by such terms as *cognitive* or *basic* or *appraisal* - and to propose a way out through the assessment of their actual epistemological status. It is our contention that a design-based approach, aimed at proposing a computational account of the emotional processes, might contribute by identifying these elements from each side of the controversies which could be reunited and structured in such a way as to enhance understanding.

3.2.2. The cognition-emotion controversy

3.2.2.1. The Zajonc-Lazarus debate

The cognition-emotion controversy, in its most recent form, was launched by Zajonc's 1979 invited address to the APA entitled: *Feeling and Thinking, Preferences Need No*

Inferences (Zajonc, 1980), which gave way to a passionate exchange, mainly with Lazarus, during the following years (Lazarus, 1981, 1982, 1984; Zajonc, 1984a, 1984b). It should be acknowledged though, that the (almost) opposite view of emotions resulting from a cognitive interpretation of physiological arousal could be traced back to the James-Lange theory (Lange and James, 1884, 1885, 1890/1967) at the end of the last century, which had been revived by Schachter and Singer's (1962) famous experiment, and which was further given a cognitivist flavor by Mandler (1975).

Zajonc's proposal was to establish affects as resulting from processes quite independent of the cognitive ones. His arguments were as follows: affect is *primary*, in that judgments usually come first as global emotional evaluations (positive versus negative) of incoming events; affect is *basic*, automatic and effortless, requiring minimal cognitive engagement and appearing as an innate reaction in complex organisms; affective reactions are generally *inescapable*, *irrevocable* and *centered on the self;* affective reactions appear as global and instantaneous, *difficult to verbalize*, and more appropriately expressed through motor behavior and nonverbal channels; affective reactions need *not depend on cognitive* processing or on elaborate feature analysis of the incoming input; affective reactions may become *separated from content*, leaving overall impressions in memory rather than accurate records of events.

All of these reasons led Zajonc to believe that the processes resulting in the immediate and syncretic reactions called emotions had to be quite different from the slower and more analytic processes involved in the production and transformation of knowledge. An important problem though, is that the main empirical data that Zajonc drew on as confirmation did not seem too convincing, dealing essentially with prior exposure effects and primacy effects. These referred to the phenomenon of briefly presented stimuli (less than ten milliseconds) eliciting preference judgments even though no recognition occurred. One should also recall that the figures used were Chinese ideographs (Moreland and Zajonc, 1977, 1979) or random polygons (Kunst-Wilson and Zajonc, 1980), which most people would probably not consider as terribly good examples of emotion eliciting stimuli!

Lazarus (1981, 1982) replied that the reifying of emotion and cognition as separate and independant psychological systems was scientifically counterproductive, an epistemological retreat to the old days of faculty psychology. He argued that Zajonc had

adhered to a rather simplistic view of information theory where all processing had to be serial, with the cognitive evaluation resting on the output side of the sequence. Lazarus's own stance was that cognitive mediation was a necessary condition for emotion. In his view, the computation of meaning had to start at the very beginning, probably in a parallel and preattentive fashion, with many interactions along the way with sensory, perceptual and emotional processes. He also stated that evaluation did not have to be entirely conscious or deliberate, and could be produced very quickly on the basis of only partial information, before the "rational" analysis of the stimulus was entirely completed.

"If one insists that cognitive apparaisal is always a precondition to emotion", Zajonc replied "one is forced to allow cognition to be reduced to such minimal processes as the firing of the retinal cells" (Zajonc, 1984, p. 121). He then criticized his opponent of so widely broadening the definition of appraisal, that it abolished all useful distinctions between sensation, perception and cognition. Zajonc contended that Lazarus's proposition could not be falsified, since it defined emotional reaction as necessarily requiring cognitive evaluation, even when this minimal appraisal remained unconscious and unobservable. He thus strongly advocated that the debate should be resolved on empirical rather than on definitional grounds. He listed further arguments for the separability of the two systems: *phylogenetic and ontogenetic primacy* of affects that seem to be present in early infancy, across a fair range of different cultures and even in some closely related species of nonhuman primates; *separate neuroanatomical structures* confirmed through preponderant activation of the limbic structures during emotional episodes, through differential activity of the two hemispheres, and through specific sensory-hypothalamic projections bypassing the usual sensory-cortical pathways; *disjoint appraisal and affect* evidenced through differences in primacy and recency effects - or in recognition and recall - for material loaded on affective content; *affective reactions without appraisal* as in conditioning to aversive stimulus obtained under anesthesia, or in autonomic discrimination without awareness of previously seen nonsense syllables; *affective states induced through noncognitive procedures,* such as drugs, hormones, electrical stimulation of the brain or even posed facial expression through precise muscular contractions, but without the subjects being explicitely informed of the actual emotions associated with each pattern.

Lazarus's (1984) answer was twofold: first, Zajonc had misunderstood Lazarus's epistemological stance as to what he meant by cognition and emotion respectively; second, the additional supporting evidence presented for the case of separability was not as

conclusive as might have appeared. He pointed out that Zajonc himself had never given a clear definition of emotion, and had loosely subsumed under the term a large variety of diverse phenomena. As an example, Lazarus contrasted this vagueness with the thorough analysis of the startle reaction that had enabled Ekman and his colleagues (Ekman, Friesen and Simons, then submitted, but later published in 1985) to classify it as a reflex and that made it clearly distinguishable from typical emotions. He also raised similar questions as to whether sensory preferences could seriously be regarded as emotions.

As a matter of fact, this question of preference judgments was further investigated at the time (Bonanno and Stillings, 1986; Seamon, Brody and Kauff, 1983a, 1983b), and appeared to be rather a matter of familiarity, attribuable to the fact that some amount of processing had already occurred. This activity increased the likelihood of the elements being selected, even though not enough features had been processed to enable full recognition. The phenomenon also proved sensitive to judgments about some other physical characteristics of the stimuli, such as brightness and darkness, even without recognition being obtained (Mandler, Nakamura and Van Zandt, 1987). Although working from a neurophysiological perspective, Squire (1992) recently compared those judgments of preference and familiarity to the phenomenon of *priming*, which refers to an increased facility for detecting stimuli as a result of prior presentation. Squire went on to suggest that repetition priming "occurs as changes in early-stage perceptual processing systems in posterior cortex, before conceptual or semantic analysis is carried out" (Squire, 1992, p. 212) and he actually claimed to have obtained direct evidence, using positron emission tomography, for the relevance of posterior cortex activity to this phenomenon.

Hence Lazarus appeared somewhat justified in being skeptical about loosely delineated terms or phenomena invoked as arguments. He rather advocated a precisely formulated definition of emotion that should include not only the physiological arousal, the phenomenological and the behavioral expressive components, but the appraisal one as well. He just could not figure out how any emotional reaction could even be triggered without the organism having beforehand evaluated the significance of the stimulus. Otherwise, any event could induce any emotion at any moment, and not much adaptive function could convincingly be ascribed to the affective processes.

3.2.2.2. What is meant by "cognitive"?

The rebuttals went on, bouncing back and forth, without any of the opponents being much convinced by the other's arguments. It should be clear by now that there was epistemological truth on both sides and that the crux of the debate rested essentially upon the different meanings Lazarus and Zajonc attributed respectively to the term "cognition" (see Scherer, 1993, pp. 326-329, and Leventhal and Scherer, 1987, pp. 5-8, for a related view). Were these definitions so incompatible, yet were they even appropriate? Lazarus himself, in a more recent article (1991b), alluded that "there has been a tendency to overextend cognition and treat it as equivalent to mind" (Lazarus, 1991b, p. 352). Obviously, with such an extension, all mental phenomena qualify as cognitions and no further understanding is gained through the use of the word. Unfortunately, many other terms still appear unduly equated with cognition, and as a consequence much conceptual dissipation still prevails.

Cognition versus information processing. In a paper aimed at specifying a diversity of sources for emotion activation, Izard (1993) raises the point that cognition has misleadingly been equated with information processing. Indeed, the popularity of the "cognitive revolution" in psychology has almost reached the status of "intellectual imperialism" (Zajonc, 1984b, p. 145) whose effect is more or less to "cannibalize" (Kirouac, 1993, p. 47) various other conceptualizations such as perception or emotion. LeDoux (1989) had actually made similar comments as to the use of computation:

> The meaning of the stimulus is not given in the physical characteristics of the stimulus but instead is determined by computations performed by the brain. As computation is the benchmark of the cognitive, the computation of affective significance could be considered a cognitive process. If the computation of stimulus significance is a cognitive process, then emotion, we must conclude, is just a class of cognition. (LeDoux, 1989, pp. 271-272)

Izard (1993) urges extreme caution in using so loose a criterion to establish a distinction between cognition and emotion, since computations occur everywhere in the organism, from the operation of DNA in genetic processing, to the neural processing within a single axon, to medullar reflexes, to fixed action patterns, to imprinting, to language and higher mental processes... Izard (1993) rather advocates a restricted use of the term:

> The cognitive category includes all mental processes that depend on acquired representations, those representations derived from the learning and experience of

> the individual. Cognition, as defined in differential emotions theory [...] and in the present model of emotion activation, begins at that point on the information-processing continuum where learning or experience produces mental representations and memory sufficient to mediate comparison processes and discrimination. (Izard, 1993, p. 71)

Conscious versus unconscious. Another important source of confusions deals with conscious awareness being or not required as a component of cognition. Concepts such as pre-processing, multi-stages of processing, levels of processing (Craik and Lockhart, 1972), or the "cognitive unconscious" (Kihlstrom, 1987), as well as the work quoted above on prior exposure effects of stimuli that were not recognized (Bonanno and Stillings, 1986; Kunst-Wilson and Zajonc, 1980; Mandler et al., 1987; Moreland and Zajonc, 1977, 1979; Seamon et al., 1983a, 1983b), all point towards the fact that cognition is never all or none and rather results from an active process of elaboration out of which only some partial products ever reach consciousness.

Automatic versus deliberate processing. A related distinction refers to the appraisal of events operating in a reflex-like fashion as opposed to being under the systematic control of the will. Although Zajonc's (1980) original address had invoked this as an argument for the separability of the two systems, most emotion theorists, from either side of the controversy, would nowadays recognize, not only the likelihood, but the existence and even the theoretical necessity of some mechanism of automatic appraisal:

> Since the interval between stimulus and emotional response is sometimes extraordinarily short, the appraisal mechanism must be capable of operating with great speed. Often the appraisal is not only quick but it happens without awareness, so I must postulate that the appraisal mechanism is able to operate automatically. (Ekman, 1977, p. 58; see also Ekman, 1992b, pp. 187-189)

> Emotion signals provide a specific communication system which can invoke the actions of some processors and switch others off. It sets the whole system into an organized emotion mode without propositional data having to be evaluated by a high-level conscious operating system which would have to reason about an appropriate action. (Oatley and Johnson-Laird, 1987, p. 33)

> The sensory motor level of [emotional] processing consists of multiple components, including a set of innate expressive-motor programmes and cerebral activating systems which are stimulated automatically, i.e. without volitional effort, by a variety of external stimuli and by internal changes of state. (Leventhal and Scherer, 1987, p. 8)

> Regardless of whether one favours a cognitive, feedback, or central theory of emotion, the core of the emotional system is thus a mechanism for computing the affective significance of stimuli. As this mechanisms is the precursor to conscious emotional experience, it operates, by definition, outside of conscious awareness. (LeDoux, 1989, p. 271)

> We may arrive at the appraisal pattern through an automatic process of knowing in the Heideggerian sense, even if the process is preconscious or unconscious, or via deliberate, self-controlled, abstact cognitive analysis. (Lazarus, 1991b, p. 362)

> Genetically coded information relevant to evolutionary significant stimuli, though unavailable to consciousness, interacts with information derived through cognitive processes to determine the specific emotion response. (Izard, 1993, p.71)

Knowledge versus appraisal. Hence, concepts of conscious awareness or of deliberate self-control fall short of providing definitive criterion for distinguishing between cognition and emotion. Yet the problem got even more complicated since Lazarus, after having fought rather fiercely to establish emotion as dependent on cognition, reinstated an additional distinction within the concept of cognition itself, by contrasting knowledge and appraisal (Lazarus and Smith, 1988; Lazarus, 1991b). Knowledge consists of *cold* cognitions about the way things are and how they work, while apparaisal has to do with the significance of events for personal well-being; it consists of ***hot*** cognitions. Knowledge, in itself, is not sufficient for - and does not necessarily lead to - emotion, while appraisal is both necessary and sufficient for the elicitation of emotion.

Primary versus secondary appraisal. The latter concept is even further split up into different kinds of appraisal. Primary appraisal addresses whether the situation is relevant for the self, in being potentially beneficial or harmful (Frijda, 1986, would use the term *concern* here). Secondary appraisal, on the other hand, defines the adaptional significance; it evaluates the individual's coping potential, i.e. whether some action is required, and whether it can actually be managed.

So far, the "cognitive landscape" is quite loaded! Indeed, what comes out of this rapid overview is that cognition, as a theoretical construct, covers quite a variable range of definitions. Depending on the authors, it could encompass practically all kinds of computations, or else be simply restricted to what has been acquired and represented in memory. It should be acknowledged though, that such abusive broadening of terms did not only happen through cognitivist enthusiasm, but was encountered on the other side as well, as evidenced in the definition of emotion "as a change in the internal systems of the

invertebrate so that it is more likely to perform a particular behavior" (McGuire, 1993, p. 158). Such an encompassing viewpoint permitted this author to ascribe plain emotions to insects and to various other invertebrates such as mollusks or cephalopods. Clearly, in each case, some rationale should be specified for adopting a well delineated range across a whole spectrum of possible interpretations.

3.2.2.3. The design stance as a possible way out

Perhaps should we briefly comment here about the epistemological status of definitions, and about the usage they are put to within the process of theorizing. Unless one adopts the stance of naive realism, it should be stressed that, in the strive to "carve nature at its joints", definitions do not correspond to the small *figure* that laboriously comes out, but rather to the *chisel* that does the carving! Definitions are not simply "out there" to be "objectively" uncovered like other empirical data, they are conceptual tools that are *fabricated* beforehand to probe and dig and cut through whatever *reality* might eventually show up to be. If this vision is appropriate, the particular shaping of definitions hence depends on the usage one wants to put them to. Before getting hold of any tools, the question of what purpose they are going to serve is obviously to guide choosing (or manufacturing) these very tools.

So, what is it that we want to accomplish with definitions of cognition or emotion? One answer that most clearly emerges from the artificial intelligence field, but that also progressively permeates through closely related disciplines, like neurosciences, constructivist psychology and most of the other cognitive sciences, is *design*. That is, one wonders about how to specify appropriate workable designs, whether based on phylogeny, ontogeny, engineering, or any combinations of those, that can meet the requirements of adaptive behavior. Now, it is quite clear, from the mass of arguments that have been thrown into the cognition-emotion debate, that, almost by definition, information processing or computation operate behind any kind of possible behaviors, be they perceptual, motor, emotional or whatever. So if there were to be some purpose in postulating different concepts for these terms, we propose (as do Johnson-Laird and Oatley, 1992, pp. 204-208) that it be to specify different *layers of design*, built on top of each other, that could help understand how very complex behaviors could have evolved at all, and how they could eventually be described through some kind of reverse engineering.

The approach we have in mind here is somewhat akin to a methodology developed in Artificial Life (Brooks, 1991a, 1991b; Kirsh, 1991b; Maes, 1991). Brooks first recalls that evolution took about three billion years, from the onset of life, to reach the level of simple vertebrates, around 500 million years ago. This was the amount of time taken (required?) to design the first articulated sensing and locomoting mechanisms. But then things moved on rather fast: 125 million more years to get to the reptiles, another 125 to get to the mammals, and another 125 to get to the first primates. Brooks firmly believes that *sensori-motricity was the hard task*, and that the higher mental processes that "classical AI" still looks after, should be far more simpler to achieve, *once the ability to move and react has been captured through proper design*. To this end, he advocates a rather stringent methodology:

> We must incrementally build up the capabilities of intelligent systems, *having complete systems at each step of the way* and thus automatically ensure that the pieces and their interfaces are valid. At each step we should build complete intelligent systems that we let loose in the real world with real sensing and real action. (Brooks, 1991b, p. 140; our italics)

Some typical systems that came out were autonomous six-legged insect-like robots that wandered around the lab floor, avoiding obstacles, looking for certain "preys" to seize and fleeing at the sight of large moving objects. One example of design portrays three successive layers, the bottom one responsible for the *avoiding* behavior, the middle one enabling quasi-random *wandering* when there is nothing to flee from, and the top one looking ahead for distant empty places to *explore*. The overall behavior seemed quite well-integrated, but it is important to stress that, *on each step, the robot could already produce its proper functional behavior*, before the next layer was installed on top of it. These principles seemed sound, from an evolutionary standpoint, in that they produced progressively more complex adaptive mechanisms through simply adding and connecting further functional layers *without having to reorganize or rebuild the entire design* for each improvement.

Now, the problem of choosing and designing a layer capable of producing a coherent adaptive behavior in itself, *but which would also provide for useful platform on which to build additional and more complex ones*, is no easy task! It certainly was difficult (and lengthy) for evolution, envisioned as some kind of designer, and so it is indeed for the theory builder who tries to provide a coherent account for complex observed behaviors. What we are proposing here, so as to elucidate the emotion-cognition debate, is a similar

vision of two broad layers that appeared successively in evolution, the first one being already functional by itself before the second developed on top of it. Such a view implies that whichever term we ascribe to the top layer should be grounded and wired in the underneath layer, and that from then on the complete system should function as a coherent interconnected whole. It also implies that the top layer could not stand - and very likely, could not even have developed - by itself. As a matter of fact, due to its greater complexity, it could only have become realizable through *bootstrapping,* because of the more primitive structures provided by the underlying layer, as powerful building blocks.

Such a hierarchical proposal, though, is no newcomer within current theories of emotions. In an influential paper, Norman (1980) thus criticized the emphasis that early research on cognition had placed on "pure cognitive systems", with the result of neglecting social and motivational factors, and many critical aspects concerning animate systems. He stressed the foremost importance of the basic regulatory processes insuring nourishment, protection and reproduction in living creatures, and proposed a three-layered design as the overall structure to handle real life demands: the regulatory system, the emotional system and the cognitive system. The first one, the most primitive, would deal with simple and vital mechanisms concerned with body states, pain and pleasure, or basic needs. The last one would have grown out of the requirements of the regulatory system, to add the resources of perception, knowledge and language. And the middle structure of emotions was seen as "an interplay between the two" (Norman, 1980, p. 11; see also his figure 5 on page 12), so that a cognitive analysis from the top layer could, at times, cause the regulatory system to generate the appropriate arousal in case of emergencies, so as to alert the whole organism and insure adaptive behavior.

Yet, the oldest and most popular formulation is probably MacLean's (1993) distinction between reptilian, paleomammalian and neomammalian structures of the brain. It was MacLean who introduced the term "limbic system" in the medical and neurophysiological litterature back in 1952, following Broca's usage of "limbic lobe" for some related, although much less complicated structures. MacLean's view was that three rather well distinguishable layers of the forebrain were actually responsible for progressively more adaptive behaviors. He postulated that various evolutionary old anatomical structures - including basal ganglia, globus pallidus and the corpus striatus - were responsible for *basic animality*, i.e. the orchestration of daily sensorimotor routines required for survival. Those were present in reptiles, birds and mammals, and MacLean called them the *R*-

complex (R for reptilian). On top of those, then appeared the *limbic structures* of amygdala, hippocampus, septum and thalamocyngulate gyrus, among others. Those seemed mainly responsible for self-preservative (search for food and defense against predators) and procreative functions. The thalamocingulate division, in particular, introduced new forms of behaviors, such as audiovocal communication (as in the separation cry), maternal-offspring caring, and play. Finally, the *neocortical* structures came out as the most recent layer, offering large memory bases for representation, and, together with them (especially in prefrontal cortex), some innovative capacities for problem solving, planning and the foreseeing of future outcomes.

Buck's (1985) *primes theory* (PRImary Motivational/Emotional Systems), as another example, described three hierarchically related special-purpose systems, each associated with a particular physiological substrate. The lowest level was responsible for bodily adaptation and homeostasis, and recruited simple automatic mechanisms, from tropisms and taxes to reflexes and fixed action patterns, to primary drives such as hunger, thirst, sleep and sex. The middle one involved mechanisms of outward expression comprising primary affects such as fear, anger, joy, sadness and the like. Finally, the third one involved an internal cognitive readout that would allow for syncretic knowledge, exploratory behavior and self-regulation of expressions.

In a similar way of thinking, Oatley and Johnson-Laird (1987; Johnson-Laird, 1988) posited needs and emotions as a middle level in between the stereotyped reactions of instinctive structures and the more sophisticated behavior of propositional reasoning and problem solving. Leventhal and Scherer (1987) likewise formulated three successive levels of emotional processing. The sensory-motor level consisted mainly of innate expressive-motor programmes. The second one, the schematic level, ingrated sensory-motor processes with image-like prototypes of emotional situations to register expectations in memory about frequent encounters. The third level, the conceptual one, could activate propositionally organised memory structures formed by comparisons of various emotional episodes.

Hence, the idea of different functional layers does not appear too unsound. But it is essential to our view that the focus be put on the *design* aspect. We should then go on to specify, at least approximatively, what each layer should - or should not - encompass. We should also try to formulate, by the same token, what kind of building blocks the middle

one has to offer, and what the top one could make of these, as it further develops. Provided these precautions are kept in mind, it does not present much difficulty that the middle layer be called "emotional". In line with the various proposals listed above, we also assume it is built on top of a simpler one, practically corresponding to the level reached by Brooks' artificial insects, and more or less akin to the instinctive structures and need mechanisms described by the ethologists (Eibl-Eibesfeldt, 1975; Tinbergen, 1980; Von Uexküll, 1956). As to the denomination chosen for the top layer, "representational", in the sense of "based upon acquired representations" (Izard (1993, p. 71), would seem to offer a better term than "cognitive". Yet, it is clear that both the needs and emotions layers are also filled will all kinds of representations at various levels of complexity. On the other hand, the concepts of rationality or consciousness still remain fairly controversial, and the corresponding adjectives would not suit either. Hence, for lack of a better term, we will adopt "deliberative" here, as a provisory denomination.

3.2.2.4. Characteristics of the emotional layer

Semi-abstractness. Now, what is it that the emotional layer has to offer, beyond what is already taken charge of by the instinctive one, but without being sophisticated enough to be counted as pertaining to the deliberative? Typical instinctive structures, like fixed action patterns (FAP), are triggered in an all-or-none fashion through the operation of template-like innate releasers, that constitute fairly simple computational devices. As an example, Herring Gull chicks will peck for food upon their parents' bill provided they see the small red spot at the base of their beaks (Tinbergen, 1980, p. 53). On the other hand, they could very well starve to death, if the beak had been covered with different colors, although they would still peck on a yellow painted stick of the proper size, if only a small red spot appeared on it. *Emotional structures, as opposed to FAP, get triggered by much more abstract situations.* Anger, for instance, would be best elicited by thwarting or goal obstruction, fear would arise in situations of threat or danger, distress would generally result from loss, be it of a beloved one or of a valued object... (Boucher, 1983; Lutz, 1988; Scherer, Wallbott and Summerfield, 1986; Scherer, 1988b; Stein, Trabasso and Liwag, 1993) But what is the *proper* stimulus for goal obstruction, danger or loss? And how could their recognition be easily - and rapidly - *computed?* Still, even simple mammals like rats or mice evidence the expected emotional reactions upon situations of threat or obstruction. Hence, from a design-oriented point of view, there is room for such a level of processing, that presumably combines the outputs of many wired-in releasers

from the underneath layer, yet without having to resort to the analytical power of the slower operating data structures from the deliberative level right above. But the semi-abstract format that characterizes that level of processing makes for useful input to the more complex operating layer.

Context-sensitivity. Also, the fact that emotions could be elicited through a large variety of situations, that similar classes of events could either be interpreted as threat or challenge, as rule-transgression or danger, depending on status and context (Archer, 1976), point towards emotional structures being *context-sensitive*, computationally speaking, while FAP seem to rest on simpler *context-free* processing, based on much less elaborate memory requirements. As a possible mechanism, Ortony, Clore and Collins (1988) "assume that there is a context-sensitive emotion-specific threshold associated with each emotion so that the emotion will only be experienced if its threshold is exceeded" (Ortony et al., 1988, p. 81).

Intentionality. Another crucial characteristic of the emotional layer is that it basically deals with intentionality and goal-oriented behavior. Indeed, most cognitive theories of emotions would subscribe to the idea that emotions usually have to do with the process of goal-attainment, and arise in response to events that are likely to bear important consequences upon the individual's goals (Frijda, 1986, 1988; Johnson-Laird, 1988; Mandler, 1975, 1984; Oatley and Johnson-Laird, 1987; Oatley, 1992; Ortony et al., 1988; Stein et al., 1993; from an artificial intelligence perspective on emotions, similar ideas are found in Bates, Loyall and Reilly, 1992; Beaudoin, 1994; Beaudoin and Sloman, 1993; Dyer, 1987; Elliott, 1992; Frijda and Swagerman, 1987; Lehnert and Vine, 1987; Sloman, 1987; Sloman and Croucher, 1981). In Frijda's words:

> Much of emotional behavior is intentional behavior [...] Many emotions can be defined by an *intentional structure:* that of maintaining or changing a given kind of situation, and doing that in certain ways, with respect to certain aspects of the given situation and with regard to given kinds of object in those situations. *They can often be more closely, more appropriately, and more specifically defined in this manner than in any other.* (Frijda, 1986, p. 98; our italics)

Oatley and Johnson-Laird (1987) also consider that:

> The cognitive system adopts an emotion mode *at a significant juncture of a plan,* [...] when the evaluation (conscious or unconscious) of the likely success of a plan changes. We assume that these junctures are both distinctive and recurring,

> so that the emotional system in mammals has evolved to recognise them and to establish distinctive responses to them. (Oatley and Johnson-Laird, 1987, p. 35; our italics)

Likewise, Stein et al. (1993) state that:

> [Our] theory [of emotions] is based upon critical assumptions about *human intentionality and goal-directed action.* A fundamental assertion is that the major life task of individuals is to monitor the status of goals and preferences that lead to states of well-being. [...] In privileging intentional action as a core part of our theory, we are assuming that people have a built-in mechanism that allows them to represent *goal-action-outcome* sequences in relationship to the maintenance of goals. (Stein et al., 1993, pp. 279-281)

This idea of a built-in mechanism to easily capture intentionality is quite congenial to our own stance, and illustrates the benefit such a computational device would offer as a first-order approximation of the more complex and abstract causal structures that the deliberative layer will have to elaborate about the world. Indeed, in recalling the old debate between Hume and Kant on the nature of causation and the nature of causal attribution, Beaudoin (1994) argues that causal inference ought to be "theory-driven" rather than based on mere statistical covariations between events, because causal relations could never be logically ascertained. As a first-hand solution to this profound epistemological problem, the intentional structures embedded in emotional reactions as some kind of evolutionary adaptive guess, already provide "implicit theories to explain and predict our own and others' actions" (Oatley, 1992, p. 71). Lemerise and Dodge (1993) also make a similar interpretation about the development of anger expressions in young children:

> It is presumed that the expression of anger requires an understanding that one can act *intentionally* to cause harm. [...] Anger is first clearly expressed at the age that infants should have at least a rudimentary understanding of the means-end relation. [...] *It is also possible that situations that elicit anger may facilitate or accelerate the consolidation of causal understanding because of the salience of those situations.* (Lemerise and Dodge, 1993, p. 539; our italics).

Presumably - and fortunately - those small theories already make for powerful communicative means between infants and parents, before higher language processes step in. But this process of attributing intentions happens even within animals: for instance, if I inadvertently stumble over my sleeping dog, his typical reaction is to stare "unbelievably" at me and then come sniff at my face as if to check whether I am angry at him (that is, if I dare infer so, whether I did it "on purpose" or not). But the point should not be too

surprising, if one thinks about threat or submission displays within various species, that convey rather clearly the emitter's *intentions* of attacking, or of appeasing and retreating.

It is perhaps also worth pointing that the very first causal models that chidren will overtly use about physical phenomena pertain to *animism*, as Piaget (1926) would characterize them, in attributing intentions to natural forces. Attribution theory of emotions (Weiner, 1982, 1985; Weiner and Graham, 1984; Wong and Weiner, 1981), on the other hand, contends that people typically engage in spontaneous causal searches (asking *why* questions) when the outcome of an event appears as negative or unexpected or detrimental to goal attainment.

Communicativity. Another critical feature of the emotional layer has to do with its communicative characteristics. This is indeed such a fundamental aspect of Oatley and Johnson-Laird's theory, that they call it the *communicative theory of emotion* (Johnson-Laird, 1988; Oatley, 1992; Oatley and Johnson-Laird, 1987, 1996):

> Emotions serve two principal functions. First, they are a form of *internal communication* that sets cognitive processors into one of a small number of characteristic modes. [...] Second, emotions are a form of *external communication*, important in the adjustment of social relations. Complex emotions characteristic of adult life arise at junctures in social plans, which often concern mutual goals and a model of self. (Oatley and Johnson-Laird, 1987, p. 48; our italics)

Recall also that the middle level in Buck's primes theory (1985) dealt mainly with the mechanisms for the outward expression of basic emotions. And clearly, this criterion is at the very core of such important theories of emotion as Tomkins's (1962, 1963, 1984), Ekman's (1982, 1984, 1992a, 1992b, 1993, 1994) or Izard's (1971, 1977, 1992, 1994), especially if one concentrates upon the communicative functions of facial expressions. But perhaps one of the best insight into the powerful novelties of the emotional layer arises from MacLean's view (1993) of the evolutionary contributions associated with the thalamocingulate division of the limbic system:

> There are three cardinal forms of behavior that characterize the evolutionary transition from reptiles to mammals: (1) nursing, in conjunction with maternal care; (2) audiovocal communication for maintaining maternal-offspring contact; and (3) play. (MacLean, 1993, p. 74)

Now, those improvements - especially the first two - deal with the implementation of new communicative tools in the service of the affectional (Harlow and Harlow, 1965) or attachment system (Bowlby, 1969, 1973, 1980; Hatfield and Rapson, 1993; Kraemer, 1992; Reite and Field, 1985) that gets intricately woven between mother and child. To most researchers in the field of attachment, this phenomenon is best seen as one of *psychobiological attunement* (Field, 1985) whose function is to promote and regulate behavioral synchrony between closely related organisms. *This essentially provides for basic foundations on which further mental capabilities can develop.* Trevarthen's analysis (1984; Trevarthen, Hubley and Sheeran, 1979) of mother-child interactions amply suggest that the emotional system acts so as to punctuate the exchange and to progressively ascribe meaning and coherence to the infant's gross and uncoordinated reactions. Minsky (1985) also sees in the attachment process the foundations of a peculiar though powerful learning mechanism:

> What is the function of childhood attachment? [...] even though there are many ways a child could learn about ordinary causes and effects, there is no way for a child to construct a coherent system of values - except by basing it upon some already existing model [...] if too wide a variety of adult models were available, it would be too hard to build a coherent personality of one's own [...] It would simplify the child's task if the attachment mechanism restricted attention to only a few role models. (Minsky, 1985, p. 176)

Trevarthen (1984) contends that the newborn is already well equipped for preferential responding to a variety of stimuli characteristically human, such as the protypical pattern of the face, the smile, the odor, the voicing, the gentle vocalizations, the rocking, the patting... On the other hand, the parents' emotional structure acts so as to tune their attentional resources to the child's expression of needs and to induce some kind of locked-in reciprocal interactions (Lewis and Rosenblum, 1974). One interesting phenomenon akin to this attachment learning process is what has been called *social referencing* (Klinnert, Campos, Sorce, Emde and Svejda, 1983; Klinnert, 1984; Sorce, Emde, Campos and Klinnert, 1985):

> Social referencing concerns the tendency of a person [...] to seek out emotional information from a significant other person in the environment and to use that information to make sense of an event that is otherwise ambiguous or beyond the person's own intrinsic appraisal capabilities. (Klinnert at al., 1983, pp. 63-64)

As an illustration of the phenomenon, it is now well documented that, by the end of the first year, the human infant who is faced with a new object will tend to look at his

mother's expression as if of guidance. A joyful or an interested stance from the mother *towards the object* will generally induce further approach and exploration, while an angry or fearful (or distressed or disgusted) face will, to the contrary, result in hesitation, worry and withdrawal. Of particular interest is the fact that such referencing is not inconsiderately made use of, but mainly encountered in ambiguous and potentially threatening situations (Sorce et al., 1985). Such an observation is also congenial with attribution theory, according to which people actually start asking "why questions" when something happens that could prove detrimental to goal attainment (Weiner, 1982, 1985; Weiner and Graham, 1984; Wong and Weiner, 1981).

Interestingly, effects very similar to "social referencing" have even been recorded with lower species. It has been demonstrated, for example, that young rhesus monkeys reared in captivity and which had not yet evidenced any fear of snakes, could acquire intense and *persisting* fear (over months) *after only a short exposure* to their wild-reared parents behaving fearfully in the presence of real, toy or model snakes (Mineka, Davidson, Cook and Keir, 1984). In chimpanzees, it has been observed that the amount of interactions between juveniles and "aunts" (nonmother female adults) was correlated with the degree of the associative relationship between the mother and these female "friends" (Evans and Tomasello, 1986). The phenomenon has been compared to social referencing, since these interactions are mainly initiated by the infants, who then have to infer from their mothers' relationships with the other female adults which of them they could securely rely upon. As another example, squirrel monkeys are known to have different alarm calls: "alarm peep" for bird predators and "yapping" for terrestrial predators. Although the perception of both calls is innate, the appropriate reponse to yapping seems to require a visual object that has to be learned: "... when yapping was played back in connection with the presentation of a reference object, this object was strictly avoided, while an object presented with a control tone was thoroughly explored" (Ploog, 1986, p. 181).

Likewise, it seems that some birds, such as the European blackbird, learn "mobbing" (collective attack against predators) mainly from watching the *emotional* reactions of conspecifics (Gould and Marler, 1987). In one typical experiment, two groups of European blackbirds were placed in separate neighbouring cages, so that they could see each other, but so that each group would see a different third species. One group would see a stuffed owl, that typically elicits the mobbing reaction, while the second one would be shown a stuffed honeycreeper, a totally harmless species. The second group would

watch the mobbing behavior of the first one for a moment, and then start attacking and mobbing the inoffensive honeycreeper. This newly acquired aversion could then be passed back to the first group, if they were simply shown the same stuffed honeycreeper, while witnessing the mobbing reaction in the second group. The response is so well established that it gets passed on from generation to generation, the young learning it by simply watching their parents.

Let us recall that MacLean (1993) also included play behavior as resulting from the apparition of the limbic structures. It is interesting to notice here that play (especially "rough and tumble" play) is one of the most powerful elicitor of joy and laughter in children and of related expressions (e.g. the *relaxed open-mouth* display) in the youngs of higher primates as well (Van Hooff, 1972). But it is also believed that playing requires mastering rather subtle *metacommunicative* signals that the current behavior is not to be interpreted as it seems, but instead as something else, like "mock-fighting" (Bateson, 1955). Well, this clearly has to do with the *intentional structure* of the emotional layer. Play, in this particular sense, is a profound and powerful practising mechanism for the *communication of intentions.* It deals with conveying such messages as: "Even if I growl and uncover my fangs, you souldn't worry because *I don't really mean it* and I won't harm you". Finally, play is also a widely acknowledged means for the learning of complex motor skills, and of social skills as well.

In this section, we have argued for an approach very similar to the one used in Artificial Life, whereby one has to specify appropriate workable designs that could meet the requirements of adaptive and evolutive systems. This led us to recast the enduring problem of distinguishing between cognition and emotion, as one of capturing the corresponding behaviors in terms of the working of successive layers of design. We thus sought to identify some of the fundamental properties which characterize the emotional layer in such a way that it is revealed to be qualitatively different from the underneath instinctive layer, but at the same time can provide for a useful intermediary step towards the deliberative layer laid on top. We then showed the emotional structures as being semi-abstract, context-dependent, intentional and communicative.The design-oriented approach requires that we further specify at least some of the machinery pertaining to the emotional layer. Indeed, what processes should there be embodied so as to evidence the properties we just mentioned? This is where the second controversy mentioned in the beginning of the article should be of help. It is likely that some of the *basic emotions* that have been

hypothesized in various theories could provide good candidates for the inner building blocks that constitute the emotional layer. But since the controversy still rides on, we have to grant that there remain many theoretical difficulties with the proposal. This is what we are going to examine more precisely within the next section.

3.2.3. Basic emotions or components of appraisal?

When Ortony and Turner's (1990) charge against the "basic emotions" concept came out in the *Psychological Review*, it created much turmoil within the community of emotion theorists. There soon was a special reply section with four articles in a subsequent edition of the journal (Ekman, 1992a; Izard, 1992; Panksepp, 1992; Turner and Ortony, 1992), and a whole issue of *Cognition and Emotion* was further devoted to the controversy (Camras, 1992; Davidson, 1992; Ekman, 1992b; Johnson-Laird and Oatley, 1992; Stein and Oatley, 1992; Stein and Trabasso, 1992; Wierzbicka, 1992). As an additional consequence, Russell's (1991b, 1991c, 1994) more recent challenge to the innate-universality hypothesis could probably be seen as further ripples propagating from the same original wave. Yet, the intensity of the reaction was somewhat surprising, since there had been a related epistemological controversy that had endured for many years, opposing *categorical* and *dimensional* approaches to the definition of emotions. A few years before, Scherer had also raised a similar - although much less elaborated - criticism:

> Most discrete emotion theories (e.g., Izard, 1977; Plutchik, 1980) can be described as "palette theories" which imply a blending of basic emotions like that of a painter mixing basic colors on a palette. The alternative conceptualization that I would advocate on the basis of the component process model could be called a "kaleidoscope theory": a unique mixture of color and light resulting from the way in which the kaleidoscope pieces - the facets elements - fall or arrange themselves in a particular case. (Scherer, 1984, pp. 309-310)

In addition, it should be mentioned that Ortony himself (1987, pp. 290-291; Ortony et al., 1988) had launched similar attacks on the concept of basic emotions before, without generating such passionate replies at the time. Ortony et al. (1988, pp. 25-29), for instance, had used quite analogous arguments, citing *the very same table that was presented in the 1990 article* as evidence of much disagreement among emotions theorists, about which candidate emotions should be considered to be basic. What is it then that created such turmoil? In the following section we are going to examine in more detail the arguments invoked by Ortony and Turner. This will lead us to examine the set of - and

criteria for - likely candidates to basicness. In consonance with the design-oriented approach proposed in the preceding section, we will then consider which categories of emotions could be envisioned as dynamically interconnected computational building blocks. To this end, we will try to understand how the opposing views of basic emotions theorists and of appraisal theorists could contribute complementary pieces to the overall design.

3.2.3.1. Ortony and Turner's provocative stance

Ortony and Turner started their pleading by stressing that there was not much agreement among theorists as to *how many* emotions were considered basic, and as to *which ones* were included in the critical set. They looked at a sample of fourteen theories distributed across about a century of research on emotions, and found the number of candidates proposed ranging between two and eleven. Note that they counted only six for Frijda's choice, based on a personal communication, although in his well known book, he had actually advocated a much longer list of seventeen (Frijda, 1986, p. 88). Ortony and Turner also pointed to the variety of terms chosen for depicting supposedly identical emotions, as in *sadness* being designated by *sorrow*, *grief*, *distress* or even *panic* (Panksepp, 1982). Although the apparent disagreement might eventually be reduced by regrouping terms according to their semantic similarity, the authors still questioned how it could be ascertained that different theorists do refer to the same emotion when using different words. They then raised doubts about the inclusion of such states as *surprise* or *interest*, that they would not themselves count as emotions, but rather as cognitive accompaniments such as attention or curiosity. Their main criterion rests on their definition of emotions as valenced reactions, and they consider surprise and interest as intrinsically unvalenced and as affectively neutral.

They go on to examine what they see as the three currently advocated meanings of basicness: basic emotions as *semantically* primitive, basic emotions as *biologically* primitive, and basic emotions as *psychologically* primitive. The first viewpoint has stimulated research (Clore et al., 1987; Conway and Bekerian, 1987; Fehr and Russell, 1984; Shaver et al., 1987; Storm and Storm, 1987), about the way ordinary people categorize emotion concepts and use them to label various items such as situations, stories, reactions of others, facial expressions... Although Ortony and Turner well recognize that informative taxonomic hierarchies could derive from cluster analysis of

such data, they call for great caution before one could imply that words, issued from complex cultural transactions, do precisely reflect the actual nature of the underlying states, and that such *basic-level* concepts do refer to basic emotions. Such an inference would have to rest upon the assumption that discrete emotions are attributable to the operation of some innate neural substrate, which amounts to the second meaning of basicness, and still appears far from being obvious.

Although there had already been a few suggestive proposals about the putative circuitry underlying various emotional behaviors in the neurophysiological litterature (Gray, 1982, 1990; LeDoux, 1989; Panksepp, 1982, 1986, 1989), Ortony and Turner prefer to see these as complex *response systems*, wherein discrete emotions do not appear well delineated, but rather intricately amalgamated, as in Panksepp's *joy-expectancy* system, or in Gray's *fight-flight* system. They also allude to the fact that, although certain emotions could likely be attributed to lower animals, it would not be "equally" the case for every "basic" candidate. Fear, for instance, could certainly be traced much further down than joy within the evolutionary scale, and would, by the same token, appear "embarrassingly" *more basic* than joy. The authors then examine the major cross-cultural evidence on the facial expression of emotions (Darwin, 1872/1965; Ekman, 1973; Ekman and Friesen, 1971, 1986; Ekman, Friesen and Ellsworth, 1982a, 1982b; Izard, 1971, 1977), since the usual interpretation is that there ought to be some innate underlying structure to account for such universality. They oppose the view that *what might be biologically wired is not a well-defined package for each emotion, but rather components of them, that could be shared by many different emotional systems.* As a matter of fact, another appraisal theorist, Craig A. Smith (1989; see also Pope and Smith, 1994, and Scherer, 1986, 1992, for a similar analysis), had proposed that some of the physiological responses in emotion could be mapped onto various dimensions of appraisal:

> [...] facial expressions are organized so that individual components convey specific information about the person's emotional state, such as how the person is appraising the situation [...] For instance, it is hypothesized that the eyebrow frown in expressions of anger, fear, sorrow, and reflection represents some appraisal property shared by those emotions. (Smith, 1989, p. 340)

Ortony and Turner likewise suggested that the various components observed in the prototypical expression of anger, for instance, resulted from different aspects of the cognitive evaluations which lead to that emotional experience. Hence, the furrowed brow would reflect one's consciousness of being blocked in goal attainment. The square mouth

would manifest the inner tendency to aggress against the interfering agent. The subject's determination to solve the problem by removing the obstacle would be expressed by the compression of the lips. Finally, the raising of the upper eyelids would signal the attention devoted to the visual environment in the search for a solution. *The important point though is that each component could also be recruited as part of another emotion, depending on the corresponding partial appraisal.*

The third meaning of basicness that Ortony and Turner analysed, refers to the idea of certain emotions being regarded as psychologically primitive. According to them, such a position would imply that no basic emotion could have another one as its component, and that all complex emotions should result from the blends of more basic ones. But there seems to be more and more evidence, from studies of emotional development (Sroufe, 1979, 1984; see also Camras, 1992; Lewis, 1993a, 1995a, and Oster, Hegley and Nagel, 1992, for more recent studies along the very same line), that discrete emotions themselves do not come right from birth as functional and well integrated wholes, but rather as undifferentiated reactions of comfort or discomfort, and that even such supposedly non-decomposable emotions as anger or fear have to be progressively differentiated from the apparently more basic reaction of distress. Here again, what Ortony and Turner propose as an alternative approach, is to envision the so-called basic emotions as made of still simpler components whose nature should reflect the underlying appraisal structure. To the physical metaphor of color blending advocated by some of the discrete emotion theorists, they opposed a chemical metaphor wherein the emotions are more or less compared to molecules made of simpler *emotionless* components: the limited set of possible appraisals.

3.2.3.2. What is meant by "basic"?

As mentioned above, the various reactions to Ortony and Turner's position seemed disproportionate, considering all the underlying controversies that populate the theoretical and empirical studies of emotions. One probable explanation for this was that the title of the article, as well as the provocative tone of the arguments, were intently meant to challenge and trivialize the opponents' view, as when the authors hinted that suscribing to a basic emotions hypothesis would amount "to adhere to an unsubstantiated and probably unsubstantiatable dogma - an air, earth, fire, and water theory of emotion" (Ortony and Turner, p. 329). Furthermore, they were attacking no less than some of the leading figures in the field, such as Plutchik, Tomkins, Ekman, Izard and Panksepp. To achieve

this, they attracted the reader's attention right from the beginning in a rather spectacular way - actually more consonant with journalistic strategy - by suggesting that the whole community of basic emotions researchers was in total confusion as to their mutual field of interest.

Just to show that such an argument is not that conclusive and could be interpreted in many different ways, we undertook a similar exercise. We started from Ortony and Turner's own table, and combined it with a very similar one that had appeared in a paper by Kemper (1987) about identifying primary emotions. So that the table would not get too encumbered, we applied the following selective criteria. Choosing Arnold's (1960) seminal contribution as a significant starting point, we only kept taxonomies that were published from 1960 on. Further, we excluded twofold classifications (e.g., pleasure vs unpleasure), since they represented a dimensional approach which appeared somewhat contradictory to the idea of identifying categories of emotions. Perhaps more arbitrarily, we also excluded one author with a psychoanalytic approach reported in Kemper's table, because the terms he used, such as tension or appetite, did not seem too consonant with the ones used by all the others, nor to reflect quite the same kind of mental states. We included Kemper's, since he subscribed to the concept of basic emotions, but not Ortony's, even though he had himself produced a list of "emotion types" (Ortony et al., 1988, pp. 15-25), since he chose to see them as deriving from the various components of appraisal.

This selection procedure left us with a set of twenty theories that are listed in Table 3.1. As to the list of emotion terms, we also applied a filtering scheme. We first grouped together terms that have usually been considered as equivalent by most theorists. *Contempt*, *guilt* and *desire* stood alone, without any synonym. There was only one other term, shown in parentheses here, that had been associated with each of: *anger* (rage), *disgust* (aversion), *shame* (shyness) and *hate* (hatred). Terror and anxiety were both included with *fear;* confidence, acceptance and attachment with *love;* astonishment, amazement and wonder with *surprise;* anticipation, curiosity and expectancy with *interest.* The last two, *joy* and *sadness*, were the ones that had the largest number of synonyms. The first was associated with happiness, enjoyment, elation and pleasure; the second one with sorrow, grief, distress, loneliness, resignation, dejection and panic (as in Panksepp's usage of the term).

Table 3.1
Basic emotions as proposed in a selection of 20 theories
(Adapted from Kemper, 1987, and from Ortony and Turner, 1990)

	Anger	Fear	Joy	Sadness	Disgust	Interest	Love	Surprise	Contempt	Shame	Guilt	Hate	Desire
Arnold, 1960	√	√		√	√		√					√	√
Ekman and Heider, 1988	√	√	√	√	√			√	√				
Emde, 1984	√	√	√	√	√	√		√		√	√		
Epstein, 1984	√	√	√	√			√						
Fehr and Russell, 1984	√	√	√	√			√						
Frijda, 1986	√	√	√	√	√	√	√	√					√
Fromme and O'Brien, 1982	√	√	√	√									
Gray, 1982	√	√	√										
Izard, 1977	√	√	√	√	√	√		√	√	√	√		
Kemper, 1987	√	√	√	√									
Malatesta and Haviland, 1982	√	√	√	√		√							
Oatley and Johnson-Laird, 1996	√	√	√	√	√		√					√	
Osgood, 1966	√	√	√	√	√	√		√					
Panksepp, 1982	√	√	√	√		√							
Plutchik, 1980	√	√	√	√	√	√	√	√					
Scott, 1980	√	√	√	√		√	√						
Shaver et al., 1987	√	√	√	√			√						
Sroufe, 1979	√	√	√										
Tomkins, 1984	√	√	√	√	√	√		√	√	√			
Trevarthen, 1984	√	√	√	√									
TOTAL	20	20	19	18	9	9	8	7	3	3	2	2	2
%	100	100	95	90	45	45	40	35	15	15	10	10	10

We classified Frijda's "confidence" with love and attachment, since the associated action tendency was "being-with", aiming at contact and interaction (Frijda, 1986, p. 88). We also included Plutchik's "acceptance" with love, since it was explicitly associated by the author with such behavioral accompaniments as "grooming, sharing and pair bonding" (Plutchik, 1980, p. 154; 1984, p. 208), with the postulated effect of strengthening affiliation. Panksepp's "panic" was presented by himself as equivalent to "crying-sadness-sorrow-grief" (Panksepp, 1982, p. 452). We included Panksepp's "expectancy" both with joy and with interest, since the author associated curiosity and exploratory behavior with it, but also a high level of pleasure, desire and excitement (Panksepp, 1982, 1986, 1989). After having applied this classification scheme, all idiosyncratic terms that were used by only one author were rejected from the table. Thus, only Osgood (1966) had referred to boredom; only Fromme and O'Brien (1982) used satisfaction and shock; only Arnold (1960) included courage, hope and despair; only Malatesta and Haviland (1982) employed expression terms such as *knitbrow* and *browflash*. Since we did not have access to Frijda's personal communication used in Ortony and Turner's table, we resorted to the longer list presented in his book (1986, p. 88). This author was also the only one to use indifference, excitement, arrogance, humility, effort, contentment and surrender (as evidenced in crying or laughter). Finally, we completed the table when more recent publications authorized additional entries. Thus we added *contempt* to Ekman's list, due to his more recent claims in this regard (Ekman and Friesen, 1986; Ekman and Heider, 1988). In a recent revision of their theory, Oatley and Johnson-Laird (1996) also added *hatred, attachment, parental love* and *sexual attraction* to their original list of five (Johnson-Laird, 1988; Oatley, 1992; Oatley and Johnson-Laird, 1987). So we inserted *hatred* in the *hate* column and fitted the other three into a single *love* compound.

The overall results seem fairly meaningful to us. As to agreement, three clusters clearly come out. The emotions listed in the four leftmost columns - anger, fear, joy and sadness - receive a clear support from about 95% of the authors cited in the table. The emotions listed in the middle - disgust, interest, love and surprise - were only mentioned by some 40% of the theorists. Finally, the five rightmost ones - contempt, shame, guilt, hate and desire - obtain but a meagre 12% of the votes. One could still argue that our criteria might have biased the pattern of results. But recall that Ortony and Turner themselves chose to resort to a personal communication for Frijda's choice, although his published list was much more elaborated, and was supported by a fairly substantiated rationale. They also decided - quite surprisingly - to include taxonomies from theories that were more akin to

the dimensional approach. And, at least in the case of Ekman, they chose not to include in their table (though they mentioned it in the text, p. 320) his most recent list of emotion terms, that had already been published at the time. Our point then is that, spectacular as their starting argument might have appeared, it was far from being terribly conclusive!

As a matter of fact, Ortony himself joined the consensus regarding the instances of emotions evidenced in the leftmost cluster of our table, referring, for example, to anger and fear as "unequivocal examples of emotions" (Ortony, 1987, p. 293). Another appraisal theorist, Klaus R. Sherer, also refers to the emotions from this cluster as "the four major fundamental emotion types" (Scherer, 1993, p. 343). Hence, there very likely existed some convincing criteria behind these categorizations, and we should now have a closer look at the arguments typically invoked to assess basicness.

3.2.3.3. Arguments in favor of basicness

Universality of expressions. Probably the strongest argument, and the most frequently mentioned, has to do with the cross-cultural universality of facial expressions associated with emotions (Ekman, 1973, 1982, 1994; Ekman and Friesen, 1971; Ekman et al., 1982a, 1982b; Ekman, Friesen, O'Sullivan, Chan, Diacoyanni-Tarlatzis, Heider, Krause, LeComte, Pitcairn, Ricci-Bitti, Scherer, Tomita and Tzavaras, 1987; Izard, 1971, 1977, 1994). In the typical experiment, subjects are read a story implying some emotional reaction on the part of the characters, and are then shown two or more photographs depicting various facial expressions. Generally those photographs have been obtained by asking other people to pose for the likely emotion they think they would had felt, had the events reported in the stories happened to them. The task is to associate the emotion presumably felt with one of the photographs. Cross-cultural data come from people of different cultures (Western vs non-Western; literate vs preliterate) having to associate the very same photographs with identical or similar emotion-evoking stories. Reciprocally, people might be asked to generate examples of events that would likely induce the same kind of expression. The usual pattern of results across replications is that recognition (or attribution) of happiness, sadness, anger, fear, disgust and surprise, occur with rather high significant accuracy. Happiness produces the greatest score, usually above 90%, and fear the lowest, sometimes under 60%. As to the high score of happiness, one should take into account that it happens to be, even cross-culturally, the emotion for which there is the least effort to control (Scherer, Wallbott, Matsumoto and Kudoh, 1988, p. 28; Scherer

and Wallbott, 1994, p. 320). As another peculiar feature of the data, there are recurring patterns of errors, fear being most frequently confused with surprise, and disgust with anger.

Etcoff and Magee (1992) devised a neat experiment to check whether subjects actually perceived facial expressions categorically. They created computer drawings from a standardized set of Ekman's photographs displaying happiness, sadness, anger, disgust, fear, surprise, and neutral expressions. Using a special averaging algorithm, the computer program then generated, for various pairs of emotions, eleven faces, evenly spaced and going from one expression to the other. Subjects would then go through the series, in random order, and had to identify each expression by pressing one of two keys. The authors found that, "for each continuum except those involving surprise, [...] faces straddling a category boundary were distinguished more accurately than faces within a category" (Etcoff and Magee, 1992, p. 235).

Ekman (1984, p. 332) had once hinted that emotional expressions should involve various other signals, including the voice as well as the face. While studying the vocal affect lexicon as it normally develops in children, Malatesta (1981) compared it with vocalizations of infants from different languages, with the vocal behavior of chidren born to deaf parents, and with the evolution of primate vocalization. This led her to suggest that there might be universal vocal signals, as is the case with facial expressions. Scherer and his collaborators (Scherer, 1986; Pittam and Scherer, 1993; Scherer, Banse, Wallbott and Goldbeck, 1991) examined the vocal cues used in emotion encoding and decoding. They found recognition accuracy averaging 60%, with sadness and anger being best recognized, and disgust barely reaching chance level. But cross-cultural research on vocal expressions of emotions is still to come.

Universality of antecedents. The ability to assign a recognizable expression to certain events also requires, as counterpart, some universality in the antecedents of discrete emotions. In other words, the same patterns of events should cross-culturally evoke or trigger the same patterns of emotions, which would actually bear witness to their functional adaptive value. The typical data come from coding and abstracting a large corpus of stories (Boucher, 1983) or of answers from widely distributed questionnaires (Scherer et al., 1986; Scherer, 1988b; Shaver et al., 1987), from ethnological observations (Lutz, 1988) or from clinical interviews with children (Stein et al., 1993).

Again, for the small set of emotions usually considered, there is fairly large agreement about the eliciting situations: social gathering with friends or relatives, or the receipt of resources would likely induce *joy*; thwarting, goal frustration or violation of cultural standards would lead to *anger*; uncertainty, threat or danger would elicit *fear*; and the loss of valued possessions, or of significant relationships, would induce *sadness.*

Innate neural circuitry. Another frequently advocated argument in favor of discrete emotions calls for some innate neural circuitry underlying each category of emotional behavior. Panksepp (1982, 1986, 1989) proposed four such systems, respectively responsible for rage, fear, panic and expectancy. Gray (1982, 1990) combined *fight* and *flight* into one single system, and also proposed the *approach* system, associated with positive emotions, and the *behavioral inhibition* system, responsible for anxiety. LeDoux (1989, 1993) argued for the amygdala as the principal structure responsible for computing the affective significance of the stimuli, especially in the case of fear conditioning. A related criterion from Ekman (1977, 1992b) is that specific emotions are triggered by some wired-in automatic appraisal mechanism. Oatley and Johnson-Laird (1996; Oatley, 1992, p. 50) also advocated the existence of a similar monitoring device that could evaluate events relevant to the current goals, and broadcast control signals to the appropriate emotional system whenever required. Even Ortony et al. (1988, p. 81), as we mentioned in the first section, proposed a similar "context-sensitive emotion-specific" threshold mechanism.

Distinctive autonomic reactions. As an additional criterion, some distinctive autonomic reactions should presumably *leak out* from the action of those underlying neural systems. Cacioppo, Klein, Berntson and Hatfield (1993) reviewed a dozen studies conducted between 1953 and 1992, aiming at establishing emotion-specific ANS activity, and using fifteen different measures: bodily tension, systolic blood pressure, facial temperature, respiration, skin conductance level, cardiac stroke volume, skin conductance responses, peripheral resistance, cardiac output, finger pulse volume, pulse transit time, body movement, heart rate, diastolic blood pressure, and finger temperature . They conclude that:

> There is little evidence for replicable autonomic differences in pairwise comparisons of the emotions on the measures of bodily tension, systolic blood pressure, facial temperature, respiration, skin conductance level, and cardiac stroke volume. [...] Too few data exist on several other measures [...] to permit us to draw strong conclusions, *leaving only heart rate, diastolic blood pressure,*

> *and finger temperature* to peruse more closely. (Cacioppo et al., 1993, p. 125; our italics)

Still, the effects did not seem that reliable for the last two: anger has indeed been shown to induce a diastolic pressure significantly higher than in any other emotion, but in only 43% of the comparisons; anger has also been significantly associated with higher finger temperature than fear, but in only 36% of the comparisons. *This leaves heart rate as the best discriminator of the emotions.* In that case, the pattern of data suggests two distinct clusters: happiness, disgust and surprise, on one side, are all associated with a lower rate than anger, fear and sadness, on the other side. No significant difference appears between candidates within each cluster, *a result that is embarrassingly more consonant with a dimensional than with a categorical approach!*

Phylogenetic continuity. Another argument in favor of basicness is phylogenetic continuity. It goes back at least to Darwin (1872/1965) and rests upon the existence of similar emotional reactions in nonhuman primates and even in lower mammals. Actually, most of the evidence about the neural circuitry mentioned above already comes from experiments with rats, cats and monkeys. There is also some evidence for similar facial expressions in other primates for fear, anger, joy and sadness (Chevalier-Skolnikoff, 1973; Redican, 1982; Van Hooff, 1972). However, no counterparts to the *surprise* expression or to the *disgust* expression have yet been reported in monkeys, although Rozin, Haidt and McCauley (1993, p. 578) consider gaping in response to distasteful foods - which has been reported in a variety of lower mammals and even in some birds - as a precursor to the disgust expression.

Ontogenetic primacy. A related argument has to do with ontogenetic primacy. Indeed, it has been thought for some time that discrete emotions emerged as structured wholes right from birth, or soon thereafter according to some maturational timetable. For instance, in one experiment Emde (1984) had asked mothers about the age of first occurrence of their infant's emotions. Between the second and the fourth months, all mothers felt that interest (mean occurrence: 1,6 month) and joy (mean: 1,8) were present. For anger, 86% reported it had appeared (mean: 1,5), 73% for surprise (mean: 1,7), 68% for fear (mean: 1,7), and 62% for distress (mean: 0,9). But the picture has been a little blurred since then, with recent research (Camras, 1992; Oster et al., 1992) revealing that even trained observers of children's expressions produce biased judgements when required to fit observations into a

priori coding formulas. As a matter of fact, when given the opportunity, they rather judge expressions of negative affects as undifferentiated distress or as blends of several negative emotions (anger, sadness, disgust).

Still there had been fairly seductive evidence coming from observations of blind children (Charlesworth and Kreutzer, 1973, for a review), and even of deaf-and-blind-born children (Eibl-Eibesfeldt, 1973), who show recognizable expressions of specific emotions such as fear, anger, happiness, sadness, and even surprise. Interestingly, although blind children could *spontaneously* express those affects, they had considerable difficulty, compared to sighted children, in acting out or posing *voluntarily* the corresponding expressions, without resorting to the imagination or recall of some eliciting events. In addition, facial activity *increases* through communication in sighted children while it *decreases* with age in the blind from the seemingly inborn expressions. Such observations led the authors to the following conclusion:

> Taken as a totality, the data on the blind are heavily in favor of Darwin's hypothesis that spontaneous expressive behaviors in the blind (and in the sighted, too, for that matter) are not dependent upon visual learning for their appearance and are, therefore, most probably a result of the operations of innate factors. (Charlesworth and Kreutzer, 1973, p. 157)

Conclusions from anatomical studies, electrical brain stimulation and brain lesions in various primate species (and even in lower mammals), likewise suggest that emotional vocal patterns are genetically determined (Ploog, 1986).

Characteristic phenomenological tone. Finally, one last criterion of basicness happens to be one of the most *pervasive* attribute of emotional states, and probably the most *elusive* one as well, that is, a characteristic phenomenological tone. Although such a criterion appears rather difficult to define "operationally", Johnson-Laird and Oatley (1989) suggest an interesting if somewhat questionable way to capture its flavor. These authors see basic emotions as *semantically unanalysable primitives,* that one has to experience by oneself as referents, in order to make sense of them:

> Basic emotion signals have no internal structure that is parsed and interpreted within the system. Hence, it follows from our theory that there is no way in which words that refer to the subjective experiences corresponding to these modes can be analysed semantically: The modes are *primitive* subjective experiences that the words denote. They are, as philosophers say, unanalysable *qualia.* If you were

> "emotion-blind" and unable to experience emotions, then you would have no idea what it was like to feel, say, sadness. (Johnson-Laird and Oatley, 1989, p. 90)

This idea led to a simple linguistic test for the identification of affective words that would denote basic emotions. If these ones are indeed unanalysable primitives, then one only has to check for decomposability of the term by using it to replace *x* in sentences such as: "I feel *x* but I don't know why". If the result *does not* make sense, as in "I feel *jealous* but I don't know why", then *x* is decomposable, for it has an analysable propositional structure, in terms of causes and effects. It could not then, according to the authors' theory, be taken as basic.

We thus have a fair number of criteria to identify potential candidates as primary emotions. But this would not resolve the controversy, since Ortony and Turner would accept most of the evidence afforded, without still being compelled to agree with the interpretation that comes with it. Certainly they disagreed with the linguistic criterion just mentioned (Ortony and Clore, 1989), arguing that this would reduce emotions to their "feeling" component, leaving aside the whole sequence that would normally include, as well, appraising the eliciting event and executing the corresponding action tendencies. They also refuted the argument about ontogeny with recent data suggesting that the so-called basic emotions do not appear in infancy as well organized wholes, but emerge progressively from undifferentiated (and hence presumably more basic) feelings of comfort or distress (see also Camras, 1992, Oster et al., 1992, and Lewis, 1993a, 1995a, for further arguments and data along this line). But Ortony and Turner would readily accept most of the other criteria. They do indeed recognize universality of expressions and of antecedents, they themselves stress the importance of automatic appraisal and favor a functional (goal-oriented) analysis of emotions, they acknowledge the neurological evidence concerning the underlying circuitry and the autonomic activity... Their basic argument is that they do not segment reality at the level of emotional "entities", but one layer underneath, at the level of appraisal components:

> We take the impressive collection of evidence on the relation between facial expressions and emotions as indicating first that emotion expressions are built up by drawing on a repertoire of biologically determined components, and second that many emotions are often, but by no means always, associated with the same limited subset of such components. (Ortony and Turner, 1990, p. 321)

Hence, the crux of the controversy between basic emotions theorists and appraisal theorists rests essentially upon the units that are chosen to explain the phenomena of emotions: basic emotions as indivisible packages, or simpler components of appraisal that can be recomposed into many diverse emotions.

3.4 How many emotions, then?

One critical consequence of the appraisal view, is that "the stimulus or event evaluation process can elicit as many different emotional states as there are distinguishable outcomes of the appraisal process" (Scherer, 1993, p. 329; see also Scherer, 1984, p. 311, and Scherer, 1988a, p. 113). Averill would even contend that "there are an indefinite number of emotions" (Averill, 1980, p. 326). Table 3.2 seeks to establish correspondence among various lists that have been analysed or proposed by current appraisal theorists. Surprisingly enough, the enumeration appears rather restrictive, and moreover, it presents striking similarities with the "candidates to basicness" collected in Table 3.1.

We compared the most recent lists proposed by six leading (groups of) appraisal theorists (Ellsworth and Smith, 1988a, 1988b; Frijda, Kuipers and ter Schure, 1989; Lazarus, 1991a; Ortony et al., 1988; Roseman, Aliki Antoniou and Jose, 1996; Scherer, 1993). Similar terms from different theories (such as liking and love, or joy and happiness) were associated and aligned in the same row. Idiosyncratic terms employed by only one researcher were rejected from the table.

Parentheses around an entry mean that this category has *already been used* within the column, but is still the label that would best correspond (according to the author's explicit association) to the entries chosen as distinct by other theories in the same row. If an author made a distinction between two or more emotions that had been merged into a single one by the others, this was signified by using "+" to join these different versions within the same cell. Over all, the only emotion terms that we did not include in our table for being idiosyncratic were: "distrust" from Frijda et al. (1989); "tranquility" and "interest" from Ellsworth and Smith (1988a, 1988b); "aesthetic emotions" from Lazarus (1991a); "happy-for" and "gloating" from Ortony et al. (1988).

Table 3.2
Lists of emotions proposed by representative appraisal theorists

Frijda et al. (1989)	Ellsworth and Smith (1988a, 1988b)	Lazarus (1991a)	Scherer (1993)	Roseman et al. (1996)	Ortony et al. (1988)	Total (agreement)
Anger	Anger	Anger	Rage + Irritation	Anger + Frustration	Anger	6
Fear	Fear	Fright	Fear + Anxiety	Fear	Fear	6
Joy	Happiness + Playfulness	Joy	Joy + Happiness	Joy	Joy + Satisfaction	6
Sadness	Sadness + Resignation	Sadness	Sadness + Despair	Sadness + Distress + Frustration	Distress + Disappointment + Fears-confirmed	6
Guilt	Guilt	Guilt	Guilt + Shame	Guilt + Shame + Regret	Shame + Remorse	6
Pride	Pride	Pride	Pride	Pride	Pride + Gratification	6
Love	Love	Love		Liking	Love + Admiration + Gratitude	5
Hope	Hope	Hope		Hope	Hope	5
Disgust	(Anger)[1]	Disgust	Disgust	Disgust	(Hate)	4
Jealousy		Jealousy			Resentment	3
(Pride)		Relief		Relief	Relief	3
(Anger)	(Anger)	(Anger)	Contempt	Contempt	Reproach	3
Surprise	Surprise			Surprise		3
Boredom	Boredom (S and E, 1985)		Boredom			3
		Compassion			Pity	2
		(Anger)		Dislike	Hate	2

[1] Ellsworth and Smith also considered *Disgust*, but chose to include it as a variant of *Anger*. This kind of merging was encountered in different authors for various emotion terms. Such occurrences were signified in the table through the use of parentheses, but were not added to the total in the rightmost column, since it would have implied counting certain emotions (e.g. anger) more than once.

We used the results presented by Frijda et al. (1989. pp. 221-222) as the starting point of our table, since their data offered the rare opportunity of relating the same affective terms to *action readiness* profiles as well as to *appraisal* profiles. Their cluster analysis revealed eleven groups of emotions according to appraisals, and also eleven groups according to action readiness, although there were slight differences between the two sets. We used the following rules to decide that a group of emotions from their data should enter our table as a single entry. First, emotions that appeared together as the intersection of two different clusters from the two sets of profiles were considered as "reliable" and deserving an entry. Thus, *[anxiety, fear, startled, distrust]* appeared together within the appraisal profiles, while *[anxiety, fear, startled, jealousy, disappointment]* appeared as a cluster within the action readiness profiles. The first three terms reappeared in both clusters, so we considered them as one valid entry which we labelled "fear". The application of this rule yielded seven additional entries labelled respectively as: anger, joy, sadness, pride, disgust, jealousy and boredom.

Second, if one term had appeared within a cluster in one set of profiles, but was found alone in the other set, it was reunited with the original cluster. This happened, for example, with *[regret, guilt, shame]* found together within the appraisal profiles, while *[shame]* was separated from the first two terms within the other set of profiles. Rule two authorized guilt, hope and love as additional entries. We finally used a kind of transitivity rule: *[surprise, moved]* appeared within the appraisal profiles, and *[fascinated, moved]* appeared within the action readiness ones; on the other hand, fascinated was part of the joy cluster in the appraisal profiles, while surprise was part of the joy cluster in the action readiness profiles. Hence we felt authorized to combine all three terms under the label "surprise". Out of the 32 terms presented to the subjects, the only one we rejected is distrust because it seemed less stable, being associated with the fear cluster within the appraisal profiles, and with the disgust group within the other set of profiles. The application of our three rules left us with a set of 12 emotions for that column. Results of the cluster analysis performed on both sets of profiles in Frijda et al.'s study, suggest merging contempt and anger within a single cluster, but also (strangely enough!) relief and pride.

We extracted the list for Ellsworth and Smith from two articles (Ellsworth and Smith, 1988a, 1988b) treating separately pleasant and unpleasant emotions. As mentioned above, we rejected *tranquility* and *interest* from their list for being too idiosyncratic. From the

result of an earlier study (Smith and Ellsworth, 1985), these authors had decided to merge both disgust and contempt with anger. They considered playfulness as a separate emotion, but other authors, such as Lazarus (1991a, p. 269) had used it as an alternative term for happiness. Resignation was also treated as different, but it corresponded closely enough to what the others termed sadness. Although different terms had also been used by other theorists as variants of sadness - such as despair, distress or disappointment - they did not correspond enough to each other so as to authorize the addition of special rows. As an example, sadness was contrasted with resignation in that the former would serve "as an appeal for help" (Ellsworth and Smith, 1988a, p. 298) while the latter would lead to accept the unavoidable, without appeal. Roseman et al. (1996), on the other hand, would associate sadness with an appetitive appraisal (rewarding absent), while distress would be seen as resulting from an aversive one (punishment present). In their analysis of unpleasant emotions, Ellsworth and Smith (1988a) had initially included relief and boredom in their list, but they dropped them from consideration since the subjects in this study almost never reported them. So we did not include relief in their list, but we kept boredom which the authors had already analysed in a previous study (Smith and Ellsworth, 1985). In their analysis of pleasant emotions, they had also included initially fear, envy, relief and sympathy. They later dropped all four from consideration, the first two for clearly being unpleasant emotions, outside the main focus of the research, and the other two for being significantly less predictable from the different appraisal dimensions than the other emotions.

Lazarus (1991a, chapters 6-7) proposed a set of twelve emotions, used the term "fright" for fear, and associated both contempt and hate with anger. He also used sympathy and pity as other terms for the emotion that he called compassion. We excluded only "aesthetic emotions" from his set, since he himself considered them to be problematic:

> No new or special emotion is implied in aesthetic emotions other than traditional ones about whose causation and process we have already explored. Therefore, there is no standard or normative *core relational theme* for aesthetic emotions, because such a theme would depend on the particular emotion generated. By the same token, there is no common set of *appraisal components* or *action tendencies*, because the emotions generated are diverse. (Lazarus, 1991a, p. 294)

Scherer explicitly proposed two variants for each of anger, fear, sadness and joy, which he called the "four fundamental emotion types" (Scherer, 1993, p.343; see also Scherer, 1986, pp. 147-148). So, for each of these four, we placed both variants within the same

cell. As a matter of fact, rage appears in the same cluster as anger in Frijda et al. (1989), and the term is also associated with irritation and anger by Lazarus (1991a) and by Ortony et al. (1988). Fear is likewise lumped with anxiety, and joy is more or less equated with happiness, in Frijda et al. (1989), in Lazarus (1991a) and in Ortony et al.(1988). Despair has also been associated with sadness by Lazarus, and with disappointment by Ortony et al. (1988). Scherer also distinguished shame and guilt, while those two are seen as members of the same cluster or as facets of the same emotion in Frijda et al. (1989), Ellsworth and Smith (1988a), Lazarus (1991a) and Ortony et al. (1988). Since these so-called "moral" or "self-conscious" emotions have been the focus of much attention recently (Tangney and Fischer, 1995), they will be analysed more thoroughly in a later section.

All seventeen emotions analysed by Roseman et al. (1996) have been kept, although some combinations of them have been placed in the same cell. Frustration lies somewhere between sadness and anger: it differs from the former by the level of control potential, and from the latter in being circumstance-caused rather than other-caused. As a matter of fact, frustration has been associated with anger in Ellsworth and Smith (1988a; see also Smith and Ellsworth, 1985, p. 833), and with disappointment in Ortony et al. (1988, p. 122). It might well be that frustration corresponds to a kind of transition between these emotions, and could thus be perceived by subjects (and theorists!) as closer to one or the other depending on minute characteristics of the situation. Tomkins (1984, p. 169) hypothesized that both sadness and anger could be caused by a steady level of stimulation, anger resulting from the highest level. In that sense, a continued, unrelieved level of stimulation leading to discomfort and distress could build up over time and then cross the threshold for anger, making of sadness an innate activator of anger. Also, when one looks at the developmental sequence, it is increasingly thought (Camras, 1992; Lemerise and Dodge, 1993; Oster and al., 1992) that anger is preceded by undifferenciated distress that progessively turns into anger. In their analysis of children's and adults' appraisal process underlying emotions, Stein and Trabasso (1992) found that anger and sadness, among negative emotions, shared the most common antecedents and consequences. Anger arose from focusing one's attention on the *conditions* or the agent responsible for the failure, while sadness resulted from focusing on the *consequences*. Russell's review (1991c) of ethnological data reports that certain languages (as Luganda, Illongot and Ifaluk) do not even establish a distinction between anger and sadness. Considering all this, we did not

provide a specific entry for frustration, and made it appear with both anger and sadness in Roseman et al.'s column.

We also decided to pack into the same cell, guilt, shame and regret, that Roseman et al. (1996) had distinguished. Tangney, Miller, Flicker and Barlow (1996) found their subjects typically reporting feelings of regret (and remorse) while experiencing guilt or shame. Results from the cluster analysis of Frijda et al. (1989) had also combined these three emotions. We have already mentioned that the first two, shame and guilt, have been fitted into the same cluster in Frijda et al. (1989), in Ellsworth and Smith (1988a), in Lazarus (1991a), and in Ortony et al. (1988). As to regret, Roseman, Wiest and Swartz (1994), looking at the phenomenological aspects of various emotions, found that:

> In regret experiences, [...] subjects felt a sinking feeling, thought about a lost opportunity and a mistake that they had made, felt like "kicking" themselves and correcting their mistakes, actually did something differently, and wanted to get a second chance... (Roseman et al., 1994, p. 213)

This description shares with guilt feelings the idea of self-hatred, of wanting to undo one's action and to make up for one's misdeeds. One difference is that the focus is on self and mis-achievement in regret, while it is on the harm suffered by others in guilt. Another difference is that regret often results from inaction, while guilt results from wrongdoing. Yet, there seems to be a temporal pattern to this experience: actual misdeeds produce more regret in the short term, but inactions generate more regret in the long term (Gilovich and Husted Medvec, 1995). On the other hand, there is also the intriguing fact that guilt feelings are often more intense following accidental harmdoing (McGraw, 1987). For Ortony and his co-workers, "regret has to do with wishing that what was done had not been done, and is perhaps best considered as a Loss emotion..." (Ortony et al., 1988, p. 154). Hence, here again it depends on whether the focus is on the irremediable loss (sadness), or upon one's responsibility (guilt) or incapacity and unfitness (shame).

Of all the theorists considered, it is clearly with Ortony et al. (1988) that we had the most difficulty in trying to establish correspondence with the others: they proposed the largest number of emotions, and they offered a fairly strong argumentation to sustain their choices. Yet, as was convincingly argued by Frijda (1993a) about the appraisal concept in general, it is not unlikely that *some of the distinctions proposed correspond to further cognitive elaborations as the emotional episode unfolds, rather than to genuine appraisal*

actually required for the selection of a particular emotion. The idea is that presumably such elaborations do not actually make for different types of emotions, but rather enrich the content of a few candidates among them, by specifying contextual variations that color them in various tones and shades. Ortony et al. (1988) thus called disappointment the emotion of being "(displeased about) the disconfirmation of the prospect of a desirable event" (Ortony et al., 1988, p. 122). But this is a clear instance of an irrevocable loss, which is usually considered as the hallmark of sadness. As a matter of fact, naturalistic studies of disappointment situations suggest that both sadness and anger could be then experienced, and even co-occur (Levine, 1996). If one thinks in terms of the evolutionary adaptive value of specifically registering such a situation, what is really added, in terms of functional provision, by the fact that this particular loss happens to disconfirm some previous expectation? We certainly do not deny the richness of the very many cognitive elaborations that could ensue, but we simply wonder whether these really belong to the kind of fast-reactive adaptive machinery that emotions seem to have been designed for.

In consonance with these reflections, we dared recombine many of the proposed types of emotions into fewer categories. Satisfaction, the emotion felt when someone is "(pleased about) the confirmation of the prospect of a desirable event" (Ortony et al., 1988, p. 118), was combined with joy, as Ellsworth and Smith (1988b, p. 314) had likewise suggested. As mentioned above, we also combined the emotion that Ortony et al. (1988) labelled "disappointment" with distress, the generic term they used for sadness. Note that they used the term despair as another token for disappointment, just as Scherer had used it as a variant of sadness. For similar reasons, we also included with distress the emotion they called "fears-confirmed".

The authors used shame as the label for guilty feelings, and we chose to include also with it the emotion they called remorse, for which they offered "self-anger" as an alternative token (Ortony et al., 1988, p. 148). Our main reason for such merging, is that self-hatred for wrongdoing has frequently been associated with guilt through the psychological literature on emotion (Barrett, 1995, p. 27; Frijda, 1993a, p. 373; Roseman et al., 1994, p. 214). We also combined gratification with pride, since the former was associated with such tokens as "self-satisfaction" or "pleased-with oneself" (Ortony et al., 1988. p. 148). Finally, we clustered admiration and gratitude together with love, a combination which is found as well in Ellsworth and Smith (1988b, pp. 314-315). Recall that for Ortony et al. (1988), gratitude is already seen as a compound emotion resulting from the conjunction of

the eliciting conditions for both of joy and admiration. Gratitude, on the other hand, is often associated with liking. According to Roseman et al. (1996), liking is a positive emotion resulting from a pleasurable (i.e. motive-consistent) event attributable to the action of another agent, which is also consonant with the definition of gratitude found in Weiner's attributional theory of motivation (Weiner, 1982, 1985). One should be reminded as well that animal *taming* often seems to rest upon gratitude following rewards.

Ortony et al. (1988) used hate as equivalent to Roseman's dislike, and disgust as but another token for hate. They used reproach as the prototypical label for contempt, and resentment in place of what other theorists called jealousy. They also offered compassion and sympathy as other tokens for the emotion they had labelled pity. These authors refused to consider either surprise or interest as a valid instance of emotion, on the basis that neither state happens to be valenced (i.e. neither positive, nor negative). As mentioned earlier, we finally rejected from their set "happy-for" and "gloating", for being idiosyncratic, emotions that we would rather associate respectively with joy and sadness. Interestingly enough, recent cross-cultural research suggests that such feelings as "happy-for" and "sorry-for", although they appear common for Westerners, are not even registered as distinct emotions in non-Western cultures such as the Japanese (Kitayama, Markus and Matsumoto, 1995, pp. 449-450).

One conclusion to be drawn from our table, is that there are many divergent views among appraisal theorists themselves, just as was the case among basic emotions theorists. But interestingly, there is also a core set of emotions which practically everyone seems to consider as more fundamental, not so far removed actually from those proposed by basic emotions theorists. On heuristic grounds, we separated our table into two halves, with the candidate emotions for which there is almost perfect agreement in the upper part. In a forthcoming section, we will provide further arguments for redistributing many of the emotions from the bottom half (disgust, jealousy, relief, contempt, compassion, hate) into the upper one, or for bluntly eliminating others (surprise, boredom) as genuine emotions. But if there is so little difference between the list of fundamental emotions proposed by appraisal theorists and the one proposed by basic emotions theorists, one important question to elucidate is: what does the concept of appraisal encompass?

3.2.3.5. What is meant by "appraisal"?

A number of criticisms have been formulated recently about the concept of appraisal (Frijda, 1993a; Parkinson and Manstead, 1992), pointing to the fact that the term has been used in the literature with two very different meanings. In one sense, it refers to the identification of the *antecedents* or the conditions of elicitation of an emotional reaction. In another sense, it refers to the *content* of the emotional experience, and serves to characterize various contextual aspects of this experience as it develops over time. The problem then, is whether the dimensions that have been postulated theoretically, and apparently supported empirically, really have to do with the cognitive evaluations *required* for an emotion to be triggered, or whether they correspond to further elaborations and inferences that are progressively filled in as the emotional experience unfolds. The distinction is crucial and bears important consequences for both of the debates that we are dealing with in the present paper. Indeed, if the verbal reports used as proof that cognitive evaluation checks are required for emotion, correspond somewhat to cognitive accompaniments and reconstructions *after the onset* of the affective reaction, it would mean that much less cognition enters into emotion elicitation than contended by most appraisal theorists.

A related problematic view holds that components of appraisal *are* the "basic" affective units, the combination of which will suffice to recompose, according to a simple additive scheme, the various categories of emotions. Such a stance is held by Ortony and Turner (1990, pp. 321-324), and is also illustrated in statements such as the following:

> There is general agreement that the emotions, including the "basic" emotions identified by categorical theorists, can be broken down into smaller components, and that many of these components correspond to cognitive appraisals [...] Cognitive appraisal theories propose that emotions are the resultants of a set of appraisals; *what we feel is some sort of combination of appraisals.* (Ellsworth, 1991, pp. 145, 147; our italics)

The problem with such radical statements is that the features that are found useful to elicit different emotional responses are then identified with these responses themselves. Computationally speaking, it is as if the tests on which depends the activation of a procedure would be equated with the content and the structure of the procedure itself. It certainly is a risky claim! Parkinson and Manstead (1992) also raised one important methodological problem: the typical experimental paradigm used to uncover the appraisal

dimensions could have induced in the subjects the tendency to think over and elaborate further upon the content of the emotional experience that they were invited to recall from their own memory, or to interpret from reading a prototypical scenario.

As a way out of this dilemma, Frijda (1993a) suggests that a clearer distinction be made between the conditions of elicitation and the content of the emotional experience itself, a proposal which entails at least two important theoretical consequences. First, the process of emotion elicitation might be *much simpler* than considered so far, and thus result from rather elementary appraisal processes, presumably even of a prewired nature. This would explain for one thing how emotions could be triggered so rapidly in response to minute but significant changes in the environment. Second, theorists will have to fill in the differential content of the experience that distinguishes the emotions, along formats that would account for the dynamics of their unfolding, the regularities with which they manifest themselves across individuals and cultures, and the capabilities that are evidenced for these processes to integrate past experience and to be modulated from memory materials. Frijda's proposal thus helps to understand the respective focus of appraisal versus basic emotions theories. The appraisal theories would strive to specify the elicitation mechanism responsible for the activation of the appropriate emotional response along a series of critical dimensions, attuned to best capture a few recurring and vital universal antecedents. The basic emotion theories, on the other hand, would contribute to elucidate the content and the structure of the response tendencies encapsulated within a small set of well-delineated computational packages, carved and shaped through evolution so as to best face and handle the emergencies of life.

In the attempt to fulfill such a role though, there have been some clear weaknesses on the part of basic emotions theories. One of these is their being often represented as "palette theories", through the use of the color circle metaphor (Fromme and O'Brien, 1982; Plutchik, 1980, 1984). This idea originated from the work of Schlosberg (1941, 1952, 1954) who had asked subjects to make similarity judgments of facial expressions, by placing photographs side by side. The ordering that came out looked more circular than linear, with the expressions from both extremities being judged as presenting striking similarities. Comparison with the "color circle" model resulted in some far-fetched predictions, such as complex emotions emerging from the mix of basic ones, such as emotions opposite on the circle cancelling each other out, or such as distinct emotional tones being blurred with increase in intensity. Yet it did not make for much dynamism in

the representation of mental states otherwise characterized by action tendencies, goal management and behavioral expression.

Another difficulty comes from the idea of conceptualizing basic emotions as unanalysable *qualia* (Johnson-Laird and Oatley, 1989, p. 90), that have no "propositional structure" and possibly then, no identifiable causes and no intentional objects. In a recent revision of their theory (Oatley and Johnson-Laird, 1996), these authors introduced as an argument for such a conjecture some neurophysiological data according to which, during the aura of an epileptic seizure, patients frequently experience feelings which *"are free-floating, being completely unattached to any particular thing, situation or idea"* (MacLean, 1993, p. 79). Six different categories of affect have been reported as being experienced under such circumstances: anger, fear, joy, sadness, desire and affection. Oatley and Johnson-Laird argue that those primitive emotions form the basis of moods, which are considered as emotional states that could last for hours or days, and for which subjets are generally unaware of the cause that could have triggered it. But there is no strong experimental support for such an hypothesis. One piece of data comes from studies on structured diaries of emotional episodes. In one study (Oatley and Duncan, 1992), 18 out of the 285 episodes recalled (6,3%) were classified as seemingly not caused by anything in particular: five instances for happiness, three for sadness, one for anger, nine for fear, and none for disgust. In a second study (Oatley and Duncan, 1994), only 3 incidents out of 175 reported (1,7%) were thus categorized, all of them bearing on the emotion of happiness. It certainly is not terribly compelling!

Another reason to doubt such an hypothesis, comes from experiments on electrical stimulation of specific brain sites in different mammals (many detailed examples are reported in Frijda, 1986, pp. 381-386). The interesting thing is that the emotional behaviors thus elicited are generally oriented towards - or searching for - an appropriate object. Agressive reaction, for instance, will be preferably directed towards a potential prey or toward conspecifics lower in the hierarchy. Likewise, in experiments on aversive conditioning, attacks caused by the onset of a noxious stimulus are preferably directed towards a conspecific, then a stuffed animal or a contraspecific, and if none of these is present, towards any unusual object that happens to be around, such as a rubber hose or a tennis ball (Hutchinson, 1972). Oatley himself reports studies on attentional bias in emotional disorders (Broadbent and Broadbent, 1988) according to which "anxious subjects deploy their attention towards signs of danger in the environment, and these tend

to have the effect of maintaining a mood of fear" (Oatley, 1992, p. 65). As a matter of fact, in a recent revision of their theory, Oatley and Johnson-Laird (1996), do recognize that disgust necessarily has an object, and they add four more candidates to their original set of basic emotions: attachment, hatred, parental love, and sexual attraction. The authors believe that these four, although considered as basic, could nevertheless only be experienced in relation to someone or something.

All this suggests that the *process representation* of a given "basic emotion" incorporates slots for the most likely objects and variables required for its execution. Such a view could perhaps be best accomodated by a whole body of research on the categorization of emotion concepts. Proponents of this approach have argued that emotion concepts should not be defined in a classical logical way by virtue of assessing necessary and sufficient conditions, but are best envisioned as *prototypes* with fuzzy boundaries and family ressemblance between category members (Clore and Ortony, 1991; Fehr et al., 1982; Fehr and Russell, 1984; Fischer, Shaver and Carnochan, 1990; Frijda, 1993a; Russell, 1989, 1991a; Russell and Bullock, 1986; Shaver et al., 1987). Shaver et al. (1987) for instance, had 100 subjects perform a similarity sorting task on 135 nouns depicting emotional states. For each subject, a 135x135 co-occurrence matrix was constructed, with cells reflecting pairs of words that had been associated in a pile; the matrices were then added across subjects. The cluster analysis came out with a three-layered hierarchy. The *subordinate* level consisted of about two dozen piles within which words were very strongly associated. For example, a typical pile included such terms as: adoration, affection, love, fondness, liking, attraction, caring, tenderness, compassion and sentimentality. The middle level, which they called *basic level*, regrouped the bottom piles under five broader categories: *love, joy, anger, sadness* and *fear*. Finally, two *superordinate* dimensions were obtained through further merging the basic ones with respect to their average distance.

Such a classification scheme for the emotion lexicon might help reconcile the dimensional, the categorical and even the social constructionist views of emotions. The authors indeed hinted that the top level represented the most frequently evidenced *dimension* of valence (positive vs negative emotions), the middle level corresponded to the set of *basic categories* of emotions. And the bottom one reflected the larger variety of complex emotions resulting from ontogenetic development, cognitive elaboration and *cultural construction*. Cross-cultural replication of their work (Shaver, Wu and Schwartz, 1992)

suggested the inclusion of a sixth category, *shame*, found necessary to account for the Chinese data, and it seems, from recent developments in the study of self-conscious emotions (Tangney and Fischer, 1995), that some researchers have felt compelled to incorporate into the original set more of these "social" emotions as additional prototypes with basic level characteristics. According to this view, emotions would be best represented by generic *scripts* specifying the kind of *antecedents* that would typically elicit them, the kind of *responses* (physiological, experiential, cognitive, expressive and behavioral) usually appropriate to cope with the situation, and a set of self-control or *regulatory* mechanisms. Table 3.3 illustrates a simplified prototype for the emotion of hope (abstracted from Ellsworth and Smith, 1988b, pp. 328-329; Lazarus, 1991a, pp. 282-287; Lyman and Waters, 1986, pp. 31-33; Ortony et al., 1988, pp. 110-116).

Table 3.3
A simplified script for hope
(Abstracted from Ellsworth snd Smith, 1988b; Lazarus, 1991a; Lyman and Waters, 1986; Ortony, Clore and Collins, 1988)

Antecedents:
expecting some future desirable (goal congruent) but uncertain event

Response tendencies:
excitement, tension in the lungs, possibly even pain in the stomach
yearning for better, entertaining faith
signalling valued expectations to committed others
trying to dispel anxious thoughts and other negative tendencies
recall of similar situations which issued in a beneficial outcome
tendency to approach, looking for solutions or ways out of danger
persisting in one's committed action, willing to sustain effort and vigilance

Self-regulatory procedures:
trying not to be overly optimistic (so as to avoid disappointment)
looking for facts or advice to help remain realistic

Some researchers have already started to determine, from empirical data, how such prototypical templates should be filled in for the emotions of anger, fear, joy, sadness, love, shame, guilt, pride and embarrassment (Barrett, 1995; Fischer and Tangney, 1995; Fischer et al., 1990; Mascolo and Fischer, 1995; Parrot and Smith, 1991; Shaver et al., 1987; Stein et al., 1993; Tangney, 1995). These efforts cover seven of the eight emotions that fill the upper half of Table 3.2, with hope only being left out (this is why we chose to illustrate this one as the example in our Table 3.3). It is interesting to note that similar

structures have been offered by Ortony et al. (1988, chapters 5-8), specifying, for each of the 22 families of emotion that they consider: the *type specification*, some frequently encountered related *tokens*, the main *variables affecting intensity*, and some *examples*. Likewise, Lazarus (1991a, chapters 6-7) presented a similar description for the 12 emotions which he acknowledges, specifying for each of them: the *core relational theme*, the *appraisal pattern*, the presumed *action tendency*, a set of *other terms* used, the *dynamics* of typical episodes, and possible *pathological* expression of the emotion. The format of these specifications is quite similar to such structured representations as frames (Minsky, 1975, 1985), scripts (Abelson, 1981; Schank and Abelson, 1977) and schemata (Rumelhart, 1981) which have been proposed as the basis of some artificial intelligence systems, and they would certainly provide for the kind of procedural encapsulation that seems necessary in order to capture the very process nature of emotions. Not too surprisingly, most of the few existing computer simulations of emotions actually represent them in similar frame-like or script-like formats (Dyer, 1987; Elliott, 1922; Frijda and Swagerman, 1987). Scherer's expert system was itself conceptualized around similar ideas, since prototypical emotions were represented as appraisal vectors. Guessing the subject's emotion was then computed by evaluating the relative similarity (Scherer, 1993, p. 334) between those idealized standards and the input vector obtained from the set of questions asked to the subject.

3.2.4. Design for emotions: a proposal

The task of a design-oriented approach is then to identify which types or families of emotions should be taken as required subprocesses of the emotional layer postulated in the first section of the article.

3.2.4.1. The computational building blocks of the emotional layer

Fear, anger, joy and sadness. It seems mandatory to include *fear, anger, joy* and *sadness*, that are recognized by most basic emotion theorists (Table 3.1) and by most leading appraisal theorists as well (Table 3.2). As a matter of fact, among the 253 episodes spontaneously generated by subjects using Scherer's expert system (Scherer, 1993, pp. 342-343), 87,4% fell into the categories of emotions most closely corresponding to this set: grief/sadness, joy/happiness, fear/anxiety, rage/irritation. These four emotions also come out as the basic organizing clusters in many category sorting

tasks or semantic analyses of the emotion terms (Fehr and Russell, 1984; Johnson-Laird and Oatley, 1989; Osgood, 1966; Russell and Bullock, 1986; Shaver et al., 1987; Storm and Storm, 1987), words referring to them come most frequently when subjects are asked to list emotion exemplars that readily come to mind (Fehr and Russell, 1984), and those words are also the among the first ones to be acquired by young children as emotion descriptors (Bretherton and Beeghley, 1982; Storm and Storm, 1987). Moreover, they are among the few emotions that spontaneously manifest themselves during the aura of the epileptic seizure (MacLean, 1993), and all four have received reasonable support as being neurobiologically grounded (Gray, 1982, 1990; Panksepp, 1982, 1986, 1989).

Pride. Two additionnal emotions also appear as common to all the "appraisal-based" lists of Table 3.2: pride and shame/guilt. Although the first one was not even included in the list of basic emotions in Table 3.1, and shame and guilt were only marginally considered, it seems that all three are receiving increasingly more attention as *basic level* categories of emotion (Tangney and Fischer, 1995) and, as mentioned above, prototypical descriptions have already been set for them as well. We accept easily the emotion of pride as distinct and necessary, on the basis of specific appraisal and functional specification. It follows from a beneficial outcome that is attributable to the self, and is thus swiftly appraised (Ellsworth and Smith, 1988b; Roseman et al., 1996; Weiner, 1982, 1985). Developmental studies suggest that pride differentiates progressively from "joy about action-outcome contingency" to "joy about result caused by self" around 18 months, and is then typically manifested by "mastery smiles" (Mascolo and Fischer, 1995; Stipek, 1995). Although there does not seem to exist a specific facial expression, it has been frequently suggested that pride is mainly conveyed by erectness of posture (presumably in lineage with primates' ways of asserting dominance) and by search for eye contact, in confirmation of others' praise (Darwin, 1872/1965, p.263; Weisfeld and Beresford, 1982). Finally, Lazarus (1991a, pp. 271-274) points out that it serves a dual fonction by rewarding and promoting a sense of accomplishment and self-achievement, but also in encouraging and highlighting adherence to social standards (see also Barrett, 1995, pp. 42-43, and Mascolo and Fischer, 1995, p. 66, for a similar view).

Shame and/or guilt. We have mentioned above that there is accumulating evidence and arguments for including moral emotions as well into the basic set. For instance, Shaver et al. (1992), in a cross-cultural replication of their "prototype approach" to the analysis of emotion terms, had to include shame as a basic level emotion to account for the Chinese

data. There is also evidence, from developmental data, that guilt about others' upset caused by self's agressive action, or by self's negative statement, or by self's refusal to act positively, is more and more frequently observed in children between 18 and 24 months (Mascolo and Fischer, 1995; Zahn-Waxler and Kochanska, 1990). As a matter of fact, Bretherton and Beeghly (1982) also found that, by 28 month, most children had already acquired emotion descriptors for moral judgment (93% of the children observed, used positive words appropriately to connote correct behavior, and 87% of them used negative terms adequately). Although there has been no specific facial expression consistently reported for shame/guilt, apart what is shared with other unhappiness emotions such as disappointment and sadness, Ekman expects "that research on appraisal and physiology would show that they are distinct emotions" (Ekman, 1993, p. 389). Panksepp also suggested that shame and guilt might "arise from higher evolutionary elaborations of separation-distress circuitry, perhaps within frontal-cingulate areas of the brain" (Panksepp, 1992, p. 556). Finally, there are some observations of primates and even of other mammals, suggesting at least some capacity for "acknowledging" rule transgression, a recognition which typically leads to the expression of uneasiness and submissiveness. Vollmer (1977) thus reported the story of a dog, Mango, who had developed the habit of shredding newspapers and magazines, who would be regularly scolded for it, and who would then submissively "act guilty" towards her master. Once, the owner did the shredding himself, while the dog was absent. When Mango came back, she acted uneasily and submissively at the sight of the mess, as if she had done it herself! Frans de Waal (1996, p. 110) likewise reported that, in social primates, when subordinate males have succeeded in copulating while the dominant male was out of sight, they would also "act guilty" on the dominant's return, showing more avoidance and submission than if they had not had the opportunity to "cheat" on their superior.

Yet, we have reservations about considering embarrassment, regret, remorse, guilt and shame as *basically* different from each other. We certainly do not want to ignore here genuine differences that have consistently emerged from the literature on shame and guilt (Brooke, 1985; Izard, 1977, chapters 15-16; Keltner and Buswell, 1996; Lewis, 1993b, 1995b; Lindsay-Hartz, 1984; Tangney and Fischer, 1995; Wicker, Payne and Morgan, 1983). For instance, it has been reliably found that shame would happen when the individual experiences a failure in living up to his or her standard, with the resulting effect of a breach to the person's global identity. The action tendency is thus usually to hide from the others, to try escaping from the humiliation or to bluntly deny the failure. Guilt

on the other hand is associated with feeling responsible for having caused damage to (most typically, related) others. The resulting tendency is to excuse oneself, recognize responsibility and try make amend for the damage inflicted.

So there are observable differences. It is nevertheless important to question whether they call for distinguishing two categories of emotions, *relying upon two different sets of underlying processes*, or whether they could be envisioned as contextual variants resulting in different sets of action tendencies as means of coping. In comparison, it is well acknowledged that fear could activate different behavioral reactions: fleeing, expressing submission, hiding, freezing, using threat expressions, fighting, calling for help... These reactions are likely to evidence fair measurable differences, phenomenologically as well as physiologically, yet they will all be considered as enacting the same emotion, for the reason that they respond to the same pattern of appraisal (threat and danger) and fulfill the same adaptive function (security and protection). As a matter of fact, Ellsworth and Smith (1988a), Lazarus (1991a) and Ortony et al. (1988) all considered guilt and shame as variants of the same emotion, the last two researchers including embarrassment as well in the same cluster. Subjects from Frijda et al. (1989) regrouped shame, guilt and regret in the same cluster in the appraisal profiles, but separated shame from the other two in the action readiness profiles, for the specific reason that the latter motivated hiding the self. Scherer (1993) integrated embarrassment with shame, but distinguished them from guilt. Considering that an "open" evaluation check means that the corresponding appraisal is not discriminatory, there is not much difference between the two emotions: the concern relevance bears on self in shame, but on the relationship with others in guilt; and the urgency for goal significance is slightly higher in shame. In Roseman et al. (1996), the only appraisal that seemed to make a difference between the two emotions was related to what they called *characterological versus non-characterological problem source:* subjects felt shame when it was the whole nature of someone that had been revealed as inadequate, while they felt guilt when it was only a specific negative event that had been caused. Although Wicker et al. (1983) reported significant differences in the ratings of recalled experiences of shame and guilt, "in only 6 of the 68 comparisons in two studies were the two emotions significantly different at $p < .05$ *and* on the opposite sides of the neutral point. In the majority of cases the difference was one of degree but not direction" (Wicker et al., 1983, p. 38).

Our main arguments though, for compressing shame and guilt into the same emotion type, come from developmental data, assessment of individual differences and cross-cultural research. Observations of toddlers thus reveal that, when confronted with an undesirable outcome for which they could be held responsible, such as accidently breaking someone else's toy, some children will tend to show a shame-relevant pattern of response, while other ones will rather show a guilt-relevant pattern (Barrett, 1995; Barrett, Zahn-Waxler and Cole, 1993). The different behavioral attitudes appeared consistent enough, being observed as well at home, so that the former have been called "avoiders" and the latter, "amenders". These individual differences also tend to carry over life, and bear important consequences on the individuals' adaptiveness to stressful situations of social transgressions and failure of living up to moral standards. Shame-oriented behavior is seen as more detrimental and less adaptive than guilt, since it is based on avoidance or denial, instead of taking responsibility for the transgression and constructively trying to find a way out. Different measurement scales have even been developed to assess such shame-proneness or guilt-proneness (Tangney, 1990, 1991; Tangney, Burggraf and Wagner, 1995). Finally, Wallbott and Scherer (1995) have conducted cross-cultural research to relate differences in experiencing shame and guilt to certain cultural characteristics, such as: the reaction to the distribution of social power; the capacity to tolerate uncertainty; the tendency towards individualism versus collectivism; the tendency to value money and achievement versus quality of life and relationships. As an indication, Yugoslavia and Mexico would score high on collectivism, power distance, and uncertainty avoidance, while Sweden and New Zealand would score low on the same dimensions. The authors contend to have "demonstrated that some of the differences between shame and guilt experiences are related to cultural value dimensions" (Wallbott and Scherer, 1995, p.483):

> To sum up, it seems that "typical" shame experiences [...] are typical of collectivistic, high-power-distance, and high-uncertainty-avoidance cultures, whereas shame experiences in individualistic, low-power-distance, and low-uncertainty-avoidance cultures resemble the "typical" guilt pattern to a large degree. (Wallbott and Scherer, 1995, p. 481)

All this points towards shame and guilt being indeed *variants of the same core processes,* whose basic function is to regulate social interaction and to prevent deviant behavior. There is thus convincing evidence that, through socialization, these processes are ontogenetically and culturally attuned to the particular settings of social rules, conventions and standards that one is likely to be confronted with. In our opinion, the same would

hold for other variants such as embarrassment, shyness, regret, or remorse. Thus, although Parrot and Smith (1991) abstracted from data a prototype for embarrassment, they also mention that, to be complete, part of the script also had to contain elements traditionally associated with shame. Lewis (1995b) distinguishes two types of embarrassment: embarrassment elicited by self-exposure, and embarrassment associated with self evaluation, and considers the latter as "a less intense form of shame" (Lewis, 1995b, pp. 211, 215). Tangney et al. (1996) recently offered evidence that shame, guilt and embarrassment could be reliably distinguished on the basis of phenomenological features and motivation for subsequent action. Yet, they found the elicitation structure of all three emotions remarkably similar: they all occur predominantly in social contexts, most likely in the presence of people with whom one is well acquainted, and all three arise from personal failures or social transgressions that involve self-evaluation. Shame and guilt share the feelings of responsibility, regret and desires to make amends. Shame and embarrassment share the feeling of wanting to hide and reduce self-exposure, and both are accompanied by such physiological changes as blushing and increased heart rate. The authors concluded from these findings that all three emotions should probably be best considered as members of a common emotion family, with shame being "a more primitive emotion that served adaptive functions especially at earlier stages of development [...] before the cognitive capacity to experience the more differentiated feeling of guilt develops" (Tangney et al., 1996, p. 1267). Finally, many shame-prone cultures, such as the Newars of Nepal, the Indonesians, the Japanese, the Ifalukians, and the Tahitians, do not even make a lexical distinction between the two emotions (Russell, 1991, p. 430).

Love. Although most appraisal theorists recognized love or liking as an important emotion, we can see from Table 3.1 that, surprisingly, it has not been retained as a candidate by most leading basic emotions theorists, such as Tomkins, Izard, Ekman, Oatley and Johnson-Laird (in the first version of their theory), and Panksepp (although he has suggested recently that the oxyticin circuitry would constitute a much likely basis for the mediation of nurturant behaviors and maternal intent: Panksepp, 1989, p. 66; Panksepp, 1991, p. 88; Panksepp, 1992, p. 555). Still, when Fehr and Russell (1984) asked for a one-minute free listing of exemplars of emotions, love came in fourth position, with 124 subjects out of 200 mentioning it (compared to 96 for fear). Likewise Shaver et al. (1987) collected prototypicality ratings on a list of 213 emotion words, using a 4-point scale, where 4 meant "I definitively would call this an emotion". Love came out as the big winner with a rating of 3,94. It also formed a definite cluster in Storm and

Storm's (1987) study of the vocabulary of emotions. Bretherton and Beeghley (1982) reported it as one of the emotion names most frequently used by 28-month-olds. On the other hand, attachement behavior has been extensively studied by ethologists (Harlow and Harlow, 1965), by psychobiologists (Field, 1985; Kraemer, 1992; Reite and Field, 1985) and by developmental psychologists (Ainsworth, 1989; Bowlby, 1969, 1973, 1980) who all considered it as an essential cornerstone of further adaptive development. It is now believed to be grounded on a special neural circuitry, with, for instance, the oxytocin neuropeptide playing a facilitating role in the activation of maternal behavior and grooming, in the reduction of separation distress, and in promoting nurturant feelings of acceptance and social bonding (Insel, 1992; Panksepp, 1989, 1993; Pedersen, Ascher, Monroe and Prange, 1982; Reite and Field, 1985). Also recall MacLean's view (1993) that the onset of attachment behavior in phylogeny is likely attributable to the apparition and development of the limbic system. Moreover, there seem to be premilinary findings that people can indeed "distinguish facial expressions of love from expressions of joy, sadness, fear, and anger" (reported in Hatfield and Rapson, 1993, p. 601). Recently, Baumeister and Leary (1995) made a convincing claim that attachment, or "the need to belong", sould be considered a fundamental motivation. From an extensive survey of the literature, they argued that: people are eager to form social attachments and reluctant to break these bonds; a large part of cognitive and emotional processing has to do with interpersonal relationships; breaking of affectional bonds and deprivation of belongingness lead to various behavioral problems and pathologies; this need gets quickly satiated beyond a handfull of caring and fulfilling relationships, but when a major bond is broken, substitution helps to recover and is generally looked for; the need to belong appears as a universal motivation, innately grounded in the structure of the human brain. Finally, in a recent revision of their theory, Oatley and Johnson-Laird (1996) decided to add no less than three forms of attachement to their own list of basic emotions: attachment, parental love and sexual attraction. All those arguments incite us to also include in our basic set some emotion pertaining to love/attachment, perhaps though in the much simpler forms of liking (Roseman et al., 1996) or gratitude (Ortony et al., 1988; Weiner, 1982, 1985). We would also include *desire*, which appears marginally in Table 3.1, in the same cluster. This is indeed the way it is classified in Shaver et al. (1987) and in Storm and Storm (1987). Desire is associated with the passion component in Sternberg's triangular theory of love (Sernberg, 1986), the other two components being intimacy and commitment. Within Hatfield's classification of love into passionate love and

the companionate love, desire is seen as an important component of passionate love (Hatfield and Rapson, 1993).

Hope. Although five out of six among the appraisal theories represented in Table 3.2 include hope in their representative set, it is probably the emotion that has been the less studied, so little in fact that Lazarus (1991a) has even classified it as a problematic emotion. Yet, most appraisal theorists correctly predicted that subjects would regularly experience it when appraising uncertain positive outcomes, and although relevant literature on the subject is still scarce, Table 3.3 illustrates that a plausible script could nevertheless be sketched from the existing data. Scherer (1993) is the only one not to include hope in his set of emotions. But perhaps should we notice that this author does not include "certainty/uncertainty" in his list of evaluation checks, an appraisal which appears critical for hope, as well as for fear. Now it also happens that Sherer's expert system obtained "abnormally low accuracy percentages for anxiety and fear" (Scherer, 1993, p. 347), a result that led the author to think of adding the dimension of uncertainty to his own list of appraisals in the future. Finally, a rather compelling argument for considering hope as a fundamental emotion, is that its action readiness of anticipatory excitement is pretty consonant with the behavior that Panksepp (1982, 1986, 1989, 1991) attributes to what he calls the "foraging/expectancy system":

> When an animal is *expecting positive rewards,* it often literally bounds around with excitement, which appears to be a highly energized form of investigatory curiosity. [...] The behaviors are apparently designed to *seek and anticipate forthcoming events* so that the animal can optimize its opportunities to interact with desired objects. (Panksepp, 1991, pp. 83, 85; our italics)

The function thus ascribed to the expectancy system, although simpler, bears striking similarities with the ones that Lazarus (1991a, p. 283-285) associates with hope, of being a "sustainer of coping", which helps the individual to dispell negative emotional tendencies, and to stick to one's commitments, in spite of the adversity.

3.2.4.2. Synonyms, co-occuring emotions, and non-emotions

For all the above reasons, we feel it is reasonable, if not mandatory, to set our choice for the computational building blocks of the emotional layer on the eight emotions that appear in the upper half of Table 3.2. In a later section, we will sketch a fairly simple structure of appraisal to account for their elicitation. By "simple" here, we mean easily tractable, with

low computational cost and fast-operating mechanisms. As for the emotions from the bottom half, we have substantial reasons, either for considering them as closely related to certain of our candidates, or for rejecting them.

Disgust, contempt and hate. We have mentioned earlier that confusions between anger and disgust had been a recurrent pattern in the litterature on cross-cultural recognition of facial expressions (Ekman and Friesen, 1971; Ekman et al., 1987). Darwin thought that "contempt hardly differs from disgust" (Darwin (1872/1965, p. 253), with both emotions being often expressed by spitting. Oatley and Johnson-Laird (1990) felt that their treatment of disgust was less satisfactory than that for the other basic emotions, and mentioned that when people experience disgust toward a person, "the emotion is described as contempt, disdain, or hatred" (Oatley, 1992, p. 60). Izard (1977, chapter 13) also considered anger-disgust-contempt as intimately related, forming what he called the "hostility triad". Of the three, contempt was seen as the most subtle and the coldest. It is interesting to note that Scherer's model (1993, Table 3.2) also presents contempt and "cold anger" as very closely related. Indeed, if one considers that the "open" value on one dimension of appraisal renders it non-discriminatory, then the difference between contempt and "cold anger" is that the former rates "very low" instead of low on the compatibility standards, low instead of medium on the urgency and the power checks, and high instead of "very high" on the outcome probability. In other words, contempt scores as a milder form of cold anger. Among the 253 episodes generated by subjects using the expert system (Scherer, 1993, pp. 342-343), there were only three episodes (1,2%) of contempt.

Subjects in Frijda et al. (1989) closely associated disgust with anger and contempt in the appraisal profiles. Ellsworth and Smith (1988b) associated both disgust and contempt with anger, Lazarus (1991a) considered disgust as a distinct emotion, but listed both contempt and hate as other terms for anger, and Ortony et al. (1988) used disgust as a variant token for hate. Roseman often mentioned difficulties in incorporating disgust nicely within his appraisal structure, wondering for example whether it should be identified as a variant of distress (Roseman, 1984, p. 30; Roseman et al., 1994, p. 211). In the latest version of his theory, he proposed the dimension of "characterological versus non-characterological" to capture whether it is the basic nature (global character) of someone or something that is of concern or only a specific aspect or action. This new dimension would permit the distinction respectively between frustration, anger and guilt

(non-characterological) on one side, and disgust, contempt and shame (characterological), on the other side.

Disgust does not come out very strongly either in studies bearing on spontaneous categorization of emotion concepts (e.g., only 27 out 200 subjects freely mentioned it as an exemplar of emotion in Fehr and Russell, 1984), and when it does show up, it is usually tightly clustered with anger (Russell and Bullock, 1986, pp. 319-320), but also with hate and contempt (Shaver et al., 1987, p. 1069; Strong and Strong, 1987, p. 811). Moreover, as reviewed in Russell (1991c), there are no words for disgust in languages like Polish or Chewong (Malaysia), and it is not distinguished from hate in Samoan, or from anger in Ifalukian (Polynesia). Fromme and O'Brien (1982) rather consider it as an action pattern associated with drive arousal (hunger or elimination), Kemper (1987) as a mere intensifying transition to a successor emotion, and Panksepp (1982) as of reflexive nature, like the startle reaction. And no recognizable similarities have been reported in monkeys (Chevalier-Skolnikoff, 1973, p. 82; Redican, 1982, p. 265).

Although Rozin and his coworkers (Rozin and Fallon, 1987; Rozin et al., 1993) strongly argued for its basicness, we still feel uneasy about the fact that they cumulated evidence while oscillating between two very different kinds of reactions: disgust as a category of food rejection (i.e. distaste), and disgust as a moral emotion. It is indeed far from obvious that one's physical reaction to the smell of putrid meat could reasonably be put in the same behavioral category as the reaction to the perspective of wearing Hitler' sweater! On the other hand, the authors also acknowledge the intimate relation between disgust, anger, and contempt:

> We think of these three as forming a continuum of moral response. [...] as disgust becomes elaborated, it becomes a more general feeling of revulsion, even to sociomoral violations, and it begins to *shade into anger*. We propose that contempt is the middle ground between anger and disgust. (Rozin et al., 1993, p. 588; our italics).

As a matter of fact, in a recent investigation of the facial expressions of disgust, Rozin, Lowery and Ebert (1994) found that the cluster of expressions denoting "elaborated" (or moral) disgust, was predominantly associated with *upper lip retraction*, "an action that has more of an anger-offense connotation than a "rid-it-of-the-mouth" function" (Rozin et al., 1994, p. 876). It happens that this expression was also, in all three experiments which they conducted, the predominant response for anger, contempt and rights violation. In

addition, this second form of disgust that Rozin and his coworkers see as a "moral emotion", seems to require much cognitive sophistication and elaboration, and for that matter, does not achieve its adult-like form until about 7 or 8 years of age (Rozin and Fallon, 1987, pp. 34-35; Rozin et al., 1993, p. 583). This makes it better fit the kind of complex affective-cognitive structures that Izard (1993) talks about. For all these reasons, we decided to stick with anger, and not to include disgust, contempt or hate as additional building blocks of our postulated emotional layer.

Jealousy and compassion. Jealousy also raises problems as to whether it is fundamental enough to occupy our layer. It was not even mentioned in Table 3.1 but, in Table 3.2, it appeared in three different columns. In category sorting tasks or in semantic analysis of the emotion terms, jealousy is most frequently associated with the anger/hatred cluster (Shaver et al., 1987; Storm and Storm, 1987). But the category to which jealousy is best related happens to show much variability. Subjects in Frijda et al. (1989) associated it with sadness in the appraisal profiles, but with fear in the action readiness ones. Lazarus identifies anger as "the most prominent action tendency in jealousy" (Lazarus, 1991a, p. 257), and one of its hallmarks. Yet, his analysis shows that this emotion basically relies on a threat of loss of an attachment figure (a point also made by Frijda, 1993a, p. 376), so love and fear are also implied. On the other hand, if the loss seems irremediable, then a devastating sadness akin to grief is elicited. And as there is usually a threat to, or a loss of, self-esteem, then humiliation, embarrassment and shame often come out as additional component feelings. All these emotions have been recurrently found to intervene in the whole process (see for instance Salovey and Rodin, 1986). Bryson's (1991) analysis of the answers to a jealousy questionnaire by college students, found, among the constituents of the jealousy responses, the following factors: emotional devastation and need for social support; anger, confrontation and reactive retribution; intropunitiveness. Mathes, Adams and Davies (1985) investigated the emotions that are differentially experienced as the result of two different kinds of loss that White (1981) had postulated in his theory of romantic jealousy: loss of relationship rewards, and loss of self-esteem. They found that the first one results in depression, while the second one causes anxiety and anger.

Some researchers have proposed to conceptualize jealousy as a mixture or a compound of simpler "basic emotions". Johnson-Laird and Oatley, for example, would define it as "*hatred* for someone who is evaluated as supplanting oneself in relation to an *attached*

person" (Johnson-Laird and Oatley, 1989, p. 116). In the first version of their theory, the italicized terms would have to be further expanded, *hatred* to intense *disgust*, and *attached* to love, which in turn was defined as experiencing intense *happiness* in relation to another person. In the recent revision (Oatley and Johnson-Laird, 1996), the whole definition gets simpler, since the authors added hatred and love as new primitives in their basic set. In any case, jealousy requires the interaction of both of disgust/hatred and of love/attachment to occur. Sharpsteen (1991) has recently proposed that romantic jealousy be conceptualized as a blended emotion, with a prototypical representation corresponding to some mixture of the respective prototypes of anger, fear and sadness. One reservation with such a proposal is that it is not clear what computational gain such an extra representation would insure over the already existing three prototypes for handling what happens in the typical jealousy situation. If the components of the jealousy script are all strictly borrowed from the anger, fear and sadness prototypes, then what is added? One might even argue that such redundant representation would reduce the versatility of the person to cope with situations that could take rather different forms across individuals, not to imagine across cultures. It has been shown that two or more emotions often co-occur within the same situation, and that they even sometimes switch from one to the other as the episode unfolds. The best documented (most frequently encountered?) cases happen to be all pair-wise combinations of anger, fear, and sadness (anger+sadness: Oatley and Duncan, 1994, p. 376; Scherer and Tannenbaum, 1986, p. 305; Stein and Trabasso, 1992, p. 229-231; Tomkins, 1984, p. 169; anger+fear: Archer, 1976; Oatley and Duncan, 1994, p. 376; Panksepp, 1989, p. 63; fear+sadness: Elsworth and Smith, 1988a, p. 282; Stearns, 1993, p. 548). One could suppose that such a possibility offers more flexibility than if the organism were constrained by a specific script not containing more resources to handle the situation than what is already at hand with the simpler scripts. It thus seems that the result of all this, is to understand jealousy as much more complex than the other emotions which we have discussed so far, an opinion that is more and more widely shared among theorists:

> The emotion of jealousy consists of action readiness change [...] elicited by a constellation of events, a specific "story" [...] *that story defines the emotion.* (Frijda, 1986, pp. 72-73; our italics)

> Jealousy is viewed as *a label given to a complex* of interrelated emotional, cognitive and behavioral processes. [...] After threat is perceived, the person experiences a variety of emotional reactions (ERs) that may include anger, depression, vengefulness, guilt, and anxiety over perceived loss of control. (White, 1981, pp. 295, 298; our italics)

> The jealousy grieving process involves a succession of sundry emotional and cognitive reactions. [...] Such a complex process, involving changes over months, *cannot be contracted into a single emotion,* whether it is jealousy, anger, depression, or feeling hurt. (Hupka, 1991, p. 256; our italics)

A similar, though much less elaborated criticism, also applies to pity/compassion, which only appears twice in Table 3.2. Ortony and his coworkers would see pity as a rather simple emotion, elicited by being "(displeased about) an event presumed to be undesirable for someone else" (Ortony et al., 1988, p. 93). This reaction thus appears as the opposite of jealousy or resentment, elicited by being "(displeased about) an event presumed to be desirable for someone else" (Ortony et al., 1988, p. 93), and the two emotions thus have a clear negative connotation. For Lazarus though, things are less obvious, and he actually included compassion in his set of problematic emotions. The problem is that, by being goal congruent, the emotion should have positive overtones. But, at the same time, its core relational theme is "being moved by another's suffering and wanting to help" (Lazarus, 1991a, p. 289), which is akin to empathic distress, and sometimes even leads to guilt, two clearly negative enotions. Shaver et al. (1987, p. 1082) saw it as a mixture of sadness and love and, in Storm and Storm (1987, p. 811), it was included into the love/affection cluster. As we did for jealousy, we are thus tempted not to consider compassion as a single emotion and a basic building block of the emotional layer, but as a complex affective-cognitive structure (Izard, 1993) within the operation of which many satellite emotions could also be elicited.

Surprise and interest. Serious doubts have been raised about the possible candidacy of these two reactions. Ortony et al. (1988; Ortony and Turner, 1990) rejected both on the basis that they were not valenced states, which to them meant that they could not even be counted as emotions. They have devised a simple test whereby one has to indicate how confident one is that "feeling x" and "being x" would *both* refer to emotions, when x is replaced by various adjectives (Clore, Ortony and Foss, 1987; Ortony, Clore and Foss, 1987). As an example, "feeling abandoned" would generally be understood as an emotion, but "being abandoned" would rather specify a state of affair. Hence "abandoned" would *not* be considered an emotion. In the resulting analysis, both surprise and interest came out classified as *cognitive* states. Oatley and Johnson-Laird (1987; Johnson-Laird and Oatley, 1989) would exclude surprise and interest on similar grounds,

and also because they are often encountered with various other emotions, perhaps as precursors. Such a view is also shared by Stein et al. (1993):

> Surprise, curiosity, and interest are indicative of the creation of new representations [...] When surprise is felt in close proximity to other emotions, it normally precedes the other emotions. [...] Surprise and interest are different from emotional responses. [...] surprise is usually seen in the initial part of an emotional episode and often indicates the amount of discrepancy that the new information carries. (Stein et al., 1993, pp. 285-287)

When Fehr and Russell (1984) asked for a one-minute free listing of exemplars of emotions, only 8,5% of the subjects listed surprise. In Shaver et al. (1987), of the 135 emotions terms used for the similarity sorting task, all but surprise had received a score higher than 2,75 on a 4-point scale (where 4 meant: "I would definitively call this an emotion"). The authors had nevertheless decided to include it in the list, because it had so often been considered as a basic by emotion theorists. In the cluster analysis, it came out as a meagre category, isolated from all the other branches. In Storm and Storm, surprise was associated with "relatively neutral terms: arousal and cognitive salient" (Storm and Storm, 1987, Fig. 2, p. 810). In Ellsworth and Smith, it was analysed as "a hedonically neutral, short-lived emotion that serves to orient the person to significant events, typically unexpected, externally caused events" (Ellsworth and Smith, 1988a, p. 327). Roseman et al.'s data (1996, p. 261; Roseman, 1984, p. 29) also confirm that surprise is not inherently positive or negative. Finally, recall that the facial expression of surprise was frequently confused with fear in the recognition studies (Ekman and Friesen, 1971; Ekman et al., 1987; Etcoff and Magee, 1992) and had not either been clearly identified in other primates (Chevalier-Skolnikoff, 1973, p. 82; Redican, 1982, p. 265).

Boredom and relief. Boredom is another example of a reaction that does not appear too convincingly as a proper emotion, especially if one adopts the criterion of having to be valenced. Ortony et al. (1987, p. 363) indeed categorized it as a cognitive state, more or less as an opposite to interest and surprise, which were also classified into the same category. Subjects in Storm and Storm (1987, p. 813) likewise classified boredom with relatively neutral terms, mainly denoting variation in arousal or cognitive engagement, and for which the pleasantness-unpleasantness dimension was not salient. Frijda considered boredom as a deactivation state, together with fatigue and apathy, characterized by a kind of "emotional emptiness" (Frijda, 1986, p. 39). In Frijda et al. (1989, pp. 220-222), boredom and indifference fell almost perfectly on a middle position between positive and

negative emotions, on both of the appraisal and the action readiness profiles. Johnson-Laird and Oatley defined boredom as a mild form of depression resulting from the feeling "that one has no goals" (Johnson-Laird and Oatley, 1989, p. 110). Considering that the dimension of goal congruence or goal conduciveness usually correlates highly with the dimension of pleasantness-unpleasantness (Frijda et al., 1989, p. 224; Roseman et al., p. 257), such a definition amounts to considering boredom more or less as valence free. Recall also that Ellsworth and Smith (1988a, p. 282) dropped boredom (and relief as well) from their analysis because not enough subjects had reported it. In their 1985 study, boredom was associated with low attentional activity, and illustrated by experiences in which subjects "were compelled to remain in a situation where there was little or nothing to keep their minds occupied" (Smith and Ellsworth, 1985, p. 833), or had to attend to a dull lecture, or were bed-ridden for a few days after a minor medical intervention. Among the 253 episodes generated by subjects using Scherer's expert system (Scherer, 1993, pp. 342), there were three episodes (1,2%) reported for boredom, only one of which was correctly classified by the expert system. This suggests that, perhaps, the appraisal profile for boredom does not stand as very striking. For instance, if one considers that, in Scherer's model, the "open" value on one dimension of appraisal renders it non-discriminatory, then the difference between boredom and pride rests only upon "concern relevance", whose value is "body" for boredom and "self" for pride. The functional significance for an organism to have developped some machinery to appraise that kind of a difference is certainly far from obvious!

This argument of functional adaptiveness also makes us doubtful as to the relevance of conceptualizing a specific design for relief. This reaction has not been proposed by any basic emotions theorist, but is encountered in three columns in our Table 3.2. Lazarus (1991a, pp. 280-281) suggested that it is elicited by a goal incongruent condition which changes for the better rather unexpectedly. Ortony and his coworkers defined it as being "(pleased about) the disconfirmation of the prospect of an undesirable event" (Ortony et al., 1988, p. 121). Johnson-Laird and Oatley similarly classified relief as a "caused emotion: happiness as a result of something that brings to an end fear or sadness" (Johnson-Laird and Oatley, 1989, p. 118). Roseman et al. (1996) distinguished relief from joy, in that the former resulted from an *aversive* motive-consistent appraisal (punishment withheld), while the latter resulted from an *appetitive* motive-consistent appraisal (reward given). As mentioned already, Ellsworth and Smith (1988a) had initially included relief in their list, but they dropped it from consideration since the subjects in this

study almost never reported it. Opponent process theorists (Solomon, 1980; Solomon and Corbit, 1974), on the other hand, would account for relief by saying that an aversive stimulation not only generates a negative emotion, but also elicits at the same time, in a milder form, an opponent process that has the effect of reducing (regulating?) the intensity of the initial reaction. Sudden removal of the aversive stimulation would then induce the "slave companion" to manifest itself, generating an affective contrast, as if a genuine positive stimulation had occurred. Here again, we think it is more parsimonious to add relief to the joy cluster, than to include as additional states, characteristics that rather belong to the dynamics of emotional episodes (Mauro, 1992), and do not seem to buy additional adaptive value for simpler and more general processes.

3.2.4.3. The elicitation structure

We are thus left with a tentative set of eight basic building blocks out of which to set up our "emotional layer". It should be clear though that it has not been our intention to bluntly ignore genuine distinctions that have been established empirically between certain emotions, such as guilt and shame, or joy and relief, or anger and contempt... It actually happens that, in most analyses of the structure of the affective space, "there are nearly always some emotion or mood words that have unstable placements in clusters or factors: "jealousy," "contempt," and "shame" are among them" (Frijda, 1993b, p. 391). We thought that some rationale had to be proposed for such recurrent observations. With respect to Frijda's (1993a) criticism of the appraisal concept, we questioned whether the observed variations belonged to the elicitation side or whether they resulted from further elaborations on the response side of each emotion. Hence, we have argued that many such distinctions, either pertained to the choice made among different action tendencies from the response set of a given emotion, or to a small combination of emotions intervening simultaneously (or in very close succession). Our stance is actually that of a designer who, as a first approximation, wonders which limited set of processes seems parsimoniously mandatory to account for the functional adaptiveness evidenced by the emotional layer.

From this standpoint, it obviously becomes much easier to propose a coherent structure of appraisal for the elicitation of each category or family of emotions. Here again, we agree with Frijda (1993a) that the basic triggering mechanisms have to be cognitively and computationally much simpler than usually considered by most appraisal theorists.

Accross practically all of the literature surveyed, a few dimensions have emerged recurrently and consistently, such as *valence, activation, agency,* and *certainty.* For instance, in Frijda et al. (1989, p. 217), valence and agency came out of the factor analysis as the two most significant appraisal factors in experiment 1, while valence and certainty emerged as the most significant ones in experiment 2. These three dimensions are also reported as the most important ones in Roseman (1991), and they were also considered as prevalent in Elsworth and Smith (1988a, 1988b).

The two dimensions of valence (pleasantness) and activation have formed the basis of what has been called the "circumplex model" of emotion (Daly, Lancee and Polivy, 1983; Fromme and O'Brien, 1982; Plutchik, 1980, 1984; Russell, 1979, 1980; Russell et al., 1989; Schlosberg, 1941, 1952. 1954). Such a model presumes that all emotion terms will fall radially around the perimeter of a circular space, with emotions highly correlated being close to each other, with the ones showing a zero correlation lying approximatively 90 degrees apart, and with those negatively correlated falling on opposite sides of the diameter. Yet this approach has been criticized (Larsen and Diener, 1992, p. 46; Ortony et al., 1987, p. 343), mainly for the fact that the list of words used for the multidimensional analysis was loaded with such terms as: "active", "strong", "calm", "quiet", "sleepy", "tired", "bored", "challenged", "excited", "aroused", "vigorous", "nonchalant"... Words like these do not fit well with most definitions of emotions (Kleinginna and Kleinginna, 1981), they are not generally considered as representative of emotions by subjects (Shaver et al., 1987; Storm and Storm, 1987), yet they almost guarantee that a dimension of arousal will unwarrantedly emerge from the analysis. Hence, we will discard activation as a proper and useful dimension for the elicitation structure of basic emotions.

Valence. Valence on the other hand appears practically everywhere as the foremost dimension of the motivational. Ortony et al. (1988; see also Ortony et al. 1987) even use it as *the essential criterion* for a mental state to be categorized at all as emotional. Within their theory, the level of arousal or activation would rather be captured as the intensity of emotional reactions, which they saw as resulting from the effect of two sets of variables, global variables, affecting all emotions, and local variables, specific to only certain types of emotions (Ortony et al., 1988, chapter 4; see also the section on "Computational tractability", pp. 181-190). Lazarus (1991a) also distinguished two main classes of emotions, positive and negative, on the basis of their being respectively goal congruent or goal incongruent. Frijda (1987) argued that, if one was to plead for the existence of basic

emotions, only pleasure and pain (displeasure) would appear fundamental enough to figure as candidates. Although Davidson (1984, 1992) would describe this dimension as *approach/withdrawal* rather than as *positive/negative,* this researcher also agrees that such a distinction is characteristic of emotions. In his opinion, it is even grounded in the functional organization of the brain, as reflected in frontal and temporal assymetry, the left-side being more generally responsible for the activation of positively toned emotions, and the right-side being associated with negative ones. All theorists in our Table 3.2 indeed consider this dimension as a fundamental one, and most semantic analyses of emotion words (Ortony et al, 1987; Osgood 1966; Shaver et al., 1987, 1992; Storm and Storm, 1987) also come out with this basic cleavage. In a recent computational analysis of motivation, Wright (1996) presents the concept of valence as a basic process of credit-assignment which introduces competition between inner control states and provides them with the ability to "buy processing power" by gaining access to limited processing resources. This exchange of value among inner states results in some of them being more (or less) capable of controlling behavior. Moreover, this circulation of value can be self-monitored to provide information about gains and losses, a process which accounts for the phenomenology of "achievement pleasure" or "failure unpleasure", i.e. for the mental states associated to the success or failure of important goals to be "felt" respectively as positive or negative.

Agency. Agency also comes out regularly as a fundamental dimension of emotions, although it has been referred to, by different researchers, as responsibility, intentionality, situational control, or locus of causality. Smith and Ellsworth (1985; see also Ellsworth and Smith, 1988a) demonstrated that agency is the main dimension along which subjects distinguish among negative emotions. Wong and Weiner (1981) likewise showed that subjects typically start attributional search in situations of expectancy disconfirmation, or of nonattainment of goals, and that subjects then commonly center their attributional search on the dimension of agency: "why is this happening to me?" is thus usually translated into "who is responsible for it?". This essentially attributional view of the dimension of agency is actually shared by many researchers:

> Why can one be angry when, for no apparent reason, one's car will not start? In our view, this only happens if, in fact, one does attribute causal agency [...] to something [... These attributions] are *a very natural response to such situations.* Indeed, sometimes it seems as though the desire to identify some possible cause, however bizarre and unreasonable the candidate selected in the heat of the moment may be, results from *an unconscious understanding* that one's anger presupposes that there exists such a cause. (Ortony et al., 1988, p. 152; our italics)

> The subjects tend to treat the causes of their discomfort as *intentional and responsible agents*. Cars, bicycles, and rainy days are blamed; subjects may complain that it is unfair that they in particular always have to endure such misfortunes, as if indeed an unjust agent was responsible. [...] The attributions also appear from the actions. *The causal objects are treated as if they were to blame, physically as well as verbally, by scolding or hitting them.* (Frijda, 1993a, p. 363; our italics)

In this last quoted article, Frijda suggested that such an unreasonable interpretation argued against the attribution being causally involved in the elicitation of the emotion, and was rather elaborated *afterwards,* as the episode unfolded. He likewise interpreted the frequently reported fact that people experience more intense feelings of guilt *when they had no intention to cause harm* (McGraw, 1987), as an additional proof that the appraisal of responsibility might be *the result rather than the cause* of the emotion felt. We are inclined to disagree on this point, although we are in full accord that the elicitation process itself has to be fairly simple and automatic. We think that the "loose", and sometimes indiscriminate, attribution of responsibility to others - and even to inanimate objects - in case of anger, and to oneself in case of guilt, are the two sides of the same coin. It expresses some very primitive and innate tendency to look for a responsible agent, especially when something goes wrong. Such a provision might prove fairly adaptive in the context of species complex enough to have to deal mainly with other individuals, on an affiliative or dominance/submissiveness basis, or with other species, on a predator/prey basis. In other words, for individuals of those species, the most important objects in the environment are certainly other individuals. In Gibson's spirit (Gibson, 1977), one might think that these organisms would have evolved designs specifically attuned to abstract as special *affordances* the patterns of significant interactions with these other individuals. We thus contend that the agency factor operates at a much more primitive level than usually considered, so that even fairly simple organisms have a kind of *agency-bias* that induces them to automatically search for some likely responsible agent whenever something significant happens, in terms of being detrimental or unexpectedly useful to current states or goals.

We have already mentioned, while criticizing the idea of basic emotions as unanalysable qualia, some experiments on the electrical stimulation of certain brain sites in various mammals (many detailed examples are reported in Frijda, 1986, pp. 381-386). As discussed, the emotional behaviors thus elicited suggest that the animal is indeed searching for an appropriate agent, to whom to attribute whatever is happening. Agressive

reaction, for instance, will often be directed towards the experimenter or towards animals nearby, as potential agressors. On the other hand, in experiments where a caged animal is subjected to a sudden aversive stimulation, a frequent response will consist of attacks directed towards the most likely responsible agent around: preferably another animal, of the same or of another species, or towards a stuffed animal, and eventually towards an unusual object that happens to be "unduly" present (Hutchinson, 1972). A fairly simple mechanism should then be operative. Some studies on the perception of causality have indeed demonstrated that people readily interpret as intentional (attack, flight, pursuit, attraction...) certain patterns of motion involving mere geometrical shapes, such as squares or rectangles (Bassili, 1976; Michotte, 1950). Primitive as this might appear, it seems that such an "agency-detector" is nevertheless mandatory, even for the emergence of more complex interactive behaviors, such as human verbal communication. For instance, in order to account for the reciprocally intentional behavior assumed as a necessary requirement for speech acts to occur, Searle (1990) also felt that such a primitive sense of agency had to be postulated:

> Collective intentionality presupposes *a Background sense of the other as a candidate for cooperative agency* [...] I am here suggesting that we cannot explain society in terms of either conversation in particular or collective behavior in general, since each of these presuppose a form of society before they can function at all. The *biologically primitive sense of the other as a candidate for shared intentionality* is a necessary condition of all collective behavior and hence of all conversation. (Searle, 1990, 414-415; our italics)

Certainty. As mentioned above, we feel that certainty-uncertainty should also be included as a critical appraisal. Since emotions are typically related to goal attainment, and since goal-oriented behavior necessarily implies a temporal dynamics, future expectations should obviously be involved in the evaluation of incoming events. This dimension actually shows up in many empirical investigations of the appraisals spontaneously evoked by subjects (Ellsworth and Smith, 1988a; Frijda et al., 1989; Roseman, 1991; Roseman et al., 1996; Smith and Ellsworth, 1985). Scherer (1993, p. 349) also suggested that including this factor in the future might improve the performance of his expert system for the identification of episodes experienced with feelings of fear or anxiety.

Figure 3.1
The elicitation structure for basic emotions
(Adapted from Roseman et al., 1996, and Wiener, 1985)

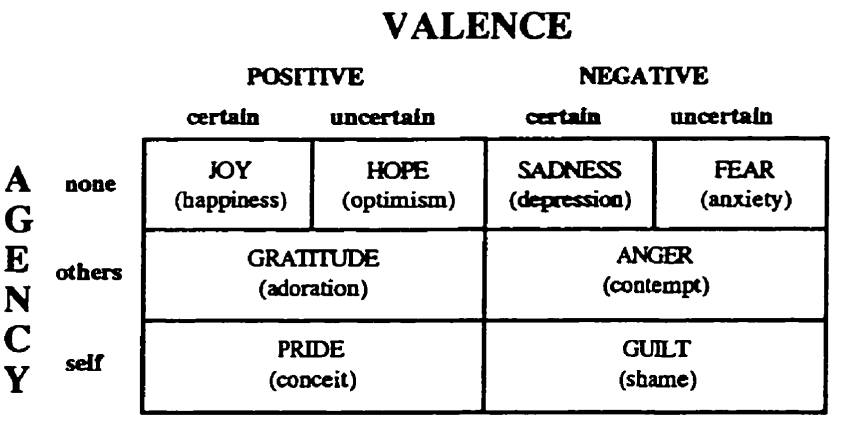


Figure 3.1 summarizes the appraisal structure for the elicitation of "basic" emotions. It corresponds closely to the one proposed by Roseman et al. (1996, Figure 2, p. 269), once the emotions that we have rejected above from our basic set have been eliminated. We found interesting the idea of distinguishing the expression and the further elaboration of the emotional episodes along the *characterological/non-characterological* dimension, although we do not think that such a distinction belongs to the elicitation side of the processes. This distinction is pretty reminiscent of the *globality/focality* component proposed by Frijda (1986, p. 213; 1993b, p. 381), to distinguish moods from emotions proper:

> Globality is the component that differentiates depression from sadness, bliss from joy, and anxiety from fear. In depression, [...] the world as a whole appears barren and devoid of intentional objects. In anxiety, the environment as a whole lacks support for orientation or a focus that might make meaningful any efforts to protect oneself. In bliss or true happiness, the environment as a whole appears accessible and invested with positive valence. (Frijda, 1986, p. 213)

So we included in parentheses the non-characterological (or global) *version* of each emotion. For instance, if self feels responsible for failing to meet social standards or prescriptions, guilt will be experienced when the transgression could be restricted to specific acts or behaviors, while shame would rather be felt if the whole self is seen as inadequate. Anger would result from specific wrongdoings attributed to someone else, but contempt would rather be experienced if this is the whole identity of the wrongdoer that is

judged as deficient. Gratitude would be associated with the receipt of some goods or favors from someone else, but if generalized over the whole person, this feeling could well turn into love and adoration for the benefactor. Pride accompanies specific positive outcomes attributed to self, but conceit (or *hubris*, as Lewis, 1993b, p. 570, would rather call it) would come out when it is the whole self that is praised and seen as responsible for all the good that is resulting. Joy or sadness would be felt in cases of specific pleasant or unpleasant happenings, but happiness or depression would rather be experienced when the pleasantness or the unpleasantness melt over longer periods of time or even over a whole way of life. Hope and fear would characterize anticipated positive or negative outcomes, but optimism or anxiety would result from generalizing these feelings over longer periods and less specific classes of events.

Not too surprisingly, given the prevalence of the agency factor and its attributional character, the two bottom rows of our Figure 3.1 show a strong convergence with the emotional structure postulated by attribution theorists (Wiener, 1982, 1985). It is actually from their position that we chose "gratitude" instead of "love" or "liking" as one of our candidate emotions. For one thing, love/attachment seems to involve a much more complicated configuration including many different emotions, some positive, and some negative, some dependent on self-agency, and some dependent on other-agency: separation distress, jealousy, feeling sorry for someone, feeling happy for someone, worrying about someone, feeling abandoned, missing someone... On the other hand, it is a well known fact that most emotions actually arise within the context of attachment and close relationships (Baumeister and Leary, 1995; Scherer, 1988b; Scherer et al., 1986), and it is also usually in situations involving relatives that the intensity of the feelings is the highest. Thus anger is most typically triggered when friends or kin fail to meet expectations (Averill, 1979, 1983; Lemerise and Dodge, 1993), and guilt is mainly experienced when one has caused harm, whether intentionally or not, to related and valued others (Baumeister, Stillwell and Heatherton, 1994).

As it was also pointed out by Scherer (1988a, p.94), there is no specific event or action that makes one like an object or a person. Hence, it does not seem that love or attachment as such fit at the same level of complexity of emotional reactions as do sadness, guilt, joy, fear or anger, although, as we have discussed earlier, there has to be some simpler emotion appropriately related to this behavioral complex. *Gratitude* thus pops out as a most likely candidate: it is everywhere related to liking, and instigating gratitude, for

instance by providing gifts, rewards or praise, is probably - across ages, cultures and even species - one of the most natural way of "making friends". It is used in the process of taming, in caretaking, in courtship (across a large variety of birds and mammals), and even in political negotiations! We have already seen that hate is most typically associated with anger. Since gratitude, in some appraisal accounts (Ortony et al., 1988, p. 151), is seen as the direct opposite of anger, a logical conclusion would be to consider love and liking (the opposite of hate), as being very closely associated with gratitude. Ellsworth and Smith (1988b, p. 315) included items designed to measure gratitude into their love scale, and found that love was predicted by appraisals of pleasantness and other-agency. In Shaver et al. (1987), two of the items which appeared most frequently in the subjects' accounts as antecedents of love were: "Other provides something that self wants" (hence gratitude) and "Self finds other attractive". Analysis of people's judgments of love showed that "caring" (which again naturally induces gratitude) is seen as playing the most important role in love, while need appears as more important in judgments of attraction (Steck, Levitan, McLane and Kelley, 1982). In other words, a person is generally found attractive on the basis that he or she can fulfill someone else's needs. All this traces a path between need, attractiveness, gratefulness, and liking. On the other hand, the eliciting appraisal for this emotion seems fairly easy to compute. Tesser, Gatewood and Driver (1968), for instance, showed that felt gratitude is a positive monotonic function of three variables: the *intentionality* attributed to the benefactor (thus directly related to our agency dimension), the *cost* incurred in providing the benefit, and the *value* of the reward to the recipient. In a similar vein, Trivers (1971) has suggested "that the emotion of gratitude has been selected to regulate human response to altruistic acts and that the emotion is sensitive to the cost/benefit ratio of such acts" (Trivers, 1971, p. 49).

As we have already mentioned, three of the four emotions that occupy the top row of Figure 3.1 (joy, fear and sadness), have already received a fair confirmation of their universal occurrence across human cultures. It is interesting to note that the four emotions from the bottom rows also evidence some striking universality. Kitayama et al. (1995) asked undergraduates from a Japanese university and undergraduates from a U.S. university to rate the frequency with which they experienced different emotions. The data were submitted to a multidimensional scale analysis, and the result showed that the emotional space in both sets could be segregated into four nonoverlapping areas, along the two dimensions of evaluation (positive/negative) and social orientation (engaged/disengaged). From the emotion terms they included, the four different clusters

obtained can be mapped directly onto the four bottom cells of our Figure 3.1: positive/engaged (e.g. "friendly feelings") onto positive/other-caused; positive/disengaged (e.g. "pride) onto positive/self-caused; negative/engaged (e.g. "guilt") onto negative/self-caused; negative/disengaged (e.g. "anger") onto negative/other-caused.

It might seem strange that the agency dimension could bear a value of "none", as is the case in the first row of Figure 3.1. The point is that joy, sadness, hope or fear are triggered either when no attributable agency can be found as a cause for the observed (expected) outcome, or when there appears to be no possible control over the corresponding event. Emotional processes are seen here as control structures, and from this functional stance, *non-controllable agency is tantamount to no agency at all.* There is, indeed, no point spending resources against overpowerful agents: better call for help then or flee (and alert relatives to flee as well). Thus, in cases of anticipated loss attributable to others, anger would be triggered *only if* there is something to be done against the agents to prevent their wrongdoing, while fear would be triggered if nothing could be envisioned to prevent it. This also explains why the certainty/uncertainty dimension only operates when no agents are involved. When other agents could actually be held responsible, gratitude or anger are triggered, whether the outcome is real or expected. When self is felt as responsible for some significant outcome, pride or guilt are experienced, even if the corresponding event has not happened yet. For instance, a parent can well feel guilty because he might eventually have to punish his or her child.

3.2.4.4. The basically social function of emotions: managing commitments

In summary, we have a substantial rationale for postulating the existence of a specific layer of design responsible for practically all emotional behavior; we also have good reason to speculate that the family (or types) of emotions listed in Figure 3.1 provide most of the required building blocks for this layer; and we have consistent empirical and theoretical material from which to specify the design of the elicitation mechanisms for these processes. Now, adopting such a computational stance to the understanding of emotional phenomena, necessarily merges into the larger framework of a functionalist approach to emotions, which characterizes most current cognitive approaches to emotions (Barrett, 1995; Ekman, 1984, 1992a, 1992b, 1993; Fisher and Tangney, 1995; Fischer et al., 1990; Frijda, 1986, 1988, 1989; Frijda and Swagerman, 1987; Harré, 1986; Izard, 1977, 1993; Lazarus, 1991a; Scherer, 1984; Leventhal and Scherer, 1987; Oatley, 1992;

Oatley and Johnson-Laird, 1987, 1996; Panksepp, 1982, 1986, 1989, 1992; Plutchick, 1980, 1984; Tomkins, 1984; Toobey and Cosmides, 1990; Trevarthen, 1984). The challenge, though, is to formulate which specific functions should be attributed to the emotional layer, so as to account for what is already known to operate as emotional episodes unfold. One very frequently asserted characteristic of emotions, is their fundamentally social nature, even if they do not always or exclusively occur within social contexts:

> In human behavior, situations involving interaction with other human beings are characteristically more heavily laden with emotions than any other situations. (Simon, 1967, p. 37)
>
> [Our] results clearly replicate earlier findings obtained in a series of intercultural questionnaire studies showing that established personal relationships are the most important source of emotional experience for the basic emotions, except for fear. (Scherer and Tannenbaum, 1986, p. 302; see also Wallbott and Scherer, 1986, pp. 72-73)
>
> Many adult plans are mutual: [...] Such plans are among the most important that we make: marriage, parenthood, employment, friendships etc. Many of our more intense and problematic emotions concern plans where mutuality has been sought for, set up, or assumed. (Oatley and Johnson-Laird, 1987, p. 43; see also Oatley, 1992, p. 10, and the whole of chapter 4)
>
> A major component of emotions is social communication and regulation. Because of the close ties between emotional appraisals, subjective feelings, action tendencies, expressions, and behaviours, emotions communicate to other peoples one's needs, wishes, intentions, understandings, and likely behaviours. This communication generally promotes both need satisfaction and social co-ordination. (Fischer et al., 1990, p. 94)
>
> Although emotions can seem to arise privately and without others being around, I think they always involve other persons, and the emotion process draws on past and present relationships with such persons. Emotions are social phenomena. (Lazarus, 1991a, p. 241)
>
> The evidence is sufficiently broad and consistent to suggest that one of the basic functions of emotion is to regulate behavior so as to form and maintain social bonds. (Baumeister and Leary, 1995, p. 508)

The social constructionist view (Averill, 1980; Harré, 1986; Kemper, 1987; Levy, 1984a, 1984b; Lutz, 1988; Rosaldo, 1984) would of course go much further in considering that emotions are "socially constituted syndromes" (Averill, 1980, p. 305) which are not so much biologically (and thus universally) determined, as they are necessarily acquired within the fabric of a given culture, with the appraisal process being itself constrained by the structure of the social surrounding. This actually corresponds to one of the

controversies that we mentioned in the beginning of this article, and it more or less replicates, for the emotion domain, the famous debate between the nativists and the empiricists that has agitated much of the history of experimental psychology, in the field of sensation and perception (Boring, 1942). Again, it is quite unlikely that the truth would lie on any of the extremes, and the computational stance which we have advocated so far, could also offer here a more satisfying account, somewhere in between the two opposing sides.

One basic point of our reasoning, is that *motivation essentially has to do with managing resources.* From an adaptive point of view, resources indeed appear as the ultimate goals that organisms should strive for in order to insure their own survival. We are defining resource here, more or less as energy is defined in physics: whatever is necessary for producing work, such as heat or electricity for machines, processing time and memory for computers, or water and food for living beings. Another key idea is that it proves useful to identify and distinguish two kinds of resources: first-order and second order. First-order resources are the ones that an organism could get directly by itself, while *second-order resources are the ones that are (and most often can only be) obtained through the help of another agent,* which is *committed* to provide them. Newborns, for instance, can only get access to food through the commitment that makes their parents provide it for them. Now, it should be made clear that it is not the other agent *per se* that is to be counted as the second-order resource, but whatever makes it that the agent is compelled to find out and pass over the first-order resource. In other words, *it is the commitment, by way of insuring access to the resource, that is to be counted itself as a (second-order) resource.*

This two-sided classification of resources offers a new and fairly simple criterion for establishing the distinction between the two broad categories of motivations: needs and emotions. Both of them could thus be seen as kinds of operating systems insuring the management of resources in organisms (and eventually in artificially-built autonomous agents). Needs are the control processes that regulate the variations in first-order resources: when there is significant gain or loss, i.e. past some predefined threshold, the hunger (or thirst or fatigue...) mechanism is triggered and acts so as to regain access to an equivalent amount of resource. In higher (i.e. social, or at least nurturing) organisms, other agents could also be called in to help to insure access to the required resources, provided these agents are committed to thus respond. *Evolution has come out with*

emotions as the control processes that regulate the variations in second-order resources, namely the set of commitments established between agents. Emotions thus operate so as to establish or create new commitments (joy, gratitude), protect, sustain or reinvest old ones (joy, hope, gratitude, pride), prevent the breaking of commitments by self or others (pride, guilt, gratitude, anger), or call on others in cases of necessity and helplessness (sadness, fear).

One theoretical consequence of thus restricting emotions to being *commitments operators* (Aubé and Senteni, 1996a, 1996b) is that we would not count as *emotions proper* some simpler reactions, such as pleasure and excitement more directly associated with physiological well-being. We would take a similar stance for the reactions classified by Rozin et al. (1994) as distaste or core disgust: the first one is typically elicited by offensive smells and bad tastes, and is expressed by nose wrinkle; the second one is best elicited in the presence of body waste products, and is rather expressed by gape and tongue extrusion. It is also our opinion that some of the reactions counted as fears, should probably not be categorized as full bloomed emotions, especially certain escaping behaviors evidenced in phylogenetically simpler species. Öhman (1986, 1993) has provided suggestive evidence that, even among human fears, the various reactions should probably be classified into different categories, some of which would rest upon much simpler and evolutionary older structures. Ekman et al. (1985), as mentioned earlier, made a convincing argument that the startle reaction is more reflex-like, and should not then be counted as an emotion. Averill also thought that "most "natural" fears (e.g., of falling, and even of bears in the woods) are not very representative of emotions in general" (Averill, 1980, p. 325). Finally, Wallbott and Scherer (1986, p. 72), and Scherer and Tannenbaum (1986, p. 302) also pointed that fear somewhat differed from all of the other "basic" emotions in being less systematically elicited by situations involving close relationships, and among these emotions, it is also the one that was the most poorly predicted by Scherer's expert system (1993, pp. 347-349).

We prefer to classify as fear reactions that necessarily involve relationships, such as attachment or conflict with other individuals. Fear is thus often induced by others' anger. This is probably one of the original functions of anger, so important actually, that there seems to exist, even in the brain of other primates, some innate detectors insuring the rapid recognition of angry faces (Dimberg, 1982; Hansen and Hansen, 1988; Hasselmo, Rolls and Baylis, 1989; Öhman, 1986; Öhman and Dimberg, 1978; Sackett, 1966). Fear

in self is also likely to induce fear in others, which is precisely the function of *alarm calls* in most birds and mammals (Eibl-Eibesfeldt, 1975; Marler, 1984). In humans, one might think of panic behavior in cases of fire or a riot. On the other hand, there seems to be a strange association between fear and attachment. Darwin had already noticed that "the feeling of affection of a dog towards his master is combined with a strong sense of submission, which is akin to fear" (Darwin, 1872/1965, p. 119). The concept of fear is indeed considered as basic to the analysis of imprinting (Hoffman, 1974; Rosenblum and Alpert, 1974), wherein it is more or less envisioned as the other side of the coin: being attached to a proper caregiver thus goes together with being fearful of other possibly dangerous individuals. The appearance of fear of strangers around seven months in the human infant is likewise interpreted (Lewis, 1993a; Konner, 1986; Sroufe, Waters and Matas, 1974), the apparition of the fear response first requiring the establishment of secure attachment. But probably the most striking illustration of the relation between the two emotions, in humans, is what has been called the "Stockholm syndrome", or the "hostage identification syndrome" (Hillman, 1983; Turner, 1985), wherein victims get strongly attached to their captors and will even come to endorse their agressors' point of view against the authorities. The onset of the phenomenon absolutely requires that the person be totally terrorized as to his or her fate, and victims who do not succomb to intense fear are not subject to the syndrome (Graham, Rawlings and Rimini, 1988; Hillman, 1983). It is now believed that such "traumatic bonding" also happens in the case of battered women, victims of incest and abused children (Dutton and Painter, 1981; Graham et al., 1988). The phenomenon might further be related to the fact that, "when children are anxious or fearful, they are especially vulnerable to passionate love" (Hatfield and Rapson, 1993, p. 597; see also Hatfield, Brinton and Cornelius, 1989).

Even though such emotions as joy or sadness do not require the attribution of agency for their elicitation, they nevertheless typically occur as well in the context of interagent relationships, especially in establishing new commitments, or in making use of the (second-order) resources they make available. Separation distress is thus indissociably related to the attachement process, and it does induce or strengthen longing and bonding. As such, it is not encountered in species where there is no special attachment between parents and offspring, or between mates, such as in amphibians and in most of the reptiles (MacLean, 1993). In species where such processes exist, they induce fear of strangers, as we have seen above. Sadness also elicits sadness: infant crying, for instance, induces crying in other children as well as empathic distress in parents (Murray, 1979; Simner,

1971). As to joy, it is quite communicative in mammal play, even across species, like between dog and master. The *relaxed open-mouth* display, which is seen as a precursor to laughter in primates (Van Hooff, 1972), is associated with play and joyful behavior; in humans, laughter has for long been considered highly contagious (Provine, 1992; Ruch, 1993). The *silent bared-teeth* display, which is seen as a precursor of smiling in primates, is a potent elicitor of affinitive behavior (Van Hooff, 1972). In humans, smiling is a universal component of greetings, "an evolutionary designed signal to smooth interaction among members of a species who must cooperate in group living" (Kraut and Johnston, 1979, p. 1551).

Our idea of seeing emotions as a *control structure* is closely related to Frijda's concept of "control precedence", already mentioned in a previous section (Frijda, 1986, pp. 77-80; 1988; Frijda and Swagerman, 1987), and to Panksepp's view (1992, p. 555) of basic emotions as "operating systems". Moreover, if emotions are control structures, then, by definition, they must be allocated a high processing priority, and that indeed is what makes them such powerful *interrupt* mechanisms (Simon, 1967). The idea is also central to the "communicative theory of emotion" (Johnson-Laird, 1988, pp. 372-376; Oatley, 1992, pp. 49-54; Oatley and Johnson-Laird, 1987, 1996) wherein emotional signals are seen as control messages broadcasted through the whole cognitive system with the function of setting certain modules into readiness to appropriately respond to the eliciting condition. According to their theory, the very same processes also result in sending control signals *externally,* by way of emotional expression, to the whole community of agents, setting them as well into readiness to contribute to resolving a problematic situation. A very similar view has equally been formulated by Minsky:

> Many of the schemes we use for self-control are the same as those we learn to use for influencing other people. We make ourselves behave by exploiting our own fears and desires, offering ourselves rewards, or threatening the loss of what we love. (Minsky, 1985, p. 45)

Borrowing from Carver and Scheier (1990), Oatley and Johnson-Laird (1996, Figure 1; see also Oatley, 1992, p. 50) proposed that the elicitation mechanism for emotions rests upon a monitoring device attached to each goal, evaluating incoming events in terms of the fluctuating probability of achieving (or failing to achieve) the corresponding goal. When the fluctuation is assumed too high, an interrupt signal is sent to the current goal, activating stored plan fragments to be assembled in order to repair the malfunction. At the

same time, a "non-propositional" emotion signal is sent to other parts of the system to set them into readiness to help respond to the problematic change. Stein, Trabasso and Liwag (1993) likewise reasoned that "people have a built-in mechanism that allows them to represent *goal-action-outcome* sequences in relationship to the maintenance of goals" (Stein et al., 1993, p. 281).

Our own stance bears some similarity to these accounts, yet it focuses on resources, typically second-order ones, rather than on goals as such, although there obviously are some intimate relationships between the two. The basic point is that goals do not explain motivation, they rather follow from it. In autonomous agents, *motivation is what generates goals*. And, as we postulated before, motivation has to do with resource (and commitments) management. In the same sense that *knowledge representation* (in terms of beliefs and goals, for instance) appears everywhere in cognitive science as the core inner variable, we contend that *commitment* is the core concept required to explain the basic functioning of emotional processes, much more so than the concepts of goal, or even of appraisal. In the "cognitive science" account of a given phenomenon, what is looked for, is the knowledge representation that has been manipulated, as well as how this manipulation has been performed. In "emotive science" accounts, we suggest that it is the commitment manipulated that should be looked for, as well as the process transformation it has incurred. Such a stance leads us to propose that the monitoring device for the elicitation of emotion should be sensitive to unexpected threats and opportunities *with regards to commitments* rather than specifically with regards to goals. Since goals are most often generated by the motivational force resulting from the strive for resources (including second-order ones), it so happens that these two critical inner variables may be confounded. Our design for emotions should thus incorporate a dynamic representation of all the commitments held by a given agent, together with a context-sensitive monitoring device assessing the variation within those commitments, as a result of ongoing events. On the other hand, the fragments of plans called for when there is a risk of breach in stored commitments or an opportunity for establishing new ones, should be tailored so as to reinstate the previous level of commitment or reinvest the ones that have just been acquired. A more detailed account of the computational specifications of this model has been presented elsewhere (Aubé and Senteni, 1996a, 1996b).

3.2.5. Conclusion

In this article, we have tried to lay down some of the required specifications for the design of a computational model of emotions. We started by stressing the foremost importance of existing controversies in the study of emotions, as ways of challenging existing conjectures, so as to produce new ones and make theories progress. Our main strategy has been to use these controversies as heuristics in probing critical facts and phenomena within the emotion domain. We specifically chose to examine more carefully two of them, namely the debate regarding the primacy of cognition or emotion, and that questioning the usefulness of postulating basic emotions. Our choice was motivated by the fact that both stand at the core of emotion theory, the first dealing with no less than the very definition of emotion, and the second having to do with specifying the elicitation mechanism and the internal structure of the emotional processes. We found it necessary to re-examine some of the critical terms used centrally in each debate, such as "cognitive", "basic" or "appraisal". In each case, we also tried to show how, by adopting a design stance, much of the controversy could be reduced, and how much of the data could thus find a coherent place in the overall puzzle.

We introduced the first section by recalling the important debate between Lazarus and Zajonc about cognition and emotion, and stressed that the crux of the controversy rested on the definitions of both terms being too encompassing. In response to the proposal to discard emotions by subsuming them under the broader class of cognitions, we borrowed from Artificial Life the idea of seeing emotions as one specific and functional layer of design built on top of the *instinctive* layer of fixed action patterns, but underneath the *deliberative* layer of complex acquired memory structures. We thus reformulated the problem of distinguishing between cognition and emotion as one of capturing the corresponding behaviors in terms of the working of successive layers that could be added, through intricate connections, on top of each other, so as to meet the requirements of adaptive and evolutive systems. The next step consisted in formulating some essential properties that could characterize the emotional layer in such a way that it would appear qualitatively different from the underneath instinctive layer, but that it could at the same time provide a useful intermediary step towards the deliberative layer laid on top. We then listed and documented some critical characteristics of emotional structures as being semi-abstract, context-dependent, intentional and communicative.

We started the second section by recalling Ortony and Turner's (1990) challenge to the concept of basic emotions. This led us to a critical review of the principal arguments in favor of basicness. We then saw that the crux of the controversy centered upon the units chosen to build a theory of emotion. Ortony and Turner, together with most appraisal theorists, advocated emotionless components of appraisal, from the composition of which an indefinite number of emotions could potentially arise. Their opponents rather favored a restricted list of "seven plus or minus two" (Russell, 1994) categories of emotions, that they saw as functionally encapsulated solutions to recurrent evolutionary issues. Yet, when we compared carefully the list of emotions proposed on both sides, we found striking similarities and convergence among the set of candidates presented. On the other hand, we recalled Frijda's (1993a) criticism that appraisal theorists had often used interchangeably the concept of appraisal in two different senses, one bearing on the elicitation of the emotional reactions, and the other referring to the content of the experience, which is usually transformed and cognitively elaborated as the emotional episode unfolds.

Applying this remark to our design-oriented stance led us to propose that the two sides of the controversy were distinguished mainly by the focus they set on the whole emotional process: appraisal theorists rather seek to specify the cognitive dimensions along which the antecedents lead to the elicitation of emotions, while basic emotions theorists concentrate their efforts mainly on the expression, the underlying circuitry or the process content of each category of emotions. We suggested that the prototype approach recently applied to the study of emotion (Fehr and Russell, 1984; Parrot and Smith, 1991; Shaver et al., 1987; Shaver et al., 1992; Tangney and Fischer, 1995) offered a promising compromise compatible with the best contributions from both sides. This approach did actually specify, for each "basic level" emotion considered, a complete script spelling out the elicitation conditions, the response tendencies, and the self-regulatory procedures represented and used by the subjects in typical situations. In the third section, we then scrutinized all of the suggestions made from each side of the controversy so as to specify which building blocks could be postulated as the required components of our emotional layer. We came out with a strong rationale favoring a reduced set of eight (families of) emotions - joy, sadness, hope, fear, gratitude, anger, pride and guilt - together with an elicitation structure based on three fundamental dimensions of appraisal: valence, agency and certainty. We also showed that the model offered a reasonable convergence, integrating a host of theoretical proposals from many different schools of thought.

This theoretical journey finally led us to specify the functions which had to be ascribed to the emotional processes, so they would fulfill a coherent role in between the instinctive and the deliberative layers of our postulated design. Considering the vast number of empirical references in the literature to the social character of emotions, either about their antecedents or about their interpersonal expression, we chose to restrict their role to this social aspect, thus rejecting many non-social reactions as pertaining to a different set of simpler motivational processes (such as needs, for instance). *Interestingly, such a proposal offers at the same time some workable compromise to another current controversy mentioned in the introduction, opposing the biopsychological and the social constructionist views of emotions.* Our point is that motivation has to do with managing resources, and that the access to some resources depends on others agents. The key to this access is the commitment that binds these agents into providing the desired resource. Hence, the commitment is a special kind of resource, which actually motivates "sociality" and glues individuals together into reciprocal relationships. This concept is indeed central to symbolic interactionist sociologists, such as Elihu Gerson, who see the negotiation of commitments as the substance of social order (Gerson, 1976, p. 799; see also Bond, 1990, and Gasser, 1991, for an application of this perspective to the field of Distributed Artificial Intelligence). But, as is the case for simple needs, where hunger, for instance, insures access to food in case of starving, there has to be a specific control mechanism insuring access to and protection of commitments. Through the evolution of social and cooperating species, this is precisely what emotions seem to have been designed for. Such a viewpoint calls for the provision of automatic built-in detectors and interrupt circuitry for critical fluctuations in commitments, but at the same time it leaves room for the social determination of threshold adjustments and for the cultural choice and carving out of action tendencies for repair, as well as for the acquisition of self-regulatory procedures.

A design stance such as the one we have proposed calls for new tools for the computational implementation of the emotional layer. We indeed postulated that a proper representation of emotions should be process-oriented, in executable script-like format, but also that it should necessarily run in parallel, since many accounts of co-occuring emotions appear in the literature. In computational terms, this would call for using such formalism as concurrent objects or actors (Agha, 1986; Briot, 1994). We also suggested that the core variable referring to the commitments held by an agent should be represented in a dynamic fashion, incorporating for each commitment a monitoring device, constantly

watching for events that could bear an impact upon the level of this second-order resource. For biological organisms, this seems to require that the sensory and perceptual processes would directly project onto the emotional system. LeDoux indeed mentioned that:

> [...] visual and auditory areas in the thalamus, in addition to projecting to neocortical receiving areas, also send fibers to the hypothalamus and amygdala [...] Moreover, direct retinal projection to the thalamus (Pickard and Silverman, 1981) and possibly the hippocampus (MacLean and Creswell, 1970) have been observed. These pathways allow crude sensory inputs to reach limbic areas rapidly and without neocortical intervention. (LeDoux, 1986, pp. 331-332)

The best provision we have in mind for the concrete implementation of such a mechanism is something akin to the exception handling system in object-oriented languages such as Smalltalk-80 (Dony, 1990; Aubé and Senteni, 1996a, 1996b). We have also illustrated how this elicitation process could rely on rather simple characteristics of the stimulation, by computing the positive or negative variations in stored (second-order) resources, by identifying a likely cause to which to ascribe the responsibility for this variation and, if no agent could be held accountable, by estimating the certainty of the appraised perturbating event. As mentioned, it is likely than individuals in social or nurturing species would have developed mechanisms attuned to rapidly catch the critical "affordances" (Gibson, 1977) of commitments being seriously at stake. We pointed to some references suggesting that such mechanisms are indeed innately wired-in (Bassili, 1976; Hansen and Hansen, 1988; Hasselmo, Rolls and Baylis, 1989; Michotte, 1950; Sackett, 1966).

On the other hand, we think that the recent progress in understanding the computational requirements of parallelism and of multi-agent architectures make the time ripe for further applications of this paradigm to theories of emotions. This why we have elsewhere proposed (Aubé and Senteni, 1996a, 1996b) that emotions should be represented as the very control processes of multiagent systems. Reciprocally, we envision an important contribution that studies of emotion could make to Distributed Artificial Intelligence (DAI) research. Based on a sociologically-minded metaphor (Gasser, 1991; Minsky, 1985), this new perspective rests on the powerful idea that real world problems generally do not get solved by single individuals, but by teams of specialists of different skills that cooperate together in their solution (Kornfield and Hewitt, 1981; Lenat, 1975), drawing mutual profit from their distributed knowledge. DAI incorporates the idea by distributing a given task across a variety of processors, which then *concurrently* search their way towards the

goal, communicating and exchanging useful information among them. Not surprisingly, this new paradigm is also confronted with its own set of problems and limitations (Bond and Gasser, 1988, p. 10). Interestingly, they all have to do with communication and control: how to best decompose and allocate a given task among agents, how to insure that communication protocols are founded on a commonly shared semantics, how to avoid harmful effects of local decisions, how to enable agents to represent plans and knowledge of other agents, how to reconcile disparate viewpoints and resolve conflicts... This looks pretty much like the problems emotions serve to solve in social organisms that have developed them, a view that is also consonant with the concept of "augmented planning" that Oatley (1992, p. 31-36) sees as the principal motivation behind the advent of emotions in phylogeny. Our contention is that studies of emotions might provide DAI with powerful insights into the *control structure* it has laboriously and painstakingly been looking for.

Our proposal also points towards useful empirical research to be pursued. Clearly, one has to question whether our postulated set of eight basic emotions should be enlarged or reduced, and we contend that any convincing arguments on this matter will have to make provision for a likely design, incorporating an evolutionary perspective. In a similar vein, if the idea of functionally intermingled layers is adequate, it is nonetheless no easy task to provide evidence for such an intricate arrangement, unless one looks at simpler organisms where the upper "deliberative" layer is still in the process of being built up. Hence the need for more comparative studies. As a research strategy, one should especially look at the border between layers. At the frontier between the instinctive and the emotional ones, the phenomena of imitation, motor mimicry and emotional contagion (Dautenhahn, 1995; Hatfield et al., 1992, 1994; Meltzoff and Gopnik, 1993; Meltzoff and Moore, 1994; Mitchell, 1987; Povinelli and Eddy, 1996) constitute a potentially promising field for further investigation. At the frontier between the emotional and the deliberative layers, the question of multiple memory systems (Sherry and Schacter, 1987; Squire, 1992; Tulving, 1985), and notably the relation between emotions and episodic memory (Conway and Bekerian, 1987; Conway, 1990), certainly calls for further elucidation. All of these should certainly make the already swarming field of emotion studies further explode in many new fascinating directions!

Chapitre 4.

La spécification du modèle proprement dit

L'objet du présent chapitre est de préciser l'articulation de notre modèle et de formuler un certain nombre de spécifications permettant éventuellement une implantation de structures de contrôle "à caractère émotionnel" au sein de systèmes informatiques multi-agents. La première section vise en particulier à relier le concept d'engagement, qui tient une place centrale dans notre conceptualisation des structures émotionnelles, à celui de coopération, sur lequel repose l'aptitude à la "socialité" (Castelfranchi, 1990) et en conséquence, l'existence même des systèmes multi-agents. Nous présenterons ensuite deux articles, déjà publiés dans la communauté d'IAD (Aubé et Senteni, 1996a, 1996b), qui occuperont respectivement les deux dernières sections du chapitre.

4.1. De la coopération aux engagements: vers une structure de contrôle pour la gestion de la réciprocité

Nous avons mentionné à quelques reprises dans les chapitres précédents le rôle explicatif que nous attribuons à la notion d'engagement dans notre effort pour "désenvelopper" et mettre à jour les rouages cachés de la "mécanique émotionnelle". Dans cette section, nous allons tenter de montrer que ce concept constitue l'un des prérequis ("requirements") du processus de coopération, sur lequel repose à son tour le fonctionnement, *sinon l'existence même,* des espèces sociales. Nous allons commencer par examiner un certain nombre de recherches portant sur l'évolution de ce processus, qui soulignent justement la nécessité d'un mécanisme qui puisse offrir des garanties de *réciprocité* entre les individus éventuellement amenés à coopérer. Nous tâcherons alors de montrer que cette réciprocité requiert l'opération de structures motivationnelles de contrôle inscrites ("designed-in") chez les organismes sociaux de façon à favoriser leur adaptation et leur survie. Dans les

deux articles qui feront l'objet des sections suivantes, nous expliciterons la définition que nous donnons des engagements, comme *ressources de deuxième ordre,* fondant précisément un type particulier de motivation, et commandant du même coup l'existence de dispositifs régulateurs destinés à en assurer la gestion et le contrôle.

4.1.1. Le problème de l'altruisme et de la coopération

La coopération, qui constitue pourtant l'une des bases essentielles de la vie sociale, reste un phénomène complexe et obscur. Quelles sont en effet les conditions qui président à son émergence dans un groupe, et qui assurent qu'elle s'y maintiendra? Dans des domaines comme le néo-darwinisme et la sociobiologie, ce phénomène est probablement même, avec l'altruisme, celui qui s'avère le plus problématique (Dawkins, 1976; Krebs, 1987; Wilson, 1975):

> If altruism is defined as behavior that enhances the net fitness of another individual at a net reduction of the fitness of the helper [...] then it should not evolve according to Darwin's principle of natural selection. In the extreme case, individuals who inherit characteristics that dispose them to sacrifice their lives for the sake of others would be less likely to survive and reproduce other individuals with such altruistic characteristics than those who were more selfish. (Krebs, 1987, p. 83)

Dans la perspective de la sélection naturelle, de la survie et de la reproduction des plus aptes, il est en effet difficile de concevoir que des organismes puissent, par pure générosité, se priver délibérément, au profit d'autres organismes, de *ressources* leur garantissant de meilleures chances de survie. Il est surtout difficile d'imaginer sur quelle base de tels comportements de sacrifice et d'abnégation, vraisemblablement "déficitaires" à long terme, peuvent avoir été sélectionnés en tant que solution adaptative au cours de l'évolution. La question de la coopération pourrait, au premier abord, sembler moins problématique. Il est cependant important de comprendre que, dans la perspective d'individus cherchant chacun à maximiser leurs propres chances de survie et de reproduction, la tendance d'un organisme à coopérer facilement avec ses congénères le rend du même coup extrêmement vulnérable à l'exploitation. D'autres individus pourraient alors simplement tirer profit de l'aide apportée, sans offrir en retour de compensation réciproque. En ce sens, l'altruisme et la coopération constituent plus qu'un défi: ils vont apparemment à l'encontre de la "logique de la sélection naturelle" et remettent ainsi en cause les fondements mêmes de la théorie de l'évolution.

Il y a quelques années, un chercheur en sciences politiques, Robert Axelrod (1984; Axelrod et Hamilton, 1981), fit une expérience pour le moins originale en organisant un *tournoi informatique* sur les stratégies de coopération entre deux individus. Intéressé à comprendre comment la coopération pouvait être possible, et surtout si elle pouvait émerger "naturellement" de l'interaction, il utilisa comme contexte expérimental le "Dilemme du Prisonnier" (DP). Il s'agit là d'un paradigme emprunté à la théorie des jeux et dont l'appellation est tirée de l'histoire suivante. Deux criminels ont commis un délit et sont appréhendés. Cependant, bien que le procureur soit capable de les faire condamner pour des larcins antérieurs, il ne possède pas suffisamment de preuves permettant de les inculper pour ce crime-ci. Il compte donc sur leurs aveux, rencontre séparément les deux criminels et leur propose un marché. Si l'un d'eux avoue, il obtiendra une remise de peine, alors que l'autre écopera de la peine maximale prévue. Si les deux avouent, le procureur n'a évidemment pas besoin d'offrir de faveur à aucun, ni non plus de maximiser la peine. Par contre, si les deux prisonniers restent solidaires et se taisent, le procureur ne peux que les inculper pour les délits mineurs antérieurs. Pour eux, ce pourrait être la meilleure solution, mais alors comment chacun peut-il être sûr que l'autre ne profitera pas de ce silence pour se retrouver seul à avouer et s'en tirer encore à meilleur compte? D'où le "dilemme", sur lequel compte évidemment le procureur.

La théorie des jeux étudie mathématiquement les conditions de décision chez deux ou plusieurs joueurs, disposant d'un répertoire de stratégies possibles, chacune assortie d'enjeux en termes de gains ou de pertes aisément quantifiables et déterminés à l'avance. La figure 4.1 illustre l'une des formes possibles que peuvent prendre ces jeux:

Figure 4.1
Le paradigme du "Dilemme du Prisonnier"

		Joueur B C	Joueur B D
Joueur A	C	(R, R)	(N, T)
	D	(T, N)	(P, P)

La question est toujours de savoir s'il existe, pour l'un ou l'autre des joueurs, une stratégie gagnante, ou du moins une combinaison probabiliste de diverses stratégies qui puisse, à long terme, s'avérer "optimale". Ce terme doit évidemment être interprété ici selon la perspective de la "rationalité économique" (Simon, 1980) où les individus sont censés chercher, chacun de leur côté, à maximiser leurs gains et à minimiser leurs pertes. Dans l'exemple illustré à la figure 4.1, chaque joueur dispose de deux stratégies à son répertoire: la coopération (C) avec le partenaire de jeu, ou la défection (D), qui exprime la tendance à tirer profit de la bonne volonté de l'adversaire à coopérer. Les lettres entre parenthèses représentent, de gauche à droite, les gains respectifs des joueurs A et B en fonction de chaque combinaison de choix: R signifie la "récompense" en cas de coopération des deux côtés; P, la "punition" en cas de défection réciproque; T, la "tentation" à profiter de la coopération de l'adversaire; et N, la "naïveté" à se laisser abuser par l'autre. Pour qu'il y ait "dilemme du prisonnier", deux conditions particulières doivent être remplies:

$$\text{Règle 1:} \qquad T > R > P > N$$

$$\text{Règle 2:} \qquad R > \frac{(T + N)}{2}$$

La première règle énonce que le meilleur gain (d'où la "tentation") résulte de la défection alors que le partenaire est disposé à coopérer. La seconde meilleure option est la coopération réciproque, la troisième, la défection réciproque, et la pire est la naïveté en face d'un partenaire non coopératif et profiteur. Comme la défection en présence de naïveté donne le meilleur avantage, et qu'il y a peu de chance qu'un même joueur reste toujours naïf, on peut imaginer que les joueurs décident de s'entendre pour alterner ces rôles. La deuxième règle stipule cependant qu'en moyenne, à long terme, il y a plus d'intérêt dans la coopération réciproque que dans cette alternance calculée entre les rôles de naïf et d'exploiteur. En outre, il découle mathématiquement de ces deux règles que la coopération réciproque est également plus profitable qu'un jeu laissé complètement au hasard, et donc que:

$$R > \frac{(T + R + P + N)}{4}$$

Comme valeurs possibles respectant ces règles, les joueurs pourraient obtenir respectivement un gain de trois (R=3) lorsqu'ils coopèrent tous les deux, mais de un seulement s'ils font tous les deux défection (P=1). Par contre, si l'un des deux seulement faisait défection pendant que l'autre coopère, il obtiendrait le gain maximum de cinq (T=5), alors que son partenaire ne recevrait rien du tout (N=0).

Il est important de comprendre que les décisions sont prises simultanément, de telle sorte que l'un des joueurs ne peut attendre de voir si l'autre est disposé à coopérer avant d'opter de son côté pour la défection. Dans une telle perspective, il est facile de voir que l'attitude "rationnelle" de chaque joueur est de choisir la défection. Le joueur A peut en effet se demander, en fonction de chaque stratégie de l'adversaire, quel serait son meilleur choix. Si B choisit de coopérer, A peut obtenir 5 au lieu de 3 en optant pour la défection. Par contre, si B choisit de ne pas coopérer, A obtient 1 au lieu de 0 en refusant lui aussi de coopérer. Comme les gains sont symétriques, cette analyse suggère que, dans un cas comme dans l'autre, aussi bien pour A que pour B, il est toujours plus profitable de ne pas coopérer, ce qui rapporte 1 point à chaque partenaire. Et pourtant, les joueurs pourraient se mériter chacun un gain de 3 en se faisant mutuellement confiance et en choisissant de coopérer. Mais si A tire de ce raisonnement la conclusion que B coopérera nécessairement, il redevient possible d'extorquer un meilleur gain (5 vs 3) en ne coopérant plus, etc. La complexité de choix qu'entraîne un telle ambiguïté rend ce paradigme si attrayant, qu'il a été utilisé pour modéliser une foule de processus complexes, aussi bien en sociologie qu'en économie ou en sciences politiques, comme les dilemmes sociaux (Dawes, 1980; Hardin, 1968) ou la course aux armements.

Le paradigme est encore plus intéressant lorsque le jeu est répété pour une longue période, car l'intérêt de développer des mécanismes permettant la confiance mutuelle et rendant la coopération moins périlleuse, devient alors manifeste. Or on ne connaît toujours pas, dans le contexte de la théorie des jeux, de meilleure stratégie possible, alors que cela a pu être démontré mathématiquement pour d'autres paradigmes. Axelrod (1984) eut l'idée ingénieuse de lancer un *tournoi informatique*, en demandant à divers experts en théorie des jeux, recrutés principalement parmi les chercheurs ayant publié sur le sujet, de proposer le meilleur programme possible susceptible de "gagner" à long terme une joute du DP, caractérisée par les gains donnés plus haut en exemple (T=5, R=3, P=1, N=0) et comportant un séquence de 200 coups. Il y eut quatorze soumissions, provenant de chercheurs appartenant à cinq disciplines différentes: psychologie, sociologie, mathématiques, économie et sciences politiques. Axelrod rajouta également à la liste un programme incorporant une stratégie purement aléatoire, et chacun des quinze programmes fut mis en compétition contre chaque autre, y compris contre lui-même. Bien que le score pour une partie puisse théoriquement varier de 0 à 1000 points, dans le cas improbable où l'un des joueurs se laisse exploiter à chaque coup, il oscillera surtout entre 600 pour deux

partenaires qui coopèrent à tous les coups, et 200 pour deux autres qui feraient au contraire systématiquement défection.

Or la stratégie gagnante, à travers tout le tournoi, fut aussi l'une des plus simples imaginables: "TIT FOR TAT" (TFT), que l'on pourrait traduire par "Donnant, Donnant". Il s'agit d'une stratégie qui consiste à se montrer coopératif au premier coup, et ensuite à reproduire exactement le jeu l'adversaire au coup précédent: coopérer s'il l'a fait, sinon faire à son tour défection. Les stratégies plus complexes - coopérant par exemple pour certains coups, puis faisant défection dans un certain pourcentage des cas, ou encore cherchant à exploiter statistiquement la propension de l'adversaire à tolérer la défection - se sont toutes montrées moins performantes. La stratégie TFT remporta pour sa part un moyenne de 504 points sur l'ensemble des parties. Or une caractéristique importante permettait de distinguer les stratégies occupant la moitié supérieure au tableau de performance: elle avaient toutes fait preuve de "bienveillance", au sens où elles n'avaient jamais été les premières à faire défection. Sur les quinze programmes en compétition, les huit meilleurs avaient *tous* cette caractéristique de "bienveillance", alors que parmi les sept derniers, *aucun* n'en avait fait preuve. Au-delà de toute interprétation morale, il faut surtout comprendre qu'aucune stratégie n'est jamais complètement significative en elle-même, et que sa performance *globale* est surtout déterminée par l'ensemble des autres stratégies auxquelles elle se trouve confrontée.

Axelrod publia les résultats détaillés du tournoi, ainsi que son analyse de la caractéristique de "bienveillance", et lança un second tournoi, qui recueillit cette fois 62 participants, provenant de six pays différents, et rassemblant, en plus des cinq disciplines déjà représentées au premier tournoi, des experts en informatique, en physique, et en théorie de l'évolution. Cette fois encore, TFT sortit grand vainqueur. En plus de la "bienveillance", l'analyse des deux tournois révéla trois autres caractéristiques contribuant à expliquer la robustesse de cette stratégie: son "intolérance" implacable à la défection, sa "miséricorde" lorsqu'il y a retour à la coopération, et sa simplicité d'application (les termes entre guillemets doivent évidemment être entendus dans un sens purement métaphorique, puisque la stratégie est appliquée de façon complètement mécanique). En effet, parce qu'il reproduit fidèlement le comportement précédent du partenaire, TFT se trouve à "sembler punir" toute défection, mais aussi à "sembler pardonner", en redevenant aussitôt conciliant, dès que l'autre se remet à coopérer. Par ailleurs, la stratégie ne requiert ni une grande mémoire, ni une analyse ou un calcul statistique complexes des compor-

tements antérieurs de l'adversaire. Elle est également suffisamment transparente pour qu'un partenaire sache facilement la reconnaître et réalise rapidement les profits qu'il y a à coopérer en même temps que les risques qu'il y a à faire défection. Les stratégies adverses qui se laissent trop facilement tenter par l'exploitation y perdent vite leur propre compte, et celles qui se montrent trop "rancunières", se retrouvent enlisées dans une boucle de "punitions" réciproques fatalement déficitaires à long terme.

Intéressé aux conditions d'émergence des attitudes coopératives, Axelrod s'interrogea ensuite sur les chances de survie et de reproduction de TFT dans une simulation à caractère "évolutionniste". Comme le succès d'une stratégie donnée est relatif à la nature des autres stratégies qui peuplent le même environnement, l'auteur imagina que chaque ronde du jeu corresponde à une "génération", et que chacune des 63 stratégies du second tournoi puisse être représentée par de multiples copies d'elles-mêmes, en nombre proportionnel à son propre score dans la partie. Une stratégie qui fait meilleure figure face à l'ensemble des autres, verra ainsi le nombre de ses "rejetons" augmenter, et verra notamment s'accroître sa chance de jouer contre des copies d'elle-même. Dans tous les cas, l'écologie de tels "organismes" risque de varier considérablement d'une génération à l'autre, puisque la nature des partenaires variera également, entraînant possiblement une modification substantielle des conditions ultérieures de succès - et de reproduction. Ainsi, une stratégie non coopérative reposant sur l'exploitation pourra paraître performante au début, mais elle contribuera aussi rapidement à la disparition des "proies" qui lui ont conféré son avantage, et se retrouvera par la suite de plus en plus souvent confrontée à d'autres partenaires aussi "voraces" qu'elle-même. Une stratégie qui s'avère performante contre une grande variété d'autres programmes, subsistera plus longtemps, mais se verra alors confrontée aux meilleures des stratégies restantes - une situation qui reflète assez fidèlement le principe de la sélection naturelle. Encore là, les stratégies bienveillantes se maintinrent dans le peloton de tête, et TFT sortit grand gagnant. Dès la quinzième génération, les stratégies occupant le tiers inférieur du tableau de performance avaient pratiquement toutes été éliminées. Vers la millième génération, TFT était de loin la meilleure, et continuait de croître à un rythme plus rapide que toutes les autres.

Axelrod se demanda finalement si TFT s'avérait une stratégie stable au plan évolutif ("evolutionary stable strategy" - ESS). On entend par là une stratégie qui ne peut être envahie par aucune autre, issue par exemple d'une mutation fortuite. L'auteur arriva même à démontrer mathématiquement que, dans le contexte du DP, lorsque la probabilité d'une

rencontre ultérieure entre les participants est suffisamment élevée, une stratégie "égoïste", faisant toujours défection, ne pouvait pas envahir une population de TFT. Ce résultat entraîne à son tour que, s'il existe déjà certains sous-groupes coopératifs au sein d'une population majoritairement égoïste, la stratégie coopérative aura tendance à augmenter et même à devenir majoritaire, plutôt qu'à être totalement éliminée. Axelrod qualifia de "structurée" une population contenant ainsi certains noyaux animés de stratégies particulières, telles ces "poches de résistance" à l'égoïsme pur. Or l'existence de comportements parentaux dans la nature, notamment chez les mammifères et chez les oiseaux, garantit automatiquement qu'il y aura au moins quelques niches coopératives, à partir desquelles une stratégie comme TFT pourrait proliférer. Axelrod crut pouvoir en tirer la conclusion que TFT était une ESS, et ne pouvait donc plus être délogée une fois installée.

Des travaux ultérieurs ont cependant démontré que, si TFT ne pouvait être envahie par une stratégie mutante, certaines *combinaisons* de stratégies (plutôt que des stratégies isolées), dont l'existence est effectivement plausible dans un environnement naturel peuplé d'une diversité d'espèces, pouvaient par contre envahir et déloger une stratégie aussi robuste que TFT (Axelrod et Dion, 1988; Boyd et Lorberbaum, 1987; Lorberbaum, 1993; Nowak et Sigmund; 1994; Nowak, May et Sigmund, 1995). Ces démonstrations reposent toutefois sur le fait qu'au moins l'une des stratégies envahissantes soit de nature bienveillante, et fort semblable elle-même à TFT. Par ailleurs diverses simulations informatiques qui ont mis en jeu une grande variété de stratégies opérant, dans une perspective évolutionniste, selon le principe de la sélection naturelle, ont généralement fait ressortir le succès et la robustesse de stratégies coopératives, basées sur la réciprocité, et constituant pour la plupart des variantes fort avoisinantes de TFT (Axelrod, 1984; Glance et Huberman, 1994; Lomborg, 1992; Nowak et Sigmund; 1994; Nowak, May et Sigmund, 1995).

Ces recherches révèlent donc que, même dans la perspective "égoïste" de la théorie de l'évolution et de la sélection naturelle, la coopération s'avère non seulement possible, mais qu'elle est même avantageuse et adaptative, *à la condition toutefois que soit mis en place un mécanisme de "contrôle" assurant une base de réciprocité entre les partenaires.* Or il est intéressant de constater que, dans l'espèce humaine à tout le moins, il semble exister une telle "norme de réciprocité" (Gouldner, 1960; Mauss, 1950):

> We owe others certain things because of what they have previously done for us, because of the history of previous interaction we have had with them. It is this kind of

> obligation which is entailed by the generalized norm of reciprocity. [...] A norm of reciprocity is, I suspect, no less universal and important an element of culture than the incest taboo. (Gouldner, 1960, p. 171)

Cette norme est considérée par les sociologues comme un "fait social" (Durkheim, 1947; Mauss, 1950), du fait de son extrême généralité à travers les sociétés et les cultures, et de par l'action coercitive qu'elle semble exercer sur les consciences, en contraignant des pratiques sociales d'échange particulières et déterminées. Elle constitue par ailleurs un concept central en théorie sociologique, puisqu'elle rend compte des phénomènes de stabilité et d'instabilité dans les systèmes sociaux:

> [...] unequal exchanges of goods and services are socially disruptive [...] Social system stability, then, presumably depends in part on the mutually contingent exchange of gratifications, that is on reciprocity as exchange. [... Reciprocity] is the pattern of exchange through which the mutual dependance of people, brought about by the division of labor, is realized. (Gouldner, 1960, pp. 167-170)

En bref, la stabilité, sinon l'existence même d'une structure sociale, repose sur la coopération, laquelle n'apparaît à son tour possible qu'en fonction de l'existence d'une norme de réciprocité, qui est effectivement constatée généralement à travers la recherche sociologique et anthropologique. Cet argument de stabilité sociale est cependant problématique du point de vue de la théorie contemporaine de l'évolution (Krebs et Davies, 1987), car il semble faire reposer l'émergence du comportement de réciprocité sur la notion de "sélection de groupe", c'est-à-dire sur ce qui est le plus favorable pour le groupe, plutôt que pour l'individu lui-même. Or, il y a de bonnes raisons de croire que ce soit l'individu plutôt que le groupe qui constitue le lieu privilégié où opère la sélection, notamment parce que le rythme de remplacement des individus est forcément plus rapide que celui des groupes, et parce que la mobilité spatiale des individus (entraînant la dispersion de leurs gènes) est également plus grande que celle des groupes. C'est donc à l'intérieur même des individus qu'il convient de rechercher les rouages, biologiques et psychologiques, qui sont responsables de la mise en place, du maintien et de la régulation d'un dispositif tel celui qui se manifeste à travers la norme de réciprocité. Comme le rappelle en effet le biologiste Trivers, à propos de l'évolution de l'altruisme réciproque:

> The human altruistic system is a sensitive, unstable one. Often it will pay to cheat [...] Given this unstable character of the system, where a degree of cheating is adaptive, natural selection will rapidly favor a complex psychological system in each individual regulating both his own altruistic and cheating tendencies and his responses to these tendencies in others. (Trivers, 1971, p. 48)

Selon ce théoricien de l'évolution, *les structures émotionnelles sont apparues et se sont développées pour remplir précisément cette fonction.* Ainsi, la tendance à créer des liens d'attachement, d'amitié ou d'intimité viserait justement à favoriser les comportements d'échange et d'altruisme réciproque, tout en réduisant les risques d'exploitation. Par ailleurs, la colère et l'indignation qui, dans une grande diversité de cultures, sont associées à des situations d'injustice ou de transgression, serviraient un triple but: freiner la tendance à coopérer lorsqu'il n'y a plus réciprocité; sensibiliser le partenaire - et même les observateurs - aux risques qu'il y a à tromper la confiance établie; et dans les cas extrêmes, effrayer, punir ou ostraciser les exploiteurs, sinon carrément les éliminer. La gratitude aurait pour fonction de sensibiliser les individus aux comportements altruistes en y attachant une gratification importante et significative, alors que la culpabilité jouerait un rôle d'autopunition, en même temps que de signal réparateur auprès du partenaire afin de permettre le rétablissement des attitudes coopératives.

Bien qu'à partir d'un point de vue différent, Herbert Simon (1990b) proposa lui aussi un mécanisme permettant de rendre compte de l'apparition de comportements altruistes, et nécessitant également *le recours aux structures émotionnelles.* Son argumentation repose essentiellement sur le principe de la rationalité limitée ("bounded rationality"), qui constitue l'un des postulats de base de la science cognitive contemporaine (Simon, 1980, 1990a). Ce concept exprime le fait qu'en dehors de situations triviales, un individu peut rarement disposer de toutes les connaissances et de tout le temps requis pour assurer qu'il adoptera, en chaque circonstance, la meilleure décision possible. Cette limitation intrinsèque plaide en faveur de ce que Simon appelle la "docilité", c'est-à-dire la disposition à recevoir des autres une partie de l'information requise, et en conséquence à accepter leur influence:

> The human species is notable, although not unique among animals, in requiring for survival many years of nurture by adults. In most human societies, the survival and fitness even of adults depends on the assistance, or at least forbearance, of other adults. Leaving aside active hostility from others, even access to food and shelter cannot be ensured in most societies without the consent of others. (Simon, 1990b, p.1666)

L'adaptabilité de l'espèce humaine repose en effet sur l'exploitation d'une multitude de faits et de connaissances que nous n'avons pas tous eu l'opportunité d'observer ou de vérifier par nous-mêmes, mais qu'il est pourtant avantageux d'accepter tels quels. Ainsi,

la plupart des enfants ont appris de leurs parents à ne pas toucher aux ronds brûlants des poêles et à ne pas courir à travers les rues achalandées, sans avoir cependant dû en éprouver par eux-mêmes les conséquences douloureuses. Même comme adultes, nous acceptons assez facilement de la part d'experts des conseils qui nous semblent judicieux ou bénéfiques, par exemple lorsqu'ils proviennent de notre médecin, de notre dentiste ou de notre conseiller fiscal.

Simon suggère donc de mettre à profit ce type de "docilité" afin d'inculquer à la majorité des individus des pratiques altruistes adaptatives, et le modèle qu'il propose à cet égard est fort simple. Supposons que les individus à tendance altruiste et ceux à tendance égoïste se répartissent respectivement dans la population avec une probabilité de p et de $(1-p)$. Supposons en outre que le nombre de rejetons obtenus normalement soit de X, que les comportements altruistes résultent en une contribution moyenne de b rejetons additionnels pour chaque membre de la société (aussi bien égoïste qu'altruiste), mais que le surplus d'efforts requis impose aux seuls altruistes un coût se traduisant par une diminution de c rejetons. Le nombre de rejetons obtenus par chacun des sous-groupes sera donc de:

$$N_A = X + bp - c, \qquad \text{pour les altruistes}$$
$$N_E = X + bp, \qquad \text{pour les égoïstes}$$

Supposons par ailleurs que la "docilité", telle qu'entendue plus haut, donne accès, pour les individus qui y souscrivent, à des ressources qui résultent en d rejetons additionnels. Puisque les comportements de coopération sont, comme on l'a vu plus haut, généralement bénéfiques à long terme, mais que la tentation de faire défection est toujours présente, supposons maintenant que l'altruisme doive être acquis par le biais de cette docilité, de sorte que la population des altruistes coïncide désormais avec celle des dociles. La répartition des deux sous-groupes est alors la suivante:

$$N_D = X + bp + d - c, \qquad \text{pour les dociles-altruistes}$$
$$N_E = X + bp, \qquad \text{pour les égoïstes}$$

Pour que l'altruisme puisse croître et s'installer dans la population, il suffit alors que le gain résultant de la docilité excède le coût entraîné par l'altruisme ($d - c > 0$). Toutefois, comme pour toute transmission de connaissance nécessitant la docilité, il n'est justement pas évident pour le bénéficiaire que les comportements prescrits soient effectivement les plus avantageux, et il faudra donc que l'apprentissage soit consolidé par une intervention

sociale appropriée des parents ou des pairs. C'est ici qu'interviennent, selon Simon, divers mécanismes de contrôle à caractère émotionnels, comme la fierté et la gratitude, la honte ou la culpabilité:

> Moreover, guilt and shame will tend to enforce even behavior that is perceived as altruistic. [...] nurturing and enforcing behavior will be learned as an essential component of the proper behavior of altruism. In enforcement are included carrots as well as sticks - praising and nurturing others who exhibit proper behavior, as well as frowning on, shunning, or otherwise punishing those who do not. (Simon, 1990b, p.1667)

Bien que moins élaborée, cette vision converge quand même passablement avec celle de Trivers, notamment au sujet de la fonction des émotions comme mécanismes régulateurs des comportements de réciprocité. Il est par ailleurs intéressant de retrouver chez Minsky une idée tout-à-fait analogue à celle de Simon, quant à l'importance et à la fonction des émotions dans la transmission de valeurs et d'idéaux - comme celui de l'altruisme - d'une génération à l'autre, par le biais de ce qu'il appelle "l'apprentissage-par-attachement", et auquel nous avons déjà fait allusion dans les chapitres 2 et 3:

> [...] our genes must build some sort of "general-purpose" machinery through which individuals can acquire and transmit goals and values from one generation to another. [... I] suggest that this is done by exploiting the kinds of personal relationships we call emotional, such as fear and affection, attachment and dependency, or hate and love. (Minsky, 1985, p. 164)

4.2. Présentation des deux articles

Il est donc possible que la coopération puisse émerger "naturellement" dans l'évolution des espèces, puisqu'elle offre des bénéfices substantiels pour la survie, en termes de ressources additionnelles souvent inaccessibles autrement. Mais elle requiert pour s'établir une attitude de réciprocité entre les partenaires, dont il est assez facile de démontrer que le maintien restera par principe fragile, en face des risques d'abus toujours plus avantageux lorsque l'exploitation s'effectue de façon unilatérale. Dans cette perspective, il existe un espace de contraintes ("requirements") favorable à l'apparition de dispositifs régulateurs de la réciprocité, et certains théoriciens, en théorie de l'évolution aussi bien qu'en intelligence artificielle, ont déjà pointé vers les processus émotionnels comme candidats potentiels pour remplir, au moins en partie, une telle fonction. L'objectif des deux articles

qui suivent est donc d'expliciter le fonctionnement de tels mécanismes, afin d'en permettre éventuellement une implantation utile au sein de systèmes informatiques multi-agents.

L'un des concepts fondamentaux autour duquel s'articule notre théorisation est celui d'engagement ("commitment"), qui occupe une place privilégiée en IAD (Bond, 1990; Bouron, 1992; Cohen et Levesque, 1990a; Dongha, 1994; Fikes, 1982; Gasser, 1991; Jennings, 1993; Shoham, 1993; Winograd et Flores, 1986). L'interprétation la plus répandue reste cependant plus individuelle que sociale, et consiste à définir cette notion essentiellement en terme de persistance dans la réalisation des buts que s'est fixés l'individu concerné. Cette notion est alors intimement liée à celle d'intention, et vise à rendre compte des contraintes qui semblent imposer, aux systèmes doués de rationalité, de s'en tenir aux intentions choisies ("Intention is choice with commitment" - Cohen et Levesque, 1990a). En effet, comme l'exprime également Bratman (1990), les concepts de plans, de buts, d'intentions et de persistance sont tous indissociablement reliés:

> Deliberation is a process that takes time and uses other resources [...] By settling on future-directed intentions, we allow present deliberation to shape later conduct [...] Future-directed intentions help facilitate both intra- and interpersonal coordination. How? Part of the answer comes from noting that future-directed intentions are typically elements of larger plans. [...] Plans, as I shall understand them, are mental states involving an appropriate sort of commitment to action. (Bratman, 1990, pp. 18-19)

Or, comme on a pu le constater au chapitre 2, cette conception des engagements en terme de persistance est aussi celle qui prévaut actuellement dans le contexte des recherches sur la motivation scolaire (Viau, 1994). Elle garde cependant un caractère plus descriptif qu'explicatif en désignant, sans toutefois véritablement l'expliquer, une contrainte notoire des comportements de planification et d'atteinte des buts. En fait, il serait plus exact de dire que le seul élément explicatif concerne la gestion des ressources et renvoie au principe de la "rationalité limitée". Il existe cependant une autre interprétation du concept d'engagement, à caractère plus "sociologique" cette fois, et cherchant à rendre compte de la mécanique sous-jacente à l'expression de la contrainte observée. L'histoire suivante, rapportée par Becker (1960 - l'année de publication explique pourquoi les montants semblent si peu élevés par rapport aux taux d'aujourd'hui!), sert à en expliquer le fonctionnement et certaines des implications:

> Suppose that you are bargaining to buy a house; you offer sixteen thousand dollars, but the seller insists on twenty thousand. Now suppose that you offer your antagonist

> in the bargaining certified proof that you have bet a third party five thousand dollars that you will not pay more than sixteen thousand dollars for the house. Your opponent must admit defeat because you would lose money by raising your bid; you have committed yourself to pay no more than you originally offered. (Becker, 1960, p. 35)

Cet exemple illustre trois caractéristiques fondamentales qu'il importe de prendre en compte lorsque l'on cherche à élucider en quoi consistent les engagements, et comment ils opèrent sur la gestion des intentions. Il y a tout d'abord un aspect essentiel de *motivation*. La personne a en effet construit son propre engagement sur la mobilisation explicite d'autres intérêts qui se cumulent à ceux qui sont déjà en jeu. Il y a ensuite une question très nette, et clairement quantifiable, de *ressource* à gagner, ou du moins à éviter de perdre. En effet, l'engagement est avant tout un "engagement de ressources", qui se retrouveront dès lors automatiquement en péril si l'engagement n'est pas tenu. Et finalement, il y a *un autre agent*, témoin et garant de l'engagement, qui en valide d'une certaine façon la réalité et en garantit la stabilité. Becker généralise alors cette troisième caractéristique en suggérant que, du simple fait de son appartenance à une organisation sociale, l'individu se retrouve automatiquement impliqué dans une multitude de paris secondaires ("side bets") du même genre qui l'engagent comme malgré lui, et contraignent son activité. Ces trois caractéristiques, bien que généralement peu développées par la suite dans la littérature, nous apparaissent fondamentales pour la compréhension du concept d'engagement, et du même coup, elles nous permettront ultérieurement une réunification des conceptions individuelle et sociale du terme, notamment par le biais de ce que nous avons choisi d'appeler *ressources de deuxième ordre*.

La perspective sociologique introduite par Becker a été approfondie par Gerson (1976), Bond (1990) et Gasser (1991). Reprenant à leur compte l'intuition de George Herbert Mead (1934) à l'effet que l'individu ne préexiste pas à la société, mais que l'un et l'autre se génèrent mutuellement dans un processus de constante interaction, ces auteurs réinterprètent les engagements comme un phénomène émergent: "Commitment from the social perspective is grounded in the action of many agents' activities taken together - it is not a matter of individual choice" (Gasser, 1991, p. 114). Ces interactions multiples induisent des contraintes sur les ressources ("mutually agreed constraints on action, belief and world state", Bond, 1990, p. 21), générant par là même l'ordre et la stabilité sociale:

> [...] patterns of commitment flow [... are] the "substance" of order. [...] Commitments and their organization must be expressed in terms of resources and constraints upon their use. [...] "What is negotiated," therefore, are patterns of

> commitment organization, as expressed in flows of ressources and constraints upon them. (Gerson, 1976, p.799)

Nous ne sommes pas très loin, comme on peut le constater, de la "norme de réciprocité" mentionnée plus haut en tant que préalable à la coopération et responsable, selon Gouldner (1960), de la stabilité sociale. Il nous semble cependant que le concept de réciprocité repose lui-même sur celui, plus général, d'engagement. C'est précisément la relation entre ces différents concepts d'engagements, de motivation et de ressources que nous avons tenté d'expliciter et d'articuler de façon plus cohérente dans les deux articles qui occupent les deux dernières sections de ce chapitre, et qui ont été présentés dans deux communautés internationales différentes, l'une en IAD, et l'autre en robotique et vie artificielle.

Nous y affirmons tout d'abord que l'autonomie d'un agent repose essentiellement sur le contrôle de sa motivation, c'est-à-dire sur sa capacité à déterminer lui-même ses propres buts. Par ailleurs, nous identifions comme but ultime celui de la survie de l'agent, laquelle repose à son tour sur l'accès aux ressources. La motivation apparaît ainsi nécessairement déterminée par la gestion des ressources. Nous distinguons ensuite deux catégories de ressources, les ressources de premier ordre, comme l'eau ou la nourriture, qui sont directement accessibles à un agent, et celles de deuxième ordre, qui ne sont accessibles que par l'intermédiaire d'un autre agent, engagé à les fournir. Il est important de comprendre que ce n'est pas l'agent médiateur qui constitue comme tel la ressource, mais plus spécifiquement ce qui l'amène ou le contraint à rendre disponible au premier agent une ressource quelconque. Autrement dit, c'est *l'engagement* proprement dit qui constitue effectivement la ressource! Cette définition particulière nous permet d'établir une distinction très nette entre besoins et émotions, deux catégories motivationnelles souvent partiellement confondues. Nous restreignons ainsi l'appellation de besoins aux mécanismes de contrôle qui gèrent les ressources de premier ordre, et réservons celle d'émotions aux mécanismes qui gèrent les ressources de second ordre. C'est aussi ce qui nous amènera à conférer aux émotions un caractère strictement "social". La taxonomie des émotions que nous présentons par la suite, sur la base de leurs mécanismes respectifs de déclenchement, est celle qui a déjà été justifiée abondamment dans le chapitre précédent.

Dans le premier article (Aubé et Senteni, 1996a), nous établissons en outre un lien original entre les actes de langage et l'expression des émotions. Là encore, c'est la définition particulière que nous proposons du concept d'engagement qui nous autorise ce

rapprochement. Nous expliquons en effet que notre conceptualisation des émotions comme mécanismes régulateurs des engagements implique l'activation de séquences d'action visant à protéger les ressources de deuxième ordre mises en péril, à compenser pour les déficits encourus, ou à générer de nouveaux engagements. Certaines de ces actions sont en fait des actes de communication, réifiés à travers l'expression même des émotions. Or c'est précisément l'originalité de la théorie des actes de langage (Austin, 1960; Searle, 1969, 1979; Vanderveken, 1990) que de concevoir les énoncés verbaux comme des véhicules efficaces des intentions, qui se traduisent alors par des actions concrètes portées sur le monde extérieur, ou parfois plus subtilement sur les croyances et les intentions mêmes des autres interlocuteurs, pour générer à leur tour de nouvelles séquences d'actions.

La théorie des actes de langage s'inscrit en fait à l'intérieur d'une théorie plus large sur l'intentionnalité (Searle, 1983) à l'intérieur de laquelle les engagements occupent une place de premier choix. Winograd et Flores (1986, voir aussi Winograd, 1988) font effectivement ressortir que les énoncés verbaux expriment ou commandent foncièrement de la part des interlocuteurs des engagements qui rendent ainsi possible la coordination sociale. Or nous avons justement tenté de démontrer, dans le chapitre précédent, que l'intentionnalité constituait précisément l'une des caractéristiques essentielles de la "couche de contrôle émotionnelle". Comme le mentionne d'ailleurs Frijda:

> Much of emotional behavior is intentional behavior [...] Many emotions can be defined by an *intentional structure:* that of maintaining or changing a given kind of situation, and doing that in certain ways, with respect to certain aspects of the given situation and with regard to given kinds of object in those situations. *They can often be more closely, more appropriately, and more specifically defined in this manner than in any other.* (Frijda, 1986, p. 98; les italiques sont de nous)

Ainsi, la colère exprime souvent l'intention d'éliminer ce qui fait obstacle à l'atteinte du but visé, et l'expression de tristesse lance généralement un appel à l'aide et à la protection. Pourtant, la théorie des actes de langages, qui porte essentiellement sur les énoncés verbaux, arrive mal à intégrer cette notion, en la confinant à la catégorie des "actes expressifs", qui se retrouvent généralement laissés pour compte dans les applications formelles de la théorie. Nous proposons alors d'introduire l'idée d'actes émotionnels, comme mode d'expression plus primitive et non verbale, néanmoins orientée vers la gestion des engagements. Cette modification permet de contourner les ambiguïtés des taxinomies existantes, et suggère du même coup des pistes nouvelles à explorer pour la

compréhension de la nature et de l'évolution de la communication chez les animaux: le chien qui couine ou jappe pour obtenir sa pitance, n'obtient de résultat que par l'engagement à le nourrir qui est ainsi sollicité chez son maître, et le grondement qui empêche l'intrus de transgresser la frontière du territoire opère également par le même dynamisme.

Par ailleurs, au moment de soumettre cet article, nous avions encore en tête l'idée d'utiliser comme banc d'essai le problème des *agendas distribués*. C'est ce qui explique que l'on retrouve à la section 4.3.5 une analyse sommaire de ce problème, ainsi qu'un aperçu des spécifications que nous étions en train d'élaborer à ce moment, en vue d'une implantation dans le langage à base d'acteurs Actalk (Briot, 1994), construit au-dessus de Smalltalk-80. Nous avons déjà exposé au chapitre premier, dans la section sur les considérations méthodologiques, les raisons qui nous ont finalement conduit à écarter cette possibilité.

Le second article (Aubé et Senteni, 1996b) commence par approfondir la question de l'autonomie en énumérant certaines des contraintes qui requièrent cette propriété dans la conceptualisation des agents, comme condition de leur adaptabilité à la complexité des systèmes ouverts (Hewitt, 1985, 1991). L'autonomie est ainsi rattachée à la motivation, en invoquant la capacité pour un système donné de déterminer ses propres buts, ce qui soulève une fois de plus la question de la gestion des ressources. La présentation du concept d'engagements comme ressources de deuxième ordre, ainsi que des émotions comme opérateurs des engagements, recoupe alors d'assez près l'analyse qui en a été faite dans l'article précédent. La figure 4.4 illustrant l'interaction entre les deux systèmes de contrôle que constituent respectivement les besoins et les émotions, offre cependant une représentation plus fidèle que celle de la figure 4.3, en les présentant comme deux couches superposées, apparues successivement au cours de l'évolution, la seconde s'appuyant sur la première dans la lignée de l'architecture de "subsumption" de Brooks (1991b). L'originalité du texte repose cependant surtout sur la section 4.4.4, "The emotion engine", qui élabore de façon plus détaillée les spécifications requises pour une éventuelle implantation.

L'utilisation du système de gestion des exceptions (Dony, 1990) comme structure de contrôle privilégiée pour la prise en charge des interruptions de calcul ("computation") en présence de certains signaux critiques avait été brièvement annoncée dans l'article précédent et est plus détaillée dans le second. Nous avons en passant pu vérifier au plan informatique que cette fonctionnalité de Smalltalk était également opérationnelle dans le

contexte de calcul à base d'acteurs, où la propagation des signaux d'interruption doit se faire en mode asynchrone, par le biais de la boîte aux lettres de chaque acteur. Par ailleurs, la spécification des signaux de déclenchement a été largement détaillée au chapitre trois, à la section 3.2.4.3, "The elicitation structure". Les actions à exécuter par les "emotional handlers" en cas de déclenchement approprié, sont celles qui, dans le chapitre trois, section 3.2.3.5, correspondent au répertoire d'actions spécifié dans le script rattaché à chaque catégorie émotionnelle. Parmi ces actions, il faut notamment inclure les "actes émotionnels" d'expression dont nous avons parlé plus haut dans la présente section.

Pour que l'implantation de notre structure de gestion des engagements soit opérationnelle et effective, il nous semble absolument incontournable que chaque membre d'un système multi-agent partage la même "ontologie". Cela signifie notamment que le câblage doit être tel que les signaux émotionnels s'avèrent mutuellement contraignants. Cette spécification relève des propriétés mêmes d'un système motivationnel. Il est important de comprendre en outre que, pour un système le moindrement complexe, un câblage purement réactif pourrait difficilement rencontrer les exigences d'adaptation. Si un animal, par exemple, devait attendre d'être en déficit grave de substances nutritives, avant même que ne soit activé le mécanisme régulateur de la faim, il y aurait de fortes chances pour qu'alors il ne dispose plus de suffisamment d'énergie pour enclencher la recherche de nourriture, qui n'est pas toujours immédiatement accessible dans un environnement ouvert et complexe. Il faut donc supposer l'existence de mécanismes internes de plaisir ou d'inconfort, capables de médiatiser ou carrément de "générer" la motivation requise, en fonction d'indicateurs d'une variation significative du "stock" de ressources internes (Lapierre et Braun, 1992; Lindsay et Norman, 1972). Réciproquement, on sait que le mécanisme de satiété ne repose pas non plus sur le signal que l'énergie requise a été régénérée, puisque le processus de digestion est beaucoup plus lent que celui d'ingestion, et que l'animal risquerait alors de s'endommager lui-même à trop ingurgiter. *La même logique doit forcément s'appliquer ici à la gestion des ressources de deuxième ordre:* au-delà d'une relation directement et immédiatement fonctionnelle entre un événement potentiellement déclencheur et la réponse émotionnelle appropriée, il faut donc concevoir l'existence de processus intermédiaires aptes à détecter des signaux constituant des indicateurs relativement fidèles des ressources qui sont compromises (par un obstacle ou un danger), ou exceptionnellement favorisées (par le biais d'opportunités inespérées).

C'est ce qui explique par exemple que l'attachement parental, qui offrira vraisemblablement des gains réciproques, mais seulement à long terme, puisse quand même se produire sur la base de signaux activant de façon quasi irrépressible cette motivation intermédiaire. Minsky postule à cet égard que le cerveau de l'enfant est équipé génétiquement pour l'émission de signaux sociaux assurant la coordination de ses comportements avec ceux des adultes qui l'entourent, "by making the agencies inside those other people's brains available for exploitation by the agencies in the infant's brain" (Minsky, 1985, p. 297). Cette exploitation des ressources de l'adulte, assurée essentiellement par le biais des messages émotionnels, s'effectue d'ailleurs avec une extrême efficacité:

> [...] adults are made to find those signals irresistible: there must be special systems in our brains that give such messages a high priority. [...] My guess is that they're wired to the remnants of the same [agents] that, when aroused, caused us as infants to cry in the first place. This leads adults to respond to babies' cries by attributing to them the same degrees of urgency that we ourselves would have to feel to make us shriek with similar intensity. (Minsky, 1985, p. 171)

Ces considérations suggèrent que le modèle informatique proposé ici est également plausible au plan neurophysiologique. Une autre originalité du second article consiste dans l'idée de "structures d'engagements à long terme", tels le pouvoir et l'affiliation. Il est important en effet de comprendre que le pouvoir d'un agent n'existe que s'il est reconnu par d'autres agents. Perdu sur une île étrangère, aucun monarque ne conserve grand pouvoir! Autrement dit, le pouvoir repose essentiellement sur les engagements, implicites ou explicites, d'autres agents à reconnaître et favoriser au premier un accès privilégié à certaines ressources (taxes et impôts, respect des lois, obéissance civile, service militaire...). L'attachement offre un autre exemple de structure d'engagements, qui s'accumulent ou diminuent au gré de l'histoire des interactions, sous l'effet des opérateurs des engagements que constituent les émotions de joie, de gratitude, de tristesse ou de colère. Une fois les engagements définis de façon dynamique sous la forme "acteurs" tel que nous le proposons, il est possible d'envisager le "stockage" de ressources de deuxième ordre, tout comme cela est naturellement possible chez les organismes vivants pour les ressources de premier ordre, par exemple sous la forme de graisse ou de réserve d'eau dans les tissus. Cela nous permet d'intégrer facilement à notre modèle les phénomènes reliés aux hiérarchies de pouvoir, généralement intimement liés au déclenchement et à l'expression des émotions, ainsi qu'à toute la richesse de nos attachements à long terme. Le texte se termine en illustrant comment ces concepts d'engagements à long terme rendent alors possible la compréhension de l'anecdote rapportée par Lorenz au début de l'article.

4.3. Emotions as commitments operators: A foundation for control structure in multi-agents systems[3]

Abstract. Typically, multi-agent systems (MAS) rely on sociological theory and use centrally the concept of commitment. A definition, as well as a tentative computational model of this concept, emerge from research in sociology and in computer science, leading naturally to the study of resource management. We claim that effective resource management in a distributed system requires a powerful control structure yet to be founded. We then propose a model of control that stems from the psychology of human emotions as part of the fundamental definition of more autonomous and adaptive multi-agent systems.

4.3.1. Introduction

In the beginning of their article, *Intention is Choice with Commitment*, Cohen and Levesque (1990a) introduce Willy, the household robot, who gets into trouble for not handling commitments properly. In the example, Willy is first asked to bring a beer, and intends to do so, but then decides to adopt some other goal. It is thus shipped back to the manufacturer for lack of commitment. Having been returned to its master, the robot is asked again for a beer. This time though, the master drops his request when he learns that the make of the beverage is not his preferred kind. Yet Willy sticks to the order, brings a beer, and ends up being sent to the shop, for the reason of being overcommitted. After further tuning, the robot finally gets home, having been supposedly wired-in with the proper sense of commitment. Again, Willy is asked for a beer, actually the last one that is left in the fridge. Understanding correctly that its commitment will be over if it is impossible to fulfill, Willy deliberately smashes the bottle against the wall. In spite of its subtle reasoning capacities, the robot ends up being... dismantled!

The question we want to address here is how such a system as Willy could become capable of adjusting *from within* to the proper level of commitment, instead of having to be corrected and tuned up totally *from the outside*. In other words, what kind of mechanism should be designed inside of an agent so that it would handle variations of commitments by itself. This question will lead us to reconsider the relations between some

---

[3] Publié sous: Aubé, M. and Senteni, A. (1996a). Emotions as commitments operators: A foundation for control structure in multi-agents systems. In W. Van de Velde and J. W. Perram, (Eds.), *Agents breaking away*, Proceedings of the 7th European Workshop on MAAMAW, Lecture Notes on Artificial Intelligence, No. 1038, (pp. 13-25). Berlin: Springer.

of the basic concepts in distributed artificial intelligence (DAI) such as: autonomy, motivation and resource management.

Communities of cooperating autonomous agents, as well as human societies, use resources whose management is at the core of any organized activity. Resources are defined here more or less as energy is in physics: it basically amounts to the cost of work or actions. Within animals or animats, needs, such as hunger and thirst, appear to be the motivational processes that insure management and regulation of directly consumable resources, be it available food or water for animals or processing time for a computer. These are *first-order resources*. Besides, there exist *second-order resources*, which an agent gets only indirectly, through having other agents *committed* to provide them: for instance, newborns could have access to food only through having their parents get the feeding resources for them.

Commitments, essentially measured in terms of resource allocation and constraints upon resources, hence appear as one critical reason for the setting up of multi-agent societies. By regulating exchange, they enable using each other as a way to access additional resources otherwise unaccessible, or too costly to obtain. From their earliest definitions, multi-agent systems (MAS) relie on sociological theory (Gasser, 1991) and use centrally the concept of commitments (Fikes, 1982). Building upon the seminal work of symbolic interactionists such as Gerson (1976), Bond (1990) and Dongha (1994) propose a definition and a computational model of the concept of commitment that leads naturally to the study of resource management. We think that effective resource management in a distributed system requires powerful control structures yet to be founded.

The main purpose of this paper is to sketch the design for such a model of control based upon resource management. The second section defines first-order and second-order resources, and revisits the concept of commitments in this context. The third section proposes a model of control that stems from the psychology of human emotions (Oatley, 1992). It introduces emotions as powerful control structures that act so as to protect agents - and societies of agents - from running out of resources, typically second-order ones. The fourth section takes a closer look at the utilisation of speech acts as communication primitives among agents and introduces the new concept of *emotional acts*, as a way to resolve ambiguities inherent to the expressive and declarative illocutionary categories. We contend that emotion-like control structures should become

part of the fundamental definition of more autonomous, and more adaptive multi-agent systems. The fifth section considers distributed meeting scheduling as a possible testbed for that kind of control structures and raises some implementation issues.

4.3.2. From resources to commitments

4.3.2.1. First-order and second-order resources

If agents are ever to become autonomous, it will only be through getting control over their motivations, by having ways of setting their own goals. The idea of providing agents with the capability to generate their own top level goals is not alien to the DAI community: it can be found in Maes' earlier work (Maes, 1991b) while the concept of *motivated agency* is introduced by Norman and Long (1995). Motivations essentially have to do with managing resources: this is what is most crucial and vital for the agents, and what becomes their ultimate goal. In the agent context, resources are defined more or less as energy is in physics: whatever enables - and is required for - producing work. It could be physical energy like heat or electricity or food, but also time, money, affiliation, power, knowledge... For instance, if a robot learns about a shorter route than usual to some target, this specific knowledge is clearly worth some economy in fuel. Basically, the concept of resource is tantamount to the cost of work or actions. However it is essential to refine it and distinguish at least two kinds of resources: first-order and second-order.

First-order resources. First-order resources are directly consumable ones that an agent already has at disposal (available food or water for animals, processing time or memory for computer). In animals, *needs* - such as hunger, thirst, fatigue, shelter, sex... - are the motivational processes (Toates, 1986) that insure management and regulation of first-order resources.

Second-order resources. Second-order resources are those that an agent gets only indirectly, through having other agents committed to giving them to him: newborns could get access to food only through having their parents provide the resources for them. Money as well should be seen as a second-order resource: any piece of money always amounts to a commitment to exchange some goods for it. And as such, it is indeed also revocable, be it only through the variation of currency!

4.3.2.2. Commitments as second-order resources

Commitments are essentially measured in terms of the allocation of resources and constraints upon resources (Gerson, 1976; Bond, 1990; Dongha, 1994). They thus appear as one critical reason for the setting up of multi-agent societies: by regulating exchange, they enable using each other as a way to access additional resources that would remain otherwise unaccessible, or would be too costly to obtain .

Typically, an agent A *commits* certain of its resources to another agent B (commits stands here for *does not give right away, but promises to give eventually*). As a consequence, B is *allocated* new resources, while A is *constrained* as to the usage of its own resources. For instance, A might be constrained not to consume all of them and to take into account some amount that has to remain available to fulfill the commitment. On the other hand, such commitment could also involve A setting the goal of acquiring for B some resources that A does not already possess.

One important corollary is that, by virtue of making resources available, *commitments themselves are to be counted as resources*. Within complex species that pursue multiple goals, and which thus absolutely require the cooperation of other agents, commitments might well become the most important of all resources! Hence, it is very likely that, in social species like ours, evolutionary forces would have favored and selected for precisely those organisms that would have become equipped with powerful mechanisms to insure control over commitments!

4.3.3. Emotions as a metaphor for control structures

4.3.3.1. The appraisal structure of emotions: valence, certainty, agency

Valence: positive and negative emotions. In higher (social, or at least nurturing) animals, *emotions* - such as anger, fear, joy, sadness, guilt... - are the motivational processes that insure management and regulation of second-order resources, namely commitments. Very much like needs, emotions are powerful control structures that act so as to protect agents - and societies of agents - from running out of resources. Hence they require being allocated a higher level of processing priority than most other mechanisms. These control processes are triggered by significant variations in the flow of resources available or

required for an agent (or a community of agents). There indeed has to be regular variations in the daily consumption of resources, without these overpowerful processes intervening within the ongoing computation. Significant variations here means that the amount of increase or decrease of resources should exceed a certain predefined (although eventually ajustable) *threshold*.

There obviously have to be two main classes of emotions, positive and negative, that deal respectively with significative *gain* or *loss* in resources (leading respectively to *joy* or *sadness*). *Positive emotions* act so as to reinvest the gain (that triggered them in the first place) in order to buy access to further resources: they operate as accumulators or *amplifiers* (Isen, 1993). *Negative emotions*, on the other hand, are used so as to fill the breach, slow down the running waste, and eventually to call for help so as to replace the incurred loss: they operate as *regulators*.

Of course, these processes are themselves costly and it might not seem too adaptive to spend so much resources when one is already losing some. Yet, even if one has to pay for the plumber, it might well be worth to do so before completely running out of water! Sometimes, this protection function of negative emotions against loss would rather be met by setting the agent in economy mode, i.e. by isolating it from the others and reducing its activity level, as often happens with people suffering from grief or severe distress.

Uncertainty and temporal dynamics. On the other hand, emotions being (as motivations) intrinsically related to the goal structure of an agent, they necessarily have to also be related to planning and expectations. That is to say, emotions must be subject to a *temporal dynamics*, and potentially then be triggered by actual as well as by *anticipated* gain or loss (leading to *hope* or *fear*). While actual gains would be reinvested to obtain further resources, anticipated gains would on the contrary insure goal persistence, by acting so as to justify sustaining the actual investment, or even spending more resources in the pursuit of the current goal. While actual loss would trigger the calling for extra resources from other agents, anticipated loss would on the other hand induce significative change in the current planning strategy. In order to do so, the required resources would generally have to be subtracted from other current goal-pursuing processes that would then get temporarily (or even definitively) suspended.

Agency: causal attribution of events to other agents. Moreover, in multi-agent societies, gains and losses are most frequently caused by the action of agents (self or others). In other words, due to the very structure of commitments as second-order resources, agency is, in those societies, the general cause that a resource management system should naturally (i.e. *evolutionarily*) get a grip upon!

Before classifying emotions according to their appraisal structure, let's first recall what is agency in the psychology of emotions. *Agency* is understood as *causal attribution* (Weiner, 1985; Shultz, 1991): if something important happens to agent A, this one tries to attribute the responsibility to some other agent B. This is a way, for A, to get a grip on who can be held responsible for the considered upcoming event in order to react efficiently.

Anger is the process that gets control over situations when a significant loss of resources is attributable to the action of another agent (Hutchinson, 1972; Averill, 1983). On the other hand, if the loss is thought to be caused by self, it is *guilt* that steps in (McGraw, 1987). When someone else is seen as responsible for some noticeable gain, *gratitude* is triggered. Finally, *pride* takes charge when it is self that appears clearly responsible for the gain.

Joy, *sadness*, *hope* or *fear* are triggered either when no attributable agency can be found as a cause for the loss or gain, or when there appears to be no possible control over the corresponding event. From the *point of view* of control structures over commitments, non-controllable agency is tantamount to no agency at all, and there is no point spending resources against overpowerful agents: better call for help then or flee (and alert relatives to flee as well)! For instance, in cases of anticipated loss attributable to others, anger would be triggered *only if* there is something to be done against the agents to prevent their wrongdoing, while fear would be triggered if nothing could be envisioned to prevent it.

4.3.3.2. A taxonomy of emotions

Figure 4.2 summarizes the preceding categorization of emotions. Although much refinement could be introduced within each category (such as distinguishing between guilt and shame), this taxonomy is pretty much congruent with the intersecting aspects of the

current *appraisal theories* in the psychology of emotions (Ortony, Clore and Collins, 1988; Roseman, 1991; Scherer, 1988a; Smith and Ellsworth, 1985).

Figure 4.2
The appraisal structure for the basic categories of emotions

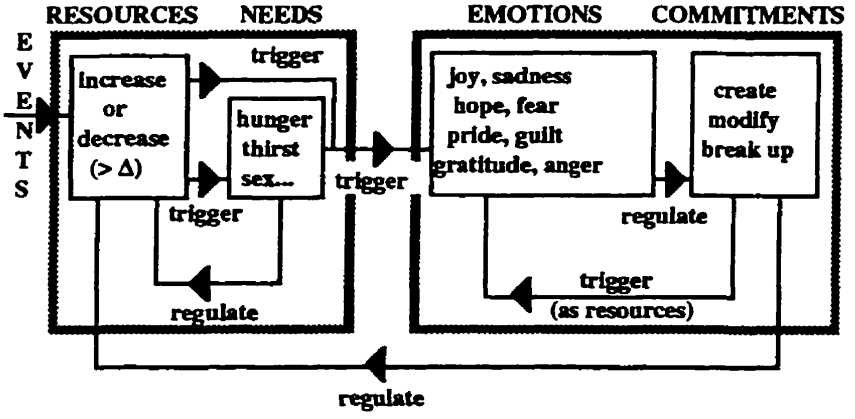
AGENCY	POSITIVE		NEGATIVE	
	actual	anticipated	actual	anticipated
none	Joy	Hope	Sadness	Fear
others	Gratitude		Anger	
self	Pride		Guilt	

The first four emotions (joy, hope, sadness and fear) act so as to set up or generate new commitments (through giving and sharing resources, strengthening affiliation, building up trust and confidence, calling for help and support...) between agents. The last four emotions (pride, guilt, gratitude and anger) act so as to handle or prevent the breaking of current commitments between agents (through praising or threatening...) or to strengthen and sustain them.

It is perhaps mandatory to stress here some theoretical consequence of having restricted emotions to being *commitments operators*. This means that we would not count as *fear proper* the escape reactions of lower organisms such as fish, amphibians and most of the reptiles (that happen not to have evolved limbic structures), but as being part of a simpler *need structure*. Our basic criterion for counting a motivational process as an emotion, is that it should provide for some communicative means in the service of interagent binding (such as alarm or distress calls, warning growls, tail waving...).

Figure 4.3 illustrates the model we have sketched above for the flow of control relative to the management of resources. It could be seen from this diagram that needs could themselves activate emotions in their attempt to regulate a significant variation of resources, for instance when a hungry baby gets distressed or even angered at not being fed.

Figure 4.3
The flow of control in the management of resources.



It also seems to be the case that first-order resources could activate emotions directly, when the variation is large enough to cross the critical threshold so as to *launch* emotional expression *evolutionarily* designed to commit others to set in and react (so they would help or flee, applaud or repent...): states of exhaustion often lead to distress, and sudden intense pain is a well documented releaser of anger (Berkowitz, 1993).

4.3.4. Emotions in the context of agent communication

4.3.4.1. Agent communication deals with commitments

In the context of MAS, communication has to be about interagent binding, resource sharing, cooperation or conflict, and then has to rest upon establishing and managing commitments. In other words, communication has to be very closely related to the emotional control mechanisms. Obviously, communications will not all be emotional: variations in commitments (as resources) do not always exceed the critical threshold (as stated above) beyond which emotions are triggered. But all emotions - as commitment operators - sure have to be communicative (i.e. they have to make use of communication methods)!

4.3.4.2. Speech acts are communication primitives

Both linguistic and DAI communities (Cohen and Levesque, 1990a; Cohen and Perrault, 1979; Singh, 1994; Winograd and Flores, 1986) agree that, in human languages, the basic units of communication are *speech acts*, i.e. various ways of specifying *intentions*. Classically Searle (1979) and later Vanderveken (1990), reformulating Austin (1962), both provide a taxonomy for these communication primitives.

4.3.4.3. A taxonomy of illocutionary speech acts

Speech acts capture categories of statements (typically centered around verbs in human languages) as various ways of specifying intentions - and, as a consequence, commitments (Winograd and Flores, 1986) among communicating agents, and thereby introduce a taxonomy of communication primitives:

- Assertives *commit* the speaker to some information being reliable.
- Directives request from someone to *commit* to accomplish some action.
- Promissives *commit* oneself to accomplish some action for someone else.
- Declaratives *commit*, and request that others do as well, to some change in status.
- Expressives express some feelings or emotions.

Expressive speech acts. However there are a few well documented problems with this taxonomy. On the one hand, Austin (1962) and Searle (1979) themselves argue that the expressive category has indeed a special character ("a very funny one"), and most further formalisms of speech acts bluntly ignored it (Cohen and Levesque, 1990a; Cohen and Perrault, 1979; Singh, 1994). We claim that expressive speech acts cluster intentions that should not belong to the same processing level as those from other categories, in particular because they correspond to different processing priorities. We would rather consider the expressive category as *emotional acts* (vs speech acts). Our contention is that, although this category also has to do with managing and conveying intentions, it simply does not stand on the same level as the others, and rather corresponds, on the verbal level, to a *vestigial* form of a more primitive underneath layer for the handling of commitments in social animals! It amounts to the verbal aspect of a much wider channel of emotional expression (emotional acts) that already flows through various facial, postural, or sound (e.g. crying, yelling, growling, laughing...) behaviors.

We contend that agent communication should *shift back from speech acts to emotional acts* whenever the threshold in the variation of second-order resources is exceeded or crossed over. Hence, communication between agents should non only reflect the verbal aspect of typical human communication but most importantly emotional nonverbal aspects as well. The choice of a platform for a testbed should take this into account: each script used by an agent for communicating with others would have to embody the specific way - in terms of speech acts or emotional acts - through which it operates upon commitments. One critical point here about the *ontology* of communicating agents is that they should be hardly able to ignore (or escape from responding to) emotional signals emitted by other agents, so that those signals could, for instance, *generate interrupts* or *reset computation*.

Declarative speech acts. On the other hand, the case for declaratives is also peculiar: they appear to be a very special device for using some authority (i.e. agents with a high status - like judges in our societies) as a third party or witness to guarantee that a whole community of agents would at once subscribe to (i.e. get bound to) a new commitment (which for instance happens when a new status is ascribed to some agent - "I name you President" - or when some object of collective usage is given a new name - "From now on, this street will be called Kennedy Avenue"). This peculiar kind of commitment, shared by a whole community at once, is closely related to what Shoham and Tennenholtz (1995) termed a *useful social law* in computational environments.

Assertive, directive and promissive speech acts. It then looks as if classical categories of speech acts did in fact encompass three different *layers* of communicating tools for setting up and managing commitments between interacting agents. In the middle lay the categories most frequently referred to and formalized within DAI literature and practice, namely assertives, directives and promissives, that deal with common ongoing transactions between agents: exchanging information, requesting and promissing, bidding and contracting...

Underneath lies the more primitive (but also more powerful, controlwise) layer of emotional acts, that sometimes emerge as expressive speech acts when they reach the verbal level. Emotional acts are triggered when there is serious risk of breaking ongoing commitments or unforeseen opportunities of establishing new ones. Serious risk here

means that the variation in second-order resources is exceeding some predefined threshold.

Finally, the top layer has to do with declaratives, that are used to set up new social laws: they hence provide a mechanism for establishing and *distributing* new commitments at once across a whole community of agents. They rest on the essential condition that the emitter's status should be high enough to enforce the new regulation, so that receivers belonging to the same community would as such be constrained to respect it.

4.3.5. Future works: what kind of implementation?

The choice of a proper testbed for the kind of control structure we have in mind is not obvious. First, the problem has to be complex enough so as to require distributing computation over a small community of agents, yet it has to remain manageable enough so we could draw useful conclusions from its implementation. It also seems interesting that it would be resonant or consonant with the kind of problems inhabiting the DAI culture. Finally, we feel that the nature of the problem should - when raised in human or animal societies - frequently, if not inherently, call for *emotional means of resolution*. Distributed meeting scheduling emerges as a rare case satisfying enough with respect to all such criteria.

In the context of meeting scheduling (agendas), relevant first-order resources are: periods of time (to be allocated for the various meetings); processing priority; knowledge (about self or others' available periods of time, or constraints, or preferences or commitments...). Relevant second-order resources are: commitments over time periods reserved for the meetings; reliability of the knowledge exchanged. Special kinds of commitments have to do with: *alliances* (or affiliation between agents that would then feel committed to cooperate with each other); *status* (i.e. allocated roles and power, that could sometimes enable obtaining, or even imposing, consensus among subordinates).

From what we said in the beginning of this article regarding motivations, it should be clear that goals (for instance, setting some specific meeting) are basically about *getting resources* (and from then on, preserving them) or about *setting commitments* (and from then on, keeping them). But this does not preclude that they could also bear upon exchanging resources (i.e. giving away some resources in order to *buy* other ones that are

required for the current activity), especially when this leads to creating new commitments. Plans are means to achieve goals, i.e. sequences of primitives or procedures (methods) owned by the agent, that could be assembled and executed so as to get to the expected results. As such, plans could be counted as part of the knowledge resource. From the above, it is easy to see that an agent's emotions are to be very closely related to its goal structure. For instance, blocking access to a required resource, would generally amount to goal blocking, which surely is one universal way (i.e. wide across cultures - and even across species!) of triggering anger (Averill, 1983; Hutchinson, 1972; Oatley, 1992).

In the context of meeting scheduling, there necessarily have to be many agents involved at the same time, each pursuing many different goals, having to keep and fulfill various commitments, sometimes requiring help and additionnal resources from a small set of intervening emotions. In terms of implementation, this means that not only agents should be represented as dynamic entities and operate in parallel, but this should be the case as well for goals, commitments and emotions! That is, all four concepts have to be represented as actors operating concurrently by sending and receiving messages asynchronously (Agha, 1986).

An agent sets a goal when asked for a new meeting, or as a subgoal of a current goal (while following a plan). Achieving goals can mainly be done through communicating with other agents and making commitments with them. Of course, some checks and bookkeeping also have to be performed within the agent's internal state, be it only to make sure that the target period is free, or that the request does not interfere with other commitments. Sometimes, meetings can be settled quite easily, when there is no conflict as to the allotted period of time. But with many people from many different groups, there are problems to be solved (and plans to be unfolded) often enough, especially as time goes by and as more periods get allocated (hence becoming unavailable for further meetings). Then an agent might think of reconsidering a commitment concerning some *minor* meeting, or try to make use of alliances and status as resources in convincing others to revoke theirs. It is typically on such occasions that emotions step in!

In the spirit of this testbed, one should envision that some requests have high enough priority (as when the meeting is about getting a doctor for an emergency or if the requester has a much higher status than self) so that reactions such as fear or distress would appear practically unavoidable. In order to be able to interrupt ongoing activity, emotions actors

should be allocated a higher processing priority than goals actors or commitments actors. Emotions are to be triggered when current goals or commitments go awry.

The implementation would use the Actalk actor platform built on top of ST-80 (Briot, 1994). It would take advantage of the Smalltalk exception handling system in which an exception is defined as a situation when computation could not be completed (Dony, 1990). When such a situation can be clearly described, it could also be foreseen and prevented by having a signal watch for its occurrence. Again, as was stated before, proper signals to raise for the triggering of emotions have to do with significant variations in the stock of resources, more precisely those involved within the critical goal or commitment (the one which caused problem). Once a signal is caught, control is to be transferred to a well-tailored handler that could execute some appropriate piece of code before resuming or halting computation.

4.3.6. Conclusion

Effective resource management in a distributed system requires a powerful control structure yet to be founded. We proposed a model of control, rooted in the psychology of human emotions, that introduces *emotion-like control structures* in the context of multi-agent systems. We pointed out how such control structures fit naturally into an inter-agent communication scheme. Their introduction improves the potential autonomy of agents by taking explicitly into account the management of the resources that are the core of their activity and that should hence be part of the fundamental definition of more autonomous and adaptive multi-agent systems.

4.4. What are emotions for? Commitments management and regulation within animals/animats encounters[4]

Abstract. The hypothesis that emotional mechanisms are wired in the repertoire of higher species is not as new as is its impact on the nature of autonomous agents underlying animats' design. In an evolutionary perspective, it looks as if, the more rational a species gets to be, the more emotional it happens to be as well. Thus, the study of emotional behavior raises the question of adaptiveness: what are emotions for, in terms of benefits for the individual? This paper attempts to answer such questions by proposing a model that links autonomy to motivation, to resource management and to emotions. The model distinguishes two kinds of resources: first-order ones, directly consumable, and second-order ones whose access is only guaranteed by a mediating agent, committed to give them. The model defines needs (such as those associated with states of hunger or thirst) as motivational processes insuring management and regulation of directly consumable first-order resources, while emotions (such as anger or fear) appear as motivational processes insuring management and regulation of second-order resources - namely commitments - within complex species for which they become the most important of all resources, and the key for achieving multiple goals. The paper concludes that emotions, like needs, are powerful control structures that have been designed in to protect agents and societies of agents from running out of resources. Such mechanisms meet some essential requirements that autonomous agents are faced with.

4.4.1. Introduction: a short story by K. Lorenz

> I have talked before about Bully, my Bulldog. He was already old, yet still brisk when I got hold of Hirschmann, the Hanover Bloodhound - or should I say when this one took possession of me. His arrival was a rude blow for poor Bully, and if I could have foreseen how much he would suffer from jealousy, I probably would have resisted Hirschmann's seduction. The atmosphere was tense for days until it exploded into one of the most violent dogs' fights I have ever witnessed [...] As I was trying to separate the fighters, Bully bit my finger accidentally. It was the end of the fight, but poor Bully had been struck with the greatest shock a dog's nervous system could ever register. He fell into depression, and although I did not punish him, but to the contrary caressed and fondled him, he remained paralysed down on the carpet, a round ball of despair unable even to get up [...] It was days before he started eating again, and even then we had to feed him from hand. For weeks, he would come to me only with an attitude of humble supplication, in total breach of his usual temper of a fierce dog that had never been servile. Now the interesting thing [is that this dog] had never bitten anyone before, and hence had never been punished for such a fault. (Lorenz, 1970, pp. 232-235 - the authors' translation)

---

[4] Publié sous: Aubé, M. and Senteni, A. (1996b). What are emotions for? Commitments management and regulation within animals/animats encounters. In P. Maes, M. Mataric, J.-A. Meyer, J. Pollack, and S. W. Wilson, (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* (pp. 264-271). Cambridge, MA: The MIT Press/Bradford Books.

Although filled with much anthropomorphism, this brief account of Bully's fate certainly does not sound too unrealistic to anyone who has shared a pet's life for some significant amount of time. The interactions thus experienced, *within and across species,* seem loaded with the emotional flavours of joy, attachment, anger, sadness, jealousy, and even, as suggested here... guilt! And as Lorenz suggests in the punch line, there is good reason to believe that Bully's complex reaction has not been conditionned, but is a natural one in such a situation, already built into the animal's repertoire.

To the behavioral scientist, one obvious question is why should it be so, what usefulness lies behind those reactions? In other words, emotional behavior raises the question of *adaptiveness:* what profit has been gained through evolution for members of a species to feel joyful or angry or fearful or guilty, what is the added survival value over individuals belonging to species - typically simpler and evolutionary older ones - that do not ever get emotional? Are there any systematic reasons for the fact that the more rational a species gets to be, the more emotional it seems to get as well?

The main purpose of this paper is to propose a coherent line of explanation of emotional behavior in order to answer the questions stated above. This attempt relies upon concepts borrowed from Artificial Life (ALife) and Distributed Artificial Intelligence (DAI) and is part of an ongoing project focusing on the use of emotional structures as a metaphor from which to specify powerful control structures for Multi-Agents Systems (MAS) (Aubé and Senteni, 1995, 1996a). An additional aim is to stress the theoretical importance of the motivational aspect of behavior and to show how this dimension could profitably be integrated with the current theorizing about rationality and goal-oriented behaviors within cognitive science.

The second section of the paper enumerates some of the requirements autonomous agents are to be faced with. The third one looks at motivation as a prerequesite for autonomy, links autonomy to resource management, and sketches two layers of control that developed on top of one another, in order to insure such management. In the following section, emotions as theoretical constructs are then deliberately restricted to the social domain, as some kind of interactional devices, that act more precisely as *commitments operators* between individuals to regulate the additional resources (termed second-order here) that other agents may provide to one another. It also considers two special kinds of

long term commitments, affiliation and status, that bear consequences upon the postulated design. Future work is forecast as a conclusion.

4.4.2. Requirements for autonomous agents

If one is to understand the architecture of living organisms so as to design artificial agents that are to prove adaptive in complex environments, what requirements are these beings likely to be faced with for them to be robust, and what kind of ontology would be apt to meet such requirements?

Open systems. One inescapable requirement is certainly that these agents will have to operate within open systems (Hewitt, 1985, 1991) wherein continuous change is the rule, and where the predictability of important events could sometimes be rather low. Hence there is the need for a sufficient amount of reactivity to those external events. The behavior of such agents is also to be restricted to "arms' length relationships", with typically no possibility for a global view over all the relevant aspects of a problem or situation. This raises the need for negotiation, so as to obtain complementary pieces of information from other agents.

The subsumption constraint. On the other hand, natural agents are to be "evolvable" creatures, submitted to the subsumption constraint (Brooks, 1991b). This means that they should not have been faced with too abrupt changes or global reorganization in their design. Hence their real-time reactive capabilities have to result from the operation of successive layers of control. So there is an argument in favor of the functional additivity of those interconnected layers, in such a way that, at each level, the design proves adaptive in itself, but that every successive layer adds new functionalities without disrupting the preceding ones.

Augmented planning. Being immersed in open systems, these agents also reach for goals through augmented planning (Oatley, 1992). They are indeed complex enough to have multiple goals of different kinds, of comparable priority, to be handled mostly in parallel. Yet they are faced with limited (at arms' length) knowledge and have to compete for scarce resources. And they have multiple agents to deal with, which raises issues of communication, cooperation, conflicts, negotiation, organization...

Self determination of goals. Being decentralized with encapsulated knowledge bases and embedded reactive behaviors, those agents have to display a fair amount of autonomy (Maes, 1991b; Castelfranchi, 1994). But mere "executive autonomy" whereby one is only free to execute whatever command comes from outside, is clearly not sufficient here. Real autonomous agents should have the capability for self determination of goals. This means that, more often than not, they will set and pursue their own interests, with the consequence that *they could not always be assumed to be sincere or benevolent.* In pursuit of their goals, they will also have to influence other agents of their kind. This implies that, by virtue of sharing a common ontology, such agents should themselves be liable to dependence and influencing.

So the stakes are fairly high, and the challenge of conceptualizing agents that are to meet such requirements, even in the simplest possible case, is no easy task. One contention of this paper is that the underpinning of motivation provides one essential line of attack for this design problem.

4.4.3. Motivation is about resource management

In specifying "a principled theory of agency and autonomy", d'Inverno and Luck (1996) defined autonomous agents as progressive refinements within a hierarchy of computational entities: *Objects* are first defined as generic *Entities* with (declarative) attributes and (procedural) capabilities. From then on, *Agents* are seen as *Objects* with goals, and finally *AutonomousAgents* are conceptualized as *Agents* with motivations. These authors further defined motivation as "any desire or preference that can lead to the generation and adoption of goals and which affects the outcome of the reasoning or behavioral task intended to satisfy those goals" (p. 73).

Indeed, the concepts of autonomy and motivation are intimately related: if agents are ever to become autonomous, it will only be through getting control over their motivations, by having ways of setting their own goals. But what is motivation about in the first place or, as Ferber (1996) would put it, where do goals ultimately come from? Our claim is that motivation basically has to do with *managing resources*, a task which is most crucial and most vital for the agents, and which hence becomes their ultimate goal. Such thinking is indeed what made behavioural ecology (Krebs and Davies, 1987) so efficient in

explaining much of animal behavior, and it seems hard to imagine that artificial autonomous agents could profitably be designed otherwise!

The next step is hence to specify what is to be counted as a resource. In the agent context, we define resources more or less as energy is defined in physics: whatever enables - and is required for - the production of work. It could be physical energy like heat or electricity or food, but it could also be other things: time, money, affiliation, power, knowledge... For instance, if a robot has to get to some target point P, this should cost it some amount of physical energy, be it electricity or fuel. Now, if the animat learns about a shorter route than usual to P, this specific knowledge is clearly worth some economy in fuel: it could hence be counted as a resource, and even be quantitatively measured in terms of the energy saved. Likewise, power over other agents is a way of harnessing their resources to benefit one's own, and could as such be counted as a proper resource. Basically then, the concept of resource is tantamount to the *cost* of work or actions.

However it seems useful to refine the concept and to distinguish at least two different kinds of resources: first-order and second-order. First-order resources are directly consumable ones that an agent already has at his disposal (available food or water for animals, processing time or memory for computers). Second-order resources, on the other hand, are those that an agent gets only indirectly, through having other agents *committed* to give them to him: newborns can get access to food only through having their parents provide the resources for them. Hence, the basic distinction between first-order and second-order, is that the latter rests essentially upon a mediating agent. The apple one has at one's disposal for lunch is first-order, while a friend that said he would bring one is second-order. More specifically, it is not the agent *per se* that is the resource, but whatever makes it so that one can count on him: namely, *the commitment is the resource!*

It should also be clear that commitments are not mere "potential resources": some food across the river, that one could try to access, could be seen as such. But the definition of second-order resources proposed here relies upon some other agent, that one counts on. *It is intrinsically bounded with multi-agent societies.* This point is quite crucial: the commitment is a resource in mediating (insuring) access to some commodity, in a similar sense that a vehicle could become a resource by allowing access to some remote commodity. Now, this does not mean that commitments have to be formulated in a conscious or reflective way - attachment structures, for instance, do involve commitments

between caretakers and infants that are presumably triggered through the activation of wired-in releasers and partially built-in structures (Reite and Field, 1985). In niches of nurturing species, getting hold of the process of establishing and regulating commitments very likely became a requirement. Indeed, for newborns of those species, it is a life or death matter!

Now commitments have received various definitions in Distributed Artificial Intelligence. On the individualistic side, they are conceptualized as persistence in intention or goal pursuing (Cohen and Levesque, 1990a, 1990b; Dongha, 1994). On the other extreme of social determinism, they are seen as an emergent property of agents being involved in many interlocking courses of action (Gerson, 1976; Gasser, 1991). And there is the middle-road interpretation of mutually agreed constraints (through promise or contract) on action, belief and world states (Fikes, 1982; Winograd and Flores, 1986; Bond, 1990). These different definitions led in turn to various "translations" or implementations of the concept: as logical conditions for intentions to endure (Cohen and Levesque, 1990a, 1990b; Dongha, 1994); as "agreed upon" actions, beliefs or goals (Fikes, 1982; Winograd and Flores, 1986); as production rules for the control of execution (Shoham, 1993); as embodied within plans, conventions or social laws (Bond, 1990; Jennings, 1993; Shoham and Tennenholtz, 1995); as quantified in terms of resources allocated to pursuing one's intentions (Gerson, 1976; Bond, 1990; Dongha, 1994).

The definition in terms of "persistence" seems too individualistic for planning agents embedded in "open" MAS. On the other hand, the more radical "emerging view" coming from social interactionists fails to offer concrete ways of representing how such interlocking gets bootstrapped in the first place within the agents' mental world. Our stance is that the middle-road view of mutually agreed constraints through promise or contract, could better meet the requirements, and even be translated into precise quantifiable terms. Commitments could indeed be measured in terms of allocation of resources and constraints upon resources (Gerson, 1976; Bond, 1990; Dongha, 1994). Typically, an agent A *commits* certain of its resources to another agent B (commits stands here for *does not give right away, but promises to give eventually*). As a consequence, B is *allocated* new resources, while A is *constrained* as to the usage of its own resources. For instance, A might be constrained not to consumate all of them and to take into account some amount that has to remain available to fulfill the commitment. On the other hand,

such commitment could also involve for A setting the goal of acquiring for B some resources that A does not already possess.

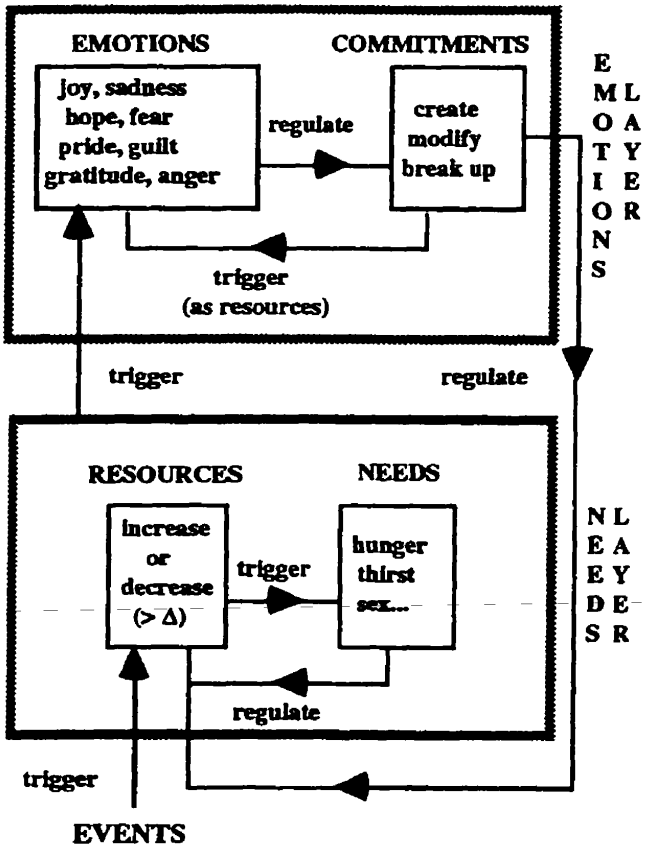
Being specifically defined in terms of resources, commitments thus appear as one critical reason for the setting up of MAS: by regulating exchange, they enable the agents to use one another as a way to access additional resources that would otherwise remain out of reach, or too costly to obtain. This refers to what (Castelfranchi, 1990) has called the *Sociality* problem: "why does an autonomous agent enter into social relations" (p. 50) and is congenial with this author's proposed answer: "Sociality was invented to multiply agents' powers of achieving their goals (using the powers of other agents)" (p. 56). *Within complex species that pursue multiple goals, and who absolutely require as such the cooperation of other agents, commitments might well become the most important of all resources! One important corollary is that, MAS thus require an effective control mechanism to regulate and control commitment fluctuations.* Hence, it is very likely that, in social species like ours, evolutionary forces would have favored and selected for precisely those organisms that would have become equipped with powerful mechanisms to insure control over commitments!

Within animals, *needs* - such as those associated with states of hunger, thirst, fatigue... - are the motivational processes (Toates, 1986) that insure management and regulation of directly consumable first-order resources, be it available food or water, rest and sleep, security or reproduction. In higher (social, or at least nurturing) animals, *emotions* - such as anger, fear, joy, sadness, guilt... - are the motivational processes that insure management and regulation of second-order resources, namely commitments. Very much like needs, emotions are powerful control structures that act so as to protect agents - and societies of agents - from running out of resources. In terms of the ***subsumption*** principle, they could be conceived as an additional layer of design that developed on top of the underneath layer of needs, using much of its control power, but taking over whenever the access to resources requires the cooperation of other agents. Figure 4.4 illustrates the operations of those two layers of control and their presumed interconnections.

As for needs, the emotional control processes are triggered by significant variations in the flow of resources, specifically second-order ones, for instance when there is a threat of breach in commitments or an opportunity for establishing new ones. Significant variations here means that the amount of increase or decrease of resources should exceed a certain

predefined (although eventually ajustable) *threshold*. Certain events are likely to have an impact upon an agent's stock of resources. When those variations cross a critical threshold, needs are called in so as to regulate them. In more complex (social or nurturing) animals, other agents can supply for the missing resources, provided they are committed to behave so. Emotions are designed in to regulate the operation of this new layer.

Figure 4.4
Two layers of control for resource management.



Now, having restricted emotions to being commitments operators, leads us to distinguish between different motivational reactions, that we would not count as "emotions proper". For instance, the escaping reactions of lower organisms like fish, amphibians and most of the reptiles, that happen not to have evolved limbic structures (MacLean, 1993), are thought to rest upon a more primitive layer for handling needs (security being involved here). Our basic criterion for counting a motivational process as an emotion, is that it

should provide for some communicative means in the service of interagent binding (such as alarm or distress calls, warning growls, tail waving...), even if the corresponding message is not actually sent in some particular situation.

In the case of a human reacting with a jump to a loud noise, or stepping aside in order to avoid a speeding car, we would say that his "fear reaction" is more complex (subsumptionwise), basically exploiting simpler old structures, but also profitting from the "additional higher" possibility of yelling as well, and thus expressing distress (i.e. calling for someone committed to help, such as a mother) or anger (i.e. warning that the transgressor should retract and amend). There is much experimental data from the psychophysiology of emotions that provide some suggestive evidence for different origins behind the repertoire of human fears (Öhman, 1986) and for the fact that some of these behaviors rest upon much simpler and evolutionary older structures (Ekman, Friesen and Simons, 1985). *It is our contention that reserving "emotions" for the control structures that are responsible for commitments management will do much to help disambiguate the term itself and specify the possible design of these structures.*

4.4.4. The emotion engine

Triggering signals. Now let us sketch how this emotion layer could operate. First we must specify the signals which are going to trigger the regulating processes. Just as needs are activated when variations in first-order resources pass beyond some threshold (Toates, 1986), emotions are triggered by significant gains or losses in commitments (second-order resources), that could be evaluated numerically in terms of the corresponding first-order resources they guarantee access to. This would lead to positive and negative emotions respectively. Positive emotions act so as to reinvest the gain (that triggered them in the first place) in order to buy access to further resources (e.g. establish new commitments): they operate as accumulators or *amplifiers*. Negative emotions, on the other hand, are used so as to fill the breach, slow down the running waste, and eventually to call for help so as to replace the incurred loss: they operate as *regulators*.

Processing priority. As to the formulation of proposals for the operation of the emotional processes themselves, Herbert Simon was the first scientist within the AI community, to dare make an offer, back in the late sixties: "If real-time needs are to be met, then provision must be made for an interrupt system [...] all the evidence points to a close

connection between the operation of the interrupt system and much of what is usually called emotional behavior" (Simon, 1967). This link between emotions and the interruption of behavior had in fact been hanging around for some time in psychological literature (Angier, 1927; Mandler, 1964), but had not found a proper grounding, perhaps for a lack of the concrete embodiment the computer metaphor could offer. Be it as it may, it is clear that the processes raised when the proper signals are triggered should receive almost the highest processing priority, to enable them to easily gain control, if the current threat or opportunity appears to be really critical. *This is why we proposed that emotional processes should operate as a control structure.* Actually the best comparison we have in mind so far is some mechanism akin to the *exception handling system* (EHS) designed in languages like Smalltalk-80 (Dony, 1990), although a proper embodiment would be likely to require that the specifications be realized as close as possible to the machine level. An *exception* is defined as any situation that "threatens" the current computation from achieving completion.

> An EHS allows users to signal exceptions and to define handlers. To signal an exception amounts to (1) identify the exceptional situation, (2) to interrupt the usual sequence, (3) to look for a relevant handler and (4) to invoke it while passing it relevant information. Handlers are defined on (or attached to, or associated with) entities for one or several exceptions (according to the language, an entity may be a program, a process, a procedure, a statement, an expression, etc). Handlers are invoked when an exception is signaled during execution or the use of a protected entity. To handle means to set the system back to a coherent state... (Dony, 1990, p. 323)

Shared ontology. For such a system to operate as a regulator of commitments, another prerequisite is that all agents belonging to the same community have been built according to the same ontology, so they can not escape from responding to the emotional acts emitted by other members. Indeed, the expression of anger only makes sense if the recipient is constrained (through some built-in design) to react to it significantly (otherwise, it would likely have already been dropped from the behavior repertoire, through natural selection!). Hence, one should envision, for every category of emotions, some built-in detector that would be triggered if the appropriate reaction is mandatory (for instance if there has been goal-blocking, threat, danger or loss...), and some list of preferred actions towards the target agent, from which the system should choose. Here again, there is much work from the psychology of emotions that suggests some universal antecedents for a small set of emotions. A more detailed account of the appraisal structure

for basic emotions and their associated triggering signals is presented in Aubé and Senteni (1996a).

Emotional handlers. For each emotion from an animat's repertoire then, the appraisal structure would specify the appropriate signals in terms of the commitments concerned (whether threatened or favored) and also provide an adequate handler as a list of prefered actions, from which repair of the injured commitment would be carried out or any new opportunities exploited. Typical reactions for fear would involve fleeing, freezing, attacking, calling for help, broadcasting alarms... *One essential component of all handlers will have to deal with emotional expression*, the quasi-intentional structure of which we have compared elsewhere to the one specified for speech acts (Aubé and Senteni, 1996a). This communicative aspect (Oatley, 1992) has profound consequences for MAS theorizing in that it offers a key to the understanding of the emerging aspect of commitments envisioned by the social interactionists (Gerson, 1976; Gasser, 1991). The basic idea is that the expression of emotions does not merely act upon the inner mental states of the agents. *By the very same process, it also acts so as to distribute the problem (threat, danger, violation of standard, rupture of commitment...) as well as its computation over the whole community of agents, thus making for a very powerful tool of distributed control.*

Conversation network. Another important feature of our specifications is that all emotions be connected into a kind of conversation network, much in the spirit of the state transition diagrams for speech acts suggested by Winograd and Flores (1986) or the "interchange protocols" of Campbell and d'Inverno (1990). The point here is that emotional reactions are themselves events that are designed to act upon commitments and should in turn trigger emotions in the others. Depending upon the specific configuration of commitments between two agents, fear could thus be answered with fear (alarm), with anger (so as to impose more fear), with sadness (as in parents being sorry for their child's fear), with guilt (if one has unwillingly scared the other) or with pride (if one has made the opponent retreat), but never with joy, hope or gratitude.

Commitments as actors. Last, but most importantly, we have to make provision for how commitments themselves should be represented, so they could play the crucial role they bear within the whole architecture. One might think of a very simple numerical representation, since we advocated that they could be precisely measured in terms of the

amount of first-order resources involved: how much has been allocated from others, or how much from self is going to be constrained. But clearly, this would not suffice, since they would certainly have to register many more attributes, such as which acquaintance is concerned, or what the time range is for the commitment to be satisfied. Such requirements suggest that commitments could be envisioned as more active entities, operating concurrently, such as *actors* (Agha, 1986), that could themselves monitor their fluctuation with all the agents one is related to, and that could send messages to the appropriate emotional handlers whenever their value attribute is critically modified.

Long-term commitments. It also seems important to specify two special kinds of long-term commitments. One is "affiliation" (alliance, attachment, friendship...) defined as some kind of contractual dependency between two agents, that could be increased or decreased depending on past history of helping and cooperation. The other is "power" (or status?), defined as socially recognized capacity for requesting help or resources from others, or for imposing consensus among subordinates. Those would be represented as kinds of "stored commitments" that could be invoked in certain situations (just as one could accumulate first-order resources to be used in periods of starvation). Except for the time scale, they share all the attributes of regular commitments. Those two seem quite useful for interpreting Bully's behavior presented in the beginning of the paper. It is fairly clear that the intrusion of Hirschmann in Bully's life was felt as a considerable threat to the long-term commitment of affection between Lorenz and Bully, hence the frequency of the fights against the agent presumed responsible. But when Bully accidentally bit his beloved master, he himself betrayed at the very same time *two* strongly established long-term commitments - attachment and submission - which triggered a corrective handler of guilt with very high priority. The interruptive capacity of this one has to do with the strength of the affiliation, but also with the combined effect of having challenged the power commitment as well. Bully's guilt had to act upon Lorenz as an expression of submission and regret, but also upon Bully's internal states, in order that he would stamp the situation firmly in memory so as not to take the slightest risk of ever repeating it again.

4.4.5. Conclusion and future work

The nature of autonomous agents underlying animats' design is strongly conditioned by the hypothesis that emotional mechanisms are deeply wired in the repertoire of higher species. The central question addressed in the paper - "What are emotions for, in terms of

benefits for the individual?" - is partially answered by the proposition of a coherent line of explanation of emotional behavior, stressing the question of adaptiveness. The first contention of the paper is that the underpinning of motivation provides a line of attack to this design problem. Autonomy and motivation are shown as intimately related: if agents are ever to get autonomous, it will only be through getting control over their motivations, by having ways of setting their own goals. The second claim of the paper is that, basically, motivations have to do with resource management. That is why behavioural ecology is so efficient in explaining animal behavior and why it would be very surprising that artificial autonomous agents could be designed otherwise. This concern leads to a closer study of the concept of resource that can be refined by distinguishing first-order and second-order. First-order resources are directly consumable ones that an agent already has at disposal. Second-order resources are those that an agent gets only indirectly, through having other agents committed to give them to him: the commitment is a resource in mediating access to some good. Being specifically defined in terms of the resources, commitments thus appear as one critical reason for setting up multi-agent societies: by regulating exchange, they enable the agents to use one another as a way to access additional resources that would otherwise remain out of reach, or would be too costly to obtain.

The third claim of the paper is that, within complex species that pursue multiple goals, commitments become the most important of all resources, one important corollary being that multi-agent societies require an effective control mechanism to regulate and control commitments fluctuations. Within animals, needs, such as those associated with states of hunger or thirst, are the motivational processes managing and regulating directly consumable first-order resources while, in higher-order species, emotions, such as anger or fear, are the motivational processes managing and regulating second-order resources or commitments. Like needs, emotions are powerful control structures protecting agents and societies of agents from running out of resources. As for needs, these control processes are triggered by significant variations in the flow of resources, for instance when there is a threat of breach in commitments or an opportunity for establishing new ones. An additional contribution is to propose one clear way of implementing the desired control structure through the use of a mechanism similar to the exception handling system of Smalltalk-80. Signals are defined within the appraisal structure of basic emotions, and the handlers involve communicative emotional acts plus a list of preferred actions or partial plans to be used for the repair of the injured commitment. One final claim stressed the

theoretical utility of specifying special kinds of long-term commitments, such as affiliation and power.

Effective resource management in a distributed system requires a powerful control structure yet to be founded. Next generation parallel programs as well as next generation autonomous agents could benefit from a principled model of control, rooted in the psychology of human emotions, that would introduce emotion-like control structures. Such a mechanism would fit naturally into an inter-agent communication scheme and its introduction would likely improve the potential autonomy of agents by taking explicitly into account the management of the resources at the core of their activity. There is a lot of work remaining to be done along these lines, if we want to be able to express these control structures through readable and reusable high level abstractions. Another thread of work would be to integrate the concerns expounded in this paper with other high-order concepts such as reflection. It is our contention that emotion-like control structures that serve to regulate resource consumption, could offer one indispensable complement to reflective structures that are also designed in to handle problematic situations. Metaphorically speaking, it is somewhat like adding thermodynamics to mechanics for a more complete account of the processes that are to be described.

Chapitre 5.

Conclusion générale

5.1. Résumé de la démarche parcourue

Au cours de la présente recherche, nous avons tenté de rendre explicites les principales fonctions adaptatives des structures émotionnelles, et de mettre à jour certains des rouages par lesquelles elles arrivent à remplir effectivement de telles fonctions. Bien que d'abord intéressé par la question de la motivation scolaire, il nous a cependant semblé qu'une percée significative en ce domaine requérait une approche théorique fondamentale. Nous avons été en cela inspiré par l'approche de l'IA qui, dans son projet de "produire de l'intelligence", s'est rapidement vu confrontée au difficile problème de la représentation des connaissances. Ce qui n'avait semblé au départ qu'un sous-problème parmi d'autres, est ainsi devenu un enjeu majeur qui, durant les années 1970, a largement envahi l'agenda de recherche de cette discipline. D'une façon analogue, il nous a semblé que la compréhension des phénomènes de motivation pourrait se trouver considérablement éclairée par le projet d'en reproduire et d'en capter le caractère adaptatif au sein de systèmes artificiels.

La thèse a été présentée sous la forme d'une série d'articles posant différents jalons d'un modèle théorique et fonctionnel des émotions, comme fondement pour les structures de contrôle dans les systèmes multi-agents, en intelligence artificielle distribuée. Le premier article (chapitre 2, section 2.2) a commencé par situer le problème de recherche dans le contexte de l'éducation, en faisant ressortir la nécessité de disposer de modèles théoriques robustes pour aborder le domaine de la motivation et des émotions. Le texte a d'abord rappelé les principales contributions de l'IA pour la compréhension des mécanismes d'apprentissage et de construction des savoirs, et fait ressortir l'importance et l'utilité des spécifications ainsi obtenues pour la conception de stratégies pédagogiques et didactiques appropriées. L'article a ensuite proposé d'appliquer d'une façon analogue l'approche et les méthodes de l'IA pour la conceptualisation des phénomènes motivationnels et

émotionnels, et a présenté un aperçu du genre de modélisation qui pourrait en résulter, ainsi que des applications qui pourraient en découler afin de rendre plus efficaces les interventions pédagogiques destinées à rehausser la motivation scolaire.

Le second article (chapitre 3, section 3.2) a pour sa part consisté en une recension approfondie en psychologie cognitive des émotions. Sa fonction essentielle, dans le cadre de la thèse, était de repérer les concepts fondamentaux pour la formulation du modèle théorique et d'accumuler un large répertoire d'exemples, aptes à supporter ou à confronter la théorisation proposée. Nous avons notamment tiré profit des principales controverses en psychologie des émotions comme révélateur des concepts-clés et des articulations théoriques les plus significatives. De cette méta-analyse a émergé l'hypothèse que les structures émotionnelles constituaient une couche particulière de contrôle, entre celle qui assure la gestion des besoins et celle qui supporte le fonctionnement rationnel. La recension a également permis d'identifier et de retenir, comme composantes de cette couche de contrôle, un petit nombre de mécanismes de base apparaissant comme responsables de la très grande majorité des comportements émotionnels, et d'articuler avec précision les conditions de leur déclenchement.

Le troisième article (chapitre 4, section 4.3) a exposé le modèle "computationnel" proprement dit, et notamment le lien fondamental que nous avons établi entre motivation, gestion des ressources et engagements. Ces derniers y ont été expressément conceptualisés comme un type particulier de ressources, reposant essentiellement sur la collaboration entre les agents. Les émotions ont alors été présentées comme des opérateurs des engagements dont la fonction était d'assurer la gestion et la régulation des variations significatives qui pouvaient se produire à l'intérieur de ce stock de ressources particulières, essentielles pour rendre possible et maintenir de façon adaptative la coopération en tant que fondement des espèces sociales. L'article est d'ailleurs précédé (chapitre 4, section 4.1) d'une section approfondissant précisément les phénomènes de coopération et de réciprocité, ainsi que les conditions favorisant leur émergence et leur évolution.

Le quatrième et dernier article (chapitre 4, section 4.4) a examiné, dans une perspective évolutionniste, les différentes contraintes qui semblent avoir présidé à l'apparition (dans le contexte de l'évolution) ou devoir conditionner la conception (par l'ingénieur cogniticien) d'agents biologiques ou informatiques véritablement doués d'autonomie. La capacité pour un organisme de déterminer ses propres motivations y a été présentée comme la

caractéristique première de l'autonomie, et la gestion des ressources y a été abordée comme le principe essentiel derrière tout mécanisme d'ordre motivationnel. Les engagements étant envisagés dans notre modèle comme un type particulier de ressources, le texte a ensuite réexaminé à cette lumière le concept tel qu'il a été utilisé en IAD. L'article s'est finalement terminé par une énumération des spécifications qui découlent du modèle, en vue d'une implantation dans un environnement informatique à base d'acteurs.

5.2. Les contributions pour l'intelligence artificielle distribuée

Le moment semble donc venu de rassembler les différentes contributions de la recherche, notamment en ce qui concerne le rôle primordial joué par le concept d'engagement pour la compréhension des phénomènes motivationnels. Ce construit hypothétique, tel que défini et utilisé dans la thèse, offre en effet une valeur descriptive et explicative exceptionnelle. Au profit de la recherche en IAD, il a le mérite d'éclairer les relations qui existent entre motivation et gestion des ressources, il offre une base pour l'explicitation des contraintes qui président à la coopération entre les agents, et il permet de formuler des spécifications précises pour une implantation éventuelle de structures de contrôle remplissant les fonctions principales des mécanismes émotionnels chez les vivants. Au plan de la formalisation des intentions et des mécanismes de communication inter-agent, il contribue à éclaircir certains aspects de la théorie des actes de langages, notamment sur la question de la catégorie des "actes expressifs".

5.2.1. Éclaircissement des liens entre autonomie, buts, motivation et gestion des ressources

Le concept d'autonomie en IAD est pratiquement indissociable de celui d'agent. Et pourtant, hormis quelques rares tentatives d'éclaircissement (Castefranchi, 1994; Luck et d'Inverno 1995; McFarland, 1995; Steels, 1995), cette notion reste encore relativement floue. Nous croyons donc que les liens que nous avons proposé au chapitre 4 entre autonomie, motivation, génération des buts et gestion des ressources, apportent un éclairage salutaire sur cette constellation de concepts. Nous appelons ainsi "motivation" l'ensemble des mécanismes qui déterminent le choix des buts adoptés par un organisme ou un système quelconque à n'importe quel moment, et nous appelons "autonomie" la capacité pour un tel système de contrôler ses propres motivations. Nous posons alors

comme motivation première et ultime de tout système adaptatif, celle de gérer ses propres ressources: chercher à y accéder, chercher à les préserver, chercher à les renouveler. Nous irions même jusqu'à prétendre que n'importe quel but généré et adopté par un système "naturel" découle toujours, en définitive, d'un problème de gestion de ressources, ou à tout le moins d'une chaîne de buts intermédiaires ordonnée en fonction d'une telle gestion.

Certes, il existe des systèmes "artificiels" capables de générer leur propre structure de buts pour la résolution d'un problème donné. Mais ce genre de problème est généralement "commandé" de l'extérieur, souvent d'ailleurs par un "employeur" humain, et on parlera plutôt alors de ce que Castelfranchi appelle "executive autonomy" (1994, p. 58), c'est-à-dire une autonomie qui porte sur le choix des *moyens*, mais non sur la capacité de déterminer les *fins* elles-mêmes. En ce sens, ces systèmes ne constituent que des annexes - ou des "prothèses" cognitives - pour l'utilisateur humain, en qui réside la véritable source d'autonomie. La conception qui est proposée ici n'est d'ailleurs pas très éloignée de la compréhension intuitive du terme: ne dit-on pas en effet d'un individu qu'il est devenu "autonome" à partir du moment où il se montre capable de subvenir lui-même à ses besoins? Mais elle entraîne également des conséquences qui dépassent largement l'interprétation populaire, notamment que la description formelle d'un processus (sa "mécanique") reste fondamentalement *incomplète* tant que les dispositifs responsables de la gestion des ressources inhérentes à son exécution (sa "thermodynamique") ne sont pas également spécifiés. Nous osons croire que là réside l'une des conditions essentielles à une vision plus intégrée des aspects motivationnels et cognitifs du comportement.

5.2.2. Réification des engagements comme ressources de deuxième ordre

Une autre contribution essentielle concerne la reconceptualisation des engagements en termes de ressources, et plus spécifiquement en termes de ce que nous avons choisi d'appeler "ressources de deuxième ordre". Il y a plusieurs mérites à ce choix théorique. Tout d'abord, il présente un caractère explicatif en faisant ressortir l'utilité, sinon la nécessité des engagements au plan adaptatif, puisqu'ils confèrent à l'individu qui en dispose une plus grande valeur de survie. Nous avons examiné comment la coopération pouvait effectivement offrir des gains appréciables, mais aussi comment elle générait du même coup un espace de risques et de vulnérabilité à l'exploitation. Une telle vision rend alors plausible, au plan évolutif, que des dispositifs de contrôle et d'accès à ce type de

ressources plus sensibles, plus complexes, mais aussi tellement bénéfiques que constituent les engagements aient pu émerger et être sélectionnés.

En second lieu, en rattachant ce type de ressources à celles de premier ordre - plus facilement quantifiables - auxquelles elles donnent accès, il devient également possible de les comptabiliser et de les ordonner rigoureusement, et de définir alors sur la variation de ces quantités, des mécanismes de déclenchement extrêmement rapides et efficaces. Cette quantification est par ailleurs essentielle lorsqu'il faut choisir entre plusieurs engagements conflictuels, mais aussi afin de faciliter le cumul des engagements à long terme, tels l'autorité ou l'affiliation, qui ne cessent de croître ou de diminuer, oscillant au gré de l'histoire des interactions et des transactions entre deux individus.

Nous avons cependant fait ressortir, en troisième lieu, qu'un simple attribut numérique, bien qu'essentiel, restait insuffisant à capter toute la complexité du concept, et qu'une représentation plus structurée et plus dynamique, par exemple en terme d'objet concurrent ou d'acteur, semblait plus appropriée. Chaque engagement doit en effet constituer une structure de données enregistrant les partenaires impliqués, les ressources en jeu, l'échéancier de respect de l'engagement, les conditions de satisfaction ou de rupture, etc. Mais il faut également envisager un mécanisme plus dynamique, surveillant avec vigilance que l'engagement n'est pas mis en péril par d'autres événements concurrents, et capable de déclencher alors, avec toute la priorité requise, les mécanismes de protection, d'interruption ou de correction appropriés, que constituent les processus émotionnels.

Finalement, en identifiant comme ressource de deuxième ordre, non pas l'agent lui-même qui donne accès à une ressource de premier ordre, mais plutôt la "contrainte interne" qui l'amène à garantir cet accès, nous croyons avoir mis le doigt sur le dispositif motivationnel qui a été progressivement inscrit à travers l'évolution pour "harnacher" la règle de réciprocité et rendre possible une coopération "sécuritaire et rentable" entre les individus. Cette vision impose à son tour certaines spécifications. Pour que la gestion de ce type particulier de ressources puisse être assurée, il importe en effet que les agents partagent au moins certains des mécanismes de reconnaissance et d'expression de ces engagements, et de réaction à leur modification. Il faut en outre que ces mécanismes opèrent de façon contraignante, qu'ils disposent justement de l'automatisme caractéristique des structures motivationnelles, assorti d'une grande priorité d'exécution, rendant quasiment impossible leur ignorance. C'est ce que nous avons voulu signifier en spécifiant que l'existence

même d'un système multi-agents reposait sur le partage d'une *ontologie* commune, et sur l'existence d'une *structure de contrôle* pour la gestion et la régulation des engagements.

5.2.3. Les émotions comme prototypes de structures de contrôle pour les systèmes multi-agents

C'est aux émotions que nous attribuons cette fonction essentielle d'assurer la gestion et la régulation des engagements. Nous avons en effet tâché de montrer que, chez les vivants, du moins chez les espèces sociales qui protègent et élèvent leurs rejetons ("nurturing species"), les mécanismes émotionnels présentaient toutes les caractéristiques d'une telle structure de contrôle. Aptes à repérer, ou même à prévoir, les variations significatives à l'intérieur du stock de ressources de deuxième ordre, les émotions disposent également d'un répertoire d'actions sur mesure, affinées à travers l'évolution, de façon à prendre en charge ces situations problématiques. Nous avons donc proposé d'exploiter les connaissances accumulées en psychologie cognitive des émotions comme "métaphore" porteuse dans la conceptualisation des structures de contrôle requises pour assurer l'autonomie et la coopération chez les systèmes multi-agents.

L'intuition de Simon (1967) de considérer les structures émotionnelles essentiellement comme mécanismes d'interruption du calcul ("computation") en cas de risques imprévus ou d'opportunités inattendues, nous a incité à regarder du côté des structures de contrôle. Par ailleurs, le système le plus sophistiqué d'interruptions dont nous ayons pris connaissance, est celui qui a été conçu et utilisé en Smalltalk-80 pour la gestion des exceptions (Dony, 1990), et il nous semblait d'autant plus approprié qu'il s'appliquait également aux environnements informatiques à base d'acteurs. Ce système permet de définir au préalable les situations qui risquent de menacer la poursuite ou l'intégrité du calcul, et de repérer ces situations en spécifiant les signaux caractéristiques de leur occurrence. Des méthodes de prise en charge ("handlers") sont également prévues, et leur exécution est automatiquement déclenchée lorsque les signaux critiques sont rencontrés. L'examen attentif des résultats de recherche en psychologie nous a alors permis de proposer, à partir des propositions de la théorie de l'appropriation ("appraisal"), une structure cohérente pour la spécification des signaux de déclenchement pour chacune des huit familles d'émotions que nous avons retenues dans notre modèle. Nous avons par ailleurs illustré comment il était possible de tirer des propositions de la théorie des émotions de base et de l'approche par prototypes, un ensemble de scripts à exécuter

("handlers") pour une prise en charge adaptative des situations problématiques ayant donné lieu à chaque catégorie d'émotion.

5.2.4. Distinction entre actes de langages et "actes émotionnels"

Parmi le répertoire d'actions utiles à exécuter en cas de situation problématique, les "actes d'expression" constituent un sous-ensemble privilégié. L'un des mérites de la théorie des actes de langages est d'ailleurs d'avoir souligné que les énoncés verbaux sont aussi des actions, souvent plus efficaces que bien d'autres gestes physiques. Ainsi, les énoncés correspondant à ordonner, à supplier ou à promettre l'exécution d'une action entraînent généralement des résultats tout aussi concrets que de poser le geste physique lui-même. Dans le cas d'une demande d'aide, cela peut même constituer, comme pour un nourrisson affamé, la seule façon d'obtenir le résultat requis. Dans la perspective de l'IAD, ou un calcul complexe peut être réparti sur toute une communauté d'agents, les actes de communication sont particulièrement importants, puisqu'il permettent de *distribuer* la résolution d'un problème sur une multiplicité de récepteurs. Un cri d'alarme, par exemple, répartit le problème du danger sur une communauté d'agents, qui peuvent accourir à la rescousse, ou faciliter la retraite des plus vulnérables.

Nous avons mentionné que la catégorie des actes expressifs, pourtant clairement identifiée par les principaux théoriciens des actes de langage (Austin, 1962; Searle, 1969; Vanderveken, 1990), a toujours été plus ou moins laissée pour compte dans les diverses élaborations ou formalisations qu'a par la suite connues la théorie. Notre analyse des structures émotionnelles a cependant fait ressortir l'importance de ces actes d'expression, comme *véhicules spécifiques d'intentions* pour chaque catégorie d'émotion. Nous avons donc proposé que les actes expressifs constituent en fait des vestiges des "actes émotionnels" sous-jacents, opérant au niveau non verbal, d'une manière analogue aux autres catégories d'actes de langages (assertifs, déclaratifs, directifs et promissifs), qui agissent plutôt au niveau verbal. Nous pensons que ces actes émotionnels gagneraient à connaître une analyse tout aussi approfondie que celle des autres actes de langage, notamment en termes de spécifications de leurs conditions de réalisation et de satisfaction. Une telle vision aiderait sans doute à mieux comprendre l'évolution de la communication animale, comme mécanisme d'expression (et parfois de partage) des intentions, favorisant la gestion de la coopération et de la réciprocité dans les échanges. Ainsi le chien qui couine et trépigne "signifie" par là ses besoins en eau, en nourriture ou en promenade à son

maître, et "obtient" généralement par là même le résultat escompté, d'où précisément l'idée d'*acte* émotionnel.

5.3. Les contributions pour la psychologie cognitive des émotions

En psychologie cognitive des émotions, le concept d'engagement, selon la conceptualisation qui en a été proposée ici, constitue probablement la variable la plus déterminante derrière le déclenchement et l'expression des émotions. Il permet en outre d'établir une distinction plus claire que jamais entre les besoins et les émotions, il aide à mieux comprendre et formuler le type de contraintes qui ont pu contribuer à l'apparition des structures émotionnelles comme instruments inédits d'adaptation dans l'évolution des mammifères et des primates, il rend compte de la nature intrinsèquement sociale des émotions, et il offre un cadre cohérent pour la répartition taxinomique des différentes catégories d'émotions. Comme on a essayé de le démontrer dans le chapitre 3, il offre également une base solide de réunification entre la "théorie des émotions de base" et la "théorie de l'appropriation". Il offre même, comme on l'a mentionné au chapitre 2, une interprétation parcimonieuse d'une foule de problèmes en psychologie sociale et en psychopathologie.

5.3.1. Les engagements comme "variable critique" derrière le déclenchement et l'expression des émotions

Nous avons souligné à diverses reprises à travers cette thèse l'importance considérable que nous accordions au concept d'engagement dans notre modèle du fonctionnement des émotions. L'une des contributions essentielles de ce travail a en effet consisté à démontrer qu'il s'agit là d'un concept-clé, absolument incontournable pour la compréhension et la mise à jour de la dynamique émotionnelle. À titre de comparaison, en science physique, chaque fois qu'un chercheur se trouve confronté à un changement de mouvement, il sait qu'il doit reconnaître là l'action d'une *force*, et il cherche alors d'où provient celle-ci. Le concept de force prend alors une valeur exceptionnelle, puisqu'il guide avec certitude le chercheur dans le processus d'explicitation de la dynamique d'un événement. L'existence de la planète Neptune a ainsi été postulée pour rendre compte des perturbation observées dans la trajectoire d'Uranus, qui restaient inexplicables sans l'intervention d'une force

extérieure. *On peut alors dire que le concept théorique de force a constitué dans sa découverte un outil aussi efficace et indispensable que le télescope.*

C'est également le concept de force qui a mené à la découverte des trous noirs, pourtant invisibles puisque leur densité est si grande que même les rayons lumineux qu'ils émettent en subissent l'attraction. Mais justement cette même force gravitationnelle induit des perturbations importantes sur le mouvement des autres corps célestes au voisinage de l'astre "caché", ce qui a permis aux astronomes d'en inférer la présence. *D'une façon analogue, nous posons le concept d'engagement comme la "variable critique" et essentielle derrière tout comportement émotionnel.* Autrement dit, chaque fois que l'on constate ce genre de réaction, on peut être assuré qu'il y a un ou plusieurs engagements en jeu, et que ceux-ci expliquent plus que tout autre variable le comportement observé. Réciproquement, si l'on sait qu'un engagement a été sérieusement modifié ou compromis, on peut généralement soupçonner qu'une réaction émotionnelle a été ou sera déclenchée, même si son expression est gardée secrète. Car bien sûr, certaines émotions peuvent être retenues, et un bon comédien peut en simuler ou en falsifier l'expression. Mais l'équivalent existe aussi en physique, par exemple avec les phénomènes d'illusion ou de prestidigitation. L'important pour notre propos, c'est que dans le domaine de la mécanique, le concept de force constitue précisément l'outil de référence qui guide l'examen et l'interprétation des observations bizarres ou irrégulières. De la même façon, nous proposons le concept d'engagement comme principal révélateur de la dynamique émotionnelle.

5.3.2. Un nouveau critère simple et fonctionnel de distinction entre les besoins et les émotions

Bien que les besoins et les émotions soient largement reconnus comme représentant deux des principales catégories motivationnelles, les critères qui ont été traditionnellement proposés pour établir leur distinction sont loin d'apparaître aussi clairs et transparents qu'on pourrait le souhaiter, et ils ne font pas non plus tellement preuve de parcimonie. Ainsi, nous avons rapporté au chapitre premier (p. 11-12), ceux qui ont été proposés par Plutchik (1980, pp. 362-363; 1984, p. 214) et qui semblent rallier plusieurs théoriciens:

- les émotions sont habituellement déclenchées par des stimulations *extérieures*, alors que les besoins résultent de changements *internes* dans certains états physiologiques;
- les émotions surgissent en *présence* d'événements déclencheurs, alors que les besoins se manifestent plutôt en *l'absence* de stimulations à caractère homéostatique;

- les émotions (comme la colère) peuvent dépendre d'une *variété* d'événements, alors que les besoins dépendent plutôt d'objets *spécifiques*, tels l'eau ou la nourriture;
- les émotions *suivent* la perception et l'évaluation d'un événement, alors que les besoins *précèdent* au contraire la recherche d'un objet nécessaire à leur satisfaction;
- les émotions dépendent d'événements qui peuvent se produire sur une base *aléatoire*, alors que les besoins présentent plutôt un caractère *rythmique*.

La difficulté avec ces critères, c'est qu'ils ne découlent pas d'une théorie cohérente, mais constituent plutôt des justifications *ad hoc*, extraites d'exemples ou de cas particuliers déjà classifiés par le théoricien. C'est comme si celui-ci avait effectué cette classification au préalable, sur la base d'une compréhension intuitive, et qu'il en avait tiré ses critères après coup, en comparant les éléments déjà attribués à chaque catégorie. Il est par ailleurs étonnant qu'il puisse exister autant de critères différents, apparemment sans lien théorique ou logique les rattachant les uns aux autres. Une conséquence indésirable de l'applications de tels critères, c'est que différents théoriciens ne s'entendront pas sur la classification de certains comportements problématiques, surtout ceux qui, comme la peur, le désir sexuel ou l'agression, apparaissent fortement déterminés par des variations d'ordre physiologique, d'aucuns les rattachant alors aux besoins et d'autres les regroupant plutôt avec les émotions (Lapierre et Braun, 1993; Toates, 1986).

Le critère que nous proposons est beaucoup plus simple, et il découle directement de notre conceptualisation de la motivation comme mécanisme de gestion des ressources, ainsi que de la distinction que nous établissons entre ressources de premier ordre et ressources de deuxième ordre. Dans notre modèle, tout système motivationnel se traduit nécessairement par l'opération d'une couche de contrôle. Nous appelons *besoins* celle qui gère l'accès aux ressources de premier ordre, et réservons le terme plus "social" d'*émotions* à celle qui gère les ressources de deuxième ordre, c'est-à-dire les engagements. En plus d'offrir une base de distinction extrêmement claire - et solidement appuyée au plan théorique - entre les deux catégories de comportements, cette vision permet également de résoudre une ambiguïté qui a longtemps confronté plusieurs chercheurs (Averill, 1980, p. 325; Öhman, 1986; Scherer, 1993, p. 347; Wallbott et Scherer, 1986, p. 72), au sujet d'une émotion comme la peur qui semble effectivement comporter des variantes fort différentes. La solution que nous proposons à ce dilemme, c'est tout simplement que ces peurs ne sont pas toutes de l'ordre des émotions, et que certaines s'avèrent beaucoup plus frustes, étant plutôt de l'ordre des besoins, telles les réactions de fuite des amphibiens. Nous préférons réserver la connotation "émotionnelle" de peur aux réactions, beaucoup plus récentes dans l'évolution, qui impliquent (consciemment ou non, explicitement ou non) l'appel à l'aide

ou l'avertissement de fuir, afin de mobiliser ceux qui sont disposés à notre égard, ou afin de prévenir ceux que nous sommes nous-mêmes enclins à protéger.

5.3.3. Une base de réunification entre "Basic Emotions Theory" et "Appraisal Theory"

La méta-analyse que nous avons effectuée au chapitre 3, au sujet des principales controverses en psychologie cognitive des émotions, nous a amené à repérer quelques incohérences parmi les théories existantes, mais peut-être surtout d'importants points de convergence. Il nous a ainsi été possible de d'évacuer le dilemme qui avait animé le débat entre Zajonc et Lazarus au sujet de l'indépendance relative entre les processus émotionnels et les processus cognitifs - ou de la préséance des uns sur les autres - en conceptualisant les structures émotionnelles comme un couche de contrôle particulière, opérant entre celle des besoins et celle du jugement rationnel, afin d'assurer la gestion des ressources de deuxième ordre que constituent les engagements.

Par ailleurs, nous avons tâché de montrer que les théories de l'appropriation et celles concernant les émotions de base divergeaient surtout en ce qu'elles s'intéressaient à des aspects différents, mais complémentaires, des mêmes processus. En s'efforçant d'élucider la nature des évaluations requises pour le déclenchement des diverses réactions émotionnelles, les théories de l'appropriation ont surtout contribué à expliciter les antécédents des émotions, et nous ont permis de spécifier concrètement quels *signaux* devaient être mis en place pour identifier rapidement et efficacement les principales situations critiques qui risquent de menacer ou de compromettre les engagements. Dans cette perspective, les théories portant sur les émotions de base et sur l'approche par prototypes, offrent de leur côté un ensemble de spécifications précises, encapsulant sous le format de *scripts*, un répertoire diversifié de méthodes de résolution et de prise en charge (*"handlers"*) des situations problématiques ainsi repérées.

5.3.4. Une taxinomie simplifiée et cohérente pour les différentes "familles" d'émotions

En comparant les résultats accumulés par les principaux représentants de chaque "camp", il nous a semblé que les listes d'émotions retenues par chacun comme les plus significatives, affichaient un degré élevé de convergence. Nous avons alors appliqué sur chaque

candidat proposé des critères de parcimonie relativement sévères. Dans la perspective de simulation qui nous habitait, il nous semblait en effet important de faire correspondre le plus précisément possible, par delà la différence des termes utilisés et sur la base de leurs fonctions communes, la similarité des processus invoqués. La section 3.2.4 du troisième chapitre explicite ainsi comment il est possible de regrouper de façon cohérente la très grande majorité des observations recueilles en huit grandes catégories émotionnelles, clairement distinguables en termes de leurs mécanismes de déclenchement, ainsi que des stratégies de gestion des engagements qu'elles ont la fonction d'assurer. Ces huit grandes familles, correspondent approximativement aux vocables de joie, de tristesse, de peur, d'espoir, de gratitude, de colère, de fierté et de culpabilité. Nous avons spécifié également comment d'autres réactions, parfois proposées comme candidates, pouvaient en fait être réintégrées à l'une ou l'autre des catégories proposées, ou correspondre, comme dans le cas de la jalousie, à une constellation d'émotions, résultant de la mise en péril simultanée de plusieurs engagements différents. Mais nous avons surtout proposé un cadre théorique fonctionnel et cohérent derrière cette classification, qui rend plus facilement envisageable l'intégration de tels processus à des systèmes artificiels complexes, lesquels pourraient éventuellement devenir capables de gérer de façon autonome leur propre accès aux ressources - y compris à celles de deuxième ordre - et se montrer alors susceptibles de s'engager à cette fin dans des comportements d'altruisme et de coopération réciproques.

5.3.5. Quelques applications en psychologie sociale et en psychopathologie

Étant donné l'importance qui est accordée à l'aspect social dans notre modèle du fonctionnement émotionnel, un intérêt additionnel de l'approche proposée est d'établir des liens étroits et fructueux avec plusieurs problèmes de recherche déjà abordés en psychologie sociale, notamment au chapitre des "motivations sociales" (Alain, 1993). Il se trouve en effet que plusieurs des phénomènes étudiés dans ce champ de recherche gravitent autour du concept d'engagement, qui tient là également un rôle de premier plan dans les schémas explicatifs. C'est le cas notamment des phénomènes de persuasion et d'influence, de l'impact des discussions en petits groupes dans la résolution des dilemmes sociaux, et des attitudes antisociales qui semblent caractériser les comportements sociopathiques. Or, comme nous allons tâcher de le montrer, les émotions ne sont jamais loin derrière, et jouent même un rôle déterminant dans l'activation et la mise en oeuvre de ces divers phénomènes.

Les phénomènes d'influence. Au début des années 1980, le chercheur Robert Cialdini (1984) publiait le bilan d'une vingtaine d'années de recherches sur les phénomènes de persuasion et d'influence. Intéressé à comprendre ce qui pouvait bien amener une personne à se laisser persuader de réaliser des actions ou d'engager des ressources, alors qu'elle n'en tirait pratiquement aucun profit immédiat, Cialdini déborda de ses études en laboratoire pour aller observer directement sur le terrain les véritables "marchands d'influence" que sont les vendeurs, les politiciens ou les agents de marketing. Son livre rassemble sous six grands principes les diverses stratégies exploitées par ces professionnels de la persuasion: la tendance à la réciprocité, la persistance dans le respect des engagements pris, l'influence de la majorité, l'affiliation, l'autorité et la rareté des ressources.

Nous avons mentionné le premier de ces principes, lorsque nous avons abordé plus haut la "norme de réciprocité", à la section 4.1.1 (p. 151), comme mécanisme de contrainte sociale visant à assurer la coopération dans les échanges. Il est important cependant de réaliser qu'une véritable *force* motivationnelle semble opérer, dès qu'un individu bénéficie d'un don ou d'un service, le poussant à retourner ultérieurement une faveur au moins équivalente, et souvent même supérieure (Mauss, 1950). Dans le cadre de notre modèle, nous faisons l'hypothèse que ce dynamisme motivationnel repose en majeure partie sur l'activation, parfois conjointe, des émotions de gratitude et de culpabilité, en raison des opportunités d'engagements, et donc de ressources nouvelles qui sont ainsi manifestées. Cialdini suggère que cette contrainte a surtout été acquise à travers l'éducation et la culture, une opinion qui recoupe d'assez près celle de sociologues comme Gouldner (1960) ou Mauss (1950), ou celle de Simon (1990b) à propos de l'inculcation et de l'acquisition des comportements altruistes. Nous croyons pour notre part que l'éducation ne peut être aussi efficace, que parce qu'elle s'appuie sur des mécanismes beaucoup plus fondamentaux déjà mis en place au cours de l'évolution. C'est sans doute la raison pour laquelle la distribution de dons et de faveurs constitue un moyen si efficace pour l'apprivoisement des animaux, probablement d'ailleurs celui qui est le plus spontanément et le plus fréquemment utilisé, jusque par les animaux eux-mêmes, notamment dans les contextes de séduction et de parades sexuelles.

Le second principe d'influence ("commitment and consistency") réfère explicitement au concept d'engagement, mais surtout entendu dans le sens "individualiste" de persistance, tel qu'invoqué dans les domaines de recherches en IAD ou en motivation scolaire. L'idée

en est, qu'une fois qu'un individu s'est fixé un certain but, il éprouve généralement de la difficulté à s'en écarter, comme si une force intérieure le contraignait à réaliser l'objectif adopté. Cependant, comme le révèlent les recherches: "It appears that commitments are most effective in changing a person's self-image and future behavior when they are active, public, and effortful" (Cialdini, 1984, p. 96). Cette citation est particulièrement éclairante, en ce qu'elle fait intervenir l'image de soi comme médiatrice de la motivation à persister, mais aussi le regard public comme puissant déterminant de cette image de soi.

Cialdini rappelle notamment que le fait d'exprimer une intention par écrit engage encore plus fortement les sujets, même lorsque le texte n'est pas formulé par la personne elle-même, et qu'il est simplement recopié. Le chercheur rapporte ainsi l'une des stratégies de "lavage de cerveau" les plus efficaces utilisées par les communistes Chinois durant la guerre de Corée, qui consistait simplement à faire recopier par des militaires américains certaines critiques sur le système politique des États-Unis. Ces prisonniers devaient ensuite participer à des débats sur tel ou tel point de cette politique. Peu à peu, de façon presque imperceptible, ils en venaient à faire leurs ces critiques, et à les défendre avec de plus en plus d'acharnement. *Menées habilement, de telles tactiques ont amené presque tous ces prisonniers à modifier leur allégeance, jusqu'à trahir d'une façon ou d'une autre leurs camarades (par exemple en révélant les projets d'évasion), sans pourtant que leurs geôliers n'aient dû avoir recours à aucuns sévices physiques.*

Or cette vision des engagements rejoint le sens plus social et plus "contractuel" que nous attribuons à ce concept. Rappelons à cet égard que, selon certains théoriciens des actes de langage (Winograd et Flores, 1986; Winograd, 1988), toute forme d'énoncé exprime nécessairement un engagement, et que c'est même la seule façon que la communication serve les interlocuteurs, puisque c'est par là qu'elle leur garantit la réalisation d'actions qui sont mutuellement désirables. À titre d'illustration, imaginez que vous demandiez l'heure dans le métro, et qu'un individu vous communique une information erronée, disons avec un décalage d'une vingtaine de minutes. Il est facile à quiconque de comprendre qu'un tel comportement puisse vous mettre sérieusement en colère. Mais pourquoi donc? C'est que, si vous avez posé la question, vous avez probablement vraiment besoin de le savoir. Votre interlocuteur *devrait donc savoir* à son tour qu'il importe que l'information donnée soit fiable et véridique. Bref, même dans un simple échange, on ne parle pas pour rien dire: il y a des engagements en jeu, et les réactions émotionnelles qui semblent alors si naturelles *font simplement leur travail* de préserver et de réguler ces ressources si précieuses. Ce que

les communistes Chinois et les vendeurs d'autos savent si bien, c'est qu'en nous faisant écrire et signer leurs textes ou leurs formulaires, ils nous associent comme malgré nous, par l'intermédiaire de la mécanique même du langage, à des engagements qui nous contraignent, et qu'ils harnachent du même coup tout l'arsenal de nos propres réactions émotionnelles (notamment dans ce cas la peur et la culpabilité) à la protection et au maintien de ces engagements plus ou moins extorqués.

Le troisième principe proposé par Cialdini, est l'influence de la majorité ("social proof"). Si tout le monde le fait, c'est sans doute que ça en vaut la peine, car sinon, il n'y aurait pas autant d'adeptes. Les résultats de recherche spécifient cependant deux qualifications importantes à ce principe:

> In general, when we are unsure of ourselves, when the situation is unclear or ambiguous, when uncertainty reigns, we are most likely to look to and accept the actions of others as correct. (Cialdini, 1984, p. 129)
> We will use the actions of others to decide on proper behavior for ourselves, especially when we view those others as similar to ourselves. (*Ibid.*, p. 142)

La façon dont ce mode d'influence opère n'est pas sans rappeler le phénomène de "social referencing" (Klinnert, Campos, Sorce, Emde et Svejda, 1983) que nous avons décrit au chapitre 3 (section 3.2.2.4, p. 80-81). Ainsi, lorsqu'un enfant se retrouve dans une situation d'incertitude et d'insécurité, il se tourne vers la personne présente à laquelle il est le plus attaché, et il ajuste ses propres réactions émotionnelles en conformité avec celle de son modèle. Cette réaction lui permet ainsi d'acquérir rapidement, par imitation et "modelage émotionnel", une foule d'informations essentielles à sa propre sécurité. Minsky (1985, p. 176) utilise de son côté le terme "attachment-learning" pour désigner un processus tout-à-fait analogue. Ici encore, on se retrouve en plein paysage émotionnel, où l'insécurité, la détresse, la gratitude et l'attachement déterminent largement les conduites adoptées et rendent compte des influences subies.

Les deux principes suivants, l'affiliation et l'autorité (ou le pouvoir), correspondent exactement aux deux catégories d'engagements à long terme que nous avons décrites dans le chapitre 4 (section 4.4.4, p. 187), et nous ne les détaillerons donc pas ici. Ils expriment le fait que l'on a généralement tendance à subir l'influence de ceux que l'on aime, ou de ceux qui ont du pouvoir. Le dernier principe énoncé, la rareté des ressources, opère en laissant croire à l'acquéreur potentiel que la marchandise offerte est pratiquement épuisée, et qu'en dépit du prix élevé, c'est probablement la dernière chance qu'il a de se la

procurer. Même s'il rencontre le critère de gestion des ressources qui le place effectivement, selon notre définition, dans l'ordre du motivationnel, ce principe ne porte pas spécifiquement sur les engagements. Il est néanmoins intéressant de constater que le mode d'opération *des cinq autres mécanismes* proposés par Cialdini est directement explicable, *et de façon bien plus parcimonieuse,* dans le cadre de notre modèle des émotions fondé sur la gestion des engagements.

La résolution des dilemmes sociaux. Un autre phénomène qui a longtemps intrigué les chercheurs en psychologie sociale, est l'impact de la communication sur la résolution des "dilemmes sociaux". On qualifie ainsi les situations où les individus doivent choisir entre des actions dont ils tireraient avantage au dépens de l'intérêt du groupe, et d'autres actions qui seraient plutôt bénéfiques à l'ensemble de la collectivité (Dawes, 1980). Un exemple pourrait être d'accepter un prix plus élevé pour l'essence, ou d'utiliser plus fréquemment les transports publics, afin de contribuer à la réduction de la pollution, qui est évidemment nocive pour tout le monde. Or depuis une cinquantaine d'années, les chercheurs savent que les discussions de groupe ont pour effet de promouvoir des attitudes collectives et coopératives (Orbell, van de Kragt et Dawes, 1988), sans toutefois comprendre quelle contrainte opère sur les consciences individuelles pour produire un tel résultat. Que se passe-t-il donc dans les esprit durant ces discussions en petits groupes, pour ainsi disposer les individus à coopérer? L'hypothèse la plus solidement confirmée jusqu'ici, est qu'à travers l'échange verbal, "group members made and honored commitments to cooperate" (Kerr et Kaufman-Gilliland, 1994, p. 513). Ce dynamisme semble opérer même lorsque le choix de collaborer n'est pas clairement énoncé, et même si l'individu n'a pas consciemment réalisé pourquoi son attitude a changé, pourvu qu'il ait pris une part active dans la discussion. Ici encore, on réalise à quel point la gestion des engagements constitue une force motivationnelle intense, et à l'instar de Simon (1990b), nous persistons à croire que les dispositifs émotionnels d'affiliation, d'autorité, de colère, de peur et de culpabilité constituent les mécanismes puissants progressivement mis en place à travers l'évolution, et récupérés par l'éducation et la culture, pour inscrire dans les esprits la disposition à coopérer et à respecter les ententes mutuelles, même implicites.

Les comportements sociopathiques. Nous avons rapporté au chapitre 2 l'histoire de Phineas Gage, ce contremaître des chemins de fers américains qui, à la fin du siècle dernier, fut victime d'une explosion au cours de laquelle une barre de métal d'un mètre de long traversa son crâne de part en part, dans la région du cortex préfrontal (Damasio,

1994a). Bien qu'il ne sembla en subir aucune séquelle au plan de ses connaissances, de ses facultés de mémorisation ou de rappel, de ses capacités de raisonnement ou d'apprentissage, sa personnalité en fut néanmoins radicalement transformée. Aux yeux de sa famille, de ses voisins ou de ses amis, il était devenu un tout autre homme: impoli, sans vergogne, ne respectant plus aucune conventions, incapable de prendre des décisions, de tenir ses engagements, ou de planifier pour l'avenir. Sa vie en fut irrémédiablement détruite, il vécut divorces sur divorces, perdit ses amis et ses emplois, et termina misérablement sa vie en clochard.

Or le neurologue Antonio Damasio et son équipe retrouvèrent ce même syndrome antisocial chez des "versions" modernes de Phineas Gage: divers patients ayant eux aussi subi des opérations ou des traumatismes majeurs dans la région du cortex préfrontal (Damasio, Tranel et Damasio, 1990). Ces individus se retrouvent alors généralement diagnostiqués comme souffrant de "désordre sociopathique": difficulté à maintenir leur persistance au travail, planification déficiente, négligence ou irresponsabilité parentale, incapacité à maintenir un attachement durable avec un partenaire sexuel ou amoureux, tendance à la manipulation et à la tricherie, comportement de parasitisme, détachement émotionnel, absence de remords ou d'empathie, faible capacité d'auto-contrôle (Damasio, Tranel et Damasio, 1990, p. 82; Patrick, 1994, p. 320) Plus récemment, une chercheure en psychologie évolutive ("evolutionary psychology"), Linda Mealey, émit l'hypothèse que ces comportements de sociopathie correspondaient essentiellement en un déficit au niveau de l'expérience et de l'expression émotionnelles, lequel entraîne alors une incapacité pour l'individu à s'engager dans des comportements d'entraide et de coopération, et à construire avec les autres des rapports basés sur la confiance réciproque:

> Whether criminal or not, sociopaths typically exhibit what is generally considered to be irresponsible and unreliable behavior; their attributes include egocentrism, an inability to form lasting personal commitments and a marked degree of impulsivity. Underlying a superficial veneer of sociability and charm, sociopaths are characterized by a deficit of the social emotions (love, shame, guilt, empathy, and remorse). (Mealey, 1995, p. 523)

Dans ce domaine de recherche également, on retrouve donc une fois de plus la gestion de la réciprocité et des engagements comme mécanisme de contrôle sur lequel repose l'équilibre personnel de l'individu, ainsi que son adaptation à son environnement social. Et là encore, on retrouve les structures émotionnelles comme les éléments de base et les rouages essentiels qui composent cette couche de contrôle.

5.4. Les contributions pour l'éducation

Finalement, dans le contexte de l'éducation et de la motivation scolaire, le concept d'engagement, tel que nous l'avons défini, offre un champ nouveau d'exploration pour mieux comprendre les bases intrinsèques et interactives sur lesquelles reposent la persistance et l'engagement des élèves. Il entraîne même une reformulation significative des modèles prévalants, notamment en ce qui concerne les indicateurs et les déterminants de la motivation scolaire. Au profit des approches sociocognitives en éducation, il rend désormais possible la formulation de modèles plus formels, exploitant éventuellement les possibilités et les mérites de la simulation informatique. Finalement, une compréhension de la dynamique motivationnelle reposant sur la gestion des engagements rend possible, à l'instar des stratégies métacognitives proposées pour l'amélioration de l'apprentissage et de l'enseignement, la conception et l'enseignement de stratégies "méta-émotionnelles" favorisant une plus grande autonomie chez les élèves en leur redonnant plus de contrôle sur leur propre motivation.

5.4.1. Impact de la redéfinition du concept d'engagement pour les théories actuelles sur la motivation scolaire

Même si l'approche que nous avons adoptée dans le cadre de cette recherche peut avoir semblé, par son caractère théorique, passablement éloignée du contexte scolaire, nous avions néanmoins fait le pari qu'une analyse approfondie de la mécanique motivationnelle pouvait entraîner de retombées inédites sur les recherches actuelles en motivation scolaire. Or notre analyse soulève effectivement la nécessité de reconsidérer certains aspects des modèles actuellement prévalants dans ce domaine, notamment en ce qui concerne la notion d'engagement. En premier lieu, la redéfinition que nous avons proposée de ce concept, ainsi que le rapprochement établi à la section précédente avec tout un champ de recherches portant sur les motivations sociales, nous incite à questionner cette interprétation strictement "individuelle", en terme de persistance dans l'atteinte des buts adoptés. Nous avons pu voir en effet que cette attitude de congruence et de "fidélité" à l'endroit des objectifs que l'on s'est fixés, est largement conditionnée par l'image de soi, ainsi que par le regard extérieur qui est porté sur cette image, à tel point que cet aspect interactif et quasiment "contractuel" du terme d'engagement, se traduit par une puissante incitation à la persévérance dans l'accomplissement d'actes auxquels nous avons ainsi publiquement

associé notre identité. L'effet motivationnel observé résulterait alors moins d'une contrainte interne déterminée par l'atteinte du but lui-même, que du réseau d'engagements interpersonnels et sociaux qui donne un sens à ce but, et où interviennent entre autres divers mécanismes comme l'apprentissage par attachement (Minsky, 1985), l'imitation et la construction identitaire.

Dans cette perspective, il ne nous apparaît pas suffisant de concentrer les interventions pédagogiques sur les croyances des élèves quant à leurs capacités de succès, sur leur conception de l'intelligence comme innée ou acquise, sur leur perception du système scolaire comme centré sur l'apprentissage ou sur la performance, sur les caractéristiques de la tâche comme leur apparaissant plus ou moins réalisable et plus ou moins contrôlable. L'idée n'est évidemment pas ici de rejeter les résultats de recherche confirmant que ces variables ont effectivement un rôle, mais plutôt de faire ressortir que *ces perceptions sont elles-mêmes sous le contrôle d'un déterminisme plus large et plus efficient,* inscrit dans la dynamique de l'acquisition des connaissances *comme activité collective.* Il nous semble en effet surprenant que les modèles prévalants (Dweck, 1986; Pintrich et Garcia, 1994; Schunk, 1991; Viau, 1994), qui s'inscrivent pour la plupart dans l'approche socio-cognitive (Bandura, 1986), n'attachent pas plus d'importance au processus d'apprentissage et de construction de savoirs *en tant qu'activités fondamentalement sociales d'échange et de partage de ressources,* où s'avère crucial l'établissement de la confiance entre les partenaires, et où des mécanismes de négociation, de résolution de conflits, d'organisation et de coopération sont constamment à l'oeuvre.

En second lieu, le concept d'engagement apparaît dans certaines de ces théories, notamment dans celle de Viau (1994) qui effectue la synthèse de plusieurs approches, du côté des *indicateurs* plutôt que du côté des *déterminants* de la dynamique motivationnelle. Bien que cet auteur souscrive au principe du "déterminisme réciproque", selon lequel les effets agissent en retour sur les causes, le fait d'avoir néanmoins regroupé l'engagement du côté des conséquences dans l'articulation du modèle révèle que son rôle comme facteur de la dynamique motivationnelle est nettement perçu comme minimal. La conséquence, au plan de l'intervention pédagogique, est que les enseignantes et les enseignants seront plutôt invités à agir sur les perceptions des élèves, laissant à la dynamique motivationnelle la fonction d'assurer par elle-même la causalité circulaire et l'impact en retour des engagements sur la motivation et les choix des élèves.

Selon notre point de vue cependant, les engagements, en tant que ressources de deuxième ordre, apparaissent absolument décisifs dans cette dynamique. Avec le même succès et la même efficacité qu'ils sont mis au service de stratégies de mobilisation et de persuasion dans tant d'autres domaines de la vie quotidienne et de l'interaction sociale, il nous semble qu'ils devraient également être mis à profit dans le contexte pédagogique de la classe et de l'école. Cela signifie qu'il y a tout intérêt, au profit même des élèves et de la qualité de leurs apprentissages, de déployer systématiquement des stratégies qui les engagent, de façon explicite et interactive, à l'endroit des enseignants et des enseignantes, de la direction de l'école, de leurs parents et de leurs amis, et même à l'égard de la société qui supporte leur croissance. Par ailleurs, notre analyse de la contrainte de réciprocité comme condition d'émergence de la coopération, entraîne qu'il faille engager tout aussi explicitement le corps professoral, et tous les autres acteurs de l'éducation, à l'endroit des élèves. À cette fin, notre modèle plaide finalement en faveur d'une exploitation plus systématique et plus intensive de la structure de contrôle "naturelle" que constituent les émotions à cet égard: oui, la colère, lorsque l'autre transgresse les règles, la culpabilité lorsque l'on fait soi-même défection, la joie pour célébrer les acquisitions, que ce soit en alliances ou en connaissances nouvelles, la tristesse lorsqu'il y a eu perte ou échec, pour signifier la carence qui handicape et le soutien des autres qui est alors requis.

Bien sûr, une telle conclusion peut surprendre. Mais que l'on songe seulement à l'utilité - voire à la nécessité - de ces structures dans le contexte familial, pour accueillir le nouveau-né qui ne parle pas encore, arriver à déchiffrer ses besoins, lui communiquer les premières règles de conduite, l'aider à construire ses propres niches de confiance et de sécurité, et l'introduire dans la communauté linguistique comme instrument privilégié d'adaptation à son milieu social. Que l'on songe également à leur fréquence d'utilisation, par un sourire, un cri, un reproche ou un geste d'excuse, sur la rue comme au travail, pour gérer la majorité de nos rapports quotidiens. Que l'on songe finalement à l'extrême utilité de ces mécanismes lorsqu'il s'agit de franchir les barrières de langue et de culture, et même pour interpréter de façon sécuritaire les comportements d'autres espèces, jusqu'à pouvoir parfois entrer en contact relativement "intime" avec elles. C'est l'omniprésence des émotions dans toutes nos activités que nous avons essayé de mieux comprendre par ce travail, en même temps que l'extraordinaire puissance d'action qu'elles rendent possible sur nous-mêmes et sur notre entourage. Il nous a semblé que de mieux comprendre comment elles s'avéraient si utiles ailleurs dans nos vies, permettrait peut-être de les mettre également au service des objectifs de l'éducation et de l'acquisition des connaissances.

5.4.2. Une base plus formelle pour les approches sociocognitives en éducation

Une autre contribution importante a été de montrer qu'il est possible de rendre compte de comportements apparemment aussi difficilement descriptibles que les réactions émotionnelles, au point d'en formuler des spécifications précises pour leur éventuelle implantation dans des agents artificiels. Nous avons également tenté de montrer que ces mécanismes ne devaient pas être inscrits dans de tels systèmes uniquement pour des questions de convivialité, c'est-à-dire simplement pour "faire vrai" et donner à la machine un apparence plus "sympathique", mais qu'ils seraient même *requis* pour la conceptualisation d'agents informatiques véritablement autonomes, capables d'établir des rapports de coopération réciproque avec d'autres agents de même facture, et de gérer efficacement cette réciprocité. Or une telle base de formalisation est également bienvenue en sciences humaines et en éducation, où le nombre de variables en jeu est si élevé et leurs interactions si complexes, qu'il semble souvent illusoire d'obtenir des modèles formels permettant des simulations réalistes et significatives. C'est notamment le cas avec les approches sociocognitives, qui, en ne se restreignant pas au seul individu comme siège de l'apprentissage, et en y incluant toute la richesse des interactions sociales où se trouve imbriqué l'enfant, semblent s'être restreintes irrémédiablement à des théories descriptives des phénomènes d'éducation. Le modèle que nous avons proposé ici illustre au contraire qu'il est également envisageable de modéliser ces interactions complexes avec les outils de la simulation informatique et d'atteindre par là un niveau explicatif. Bien que nous n'ayons pas encore réalisé nous-même une telle implantation, l'accueil et le succès que notre modèle a connus dans les différentes communautés d'intelligence artificielle où il a été présenté, nous permettent d'espérer que l'approche adoptée et les spécifications proposées ouvrent dans cette direction des pistes prometteuses.

5.4.3. Des stratégies métacognitives aux stratégies méta-émotionnelles

D'une façon quasi unanime, les chercheurs en science cognitive et en éducation estiment que l'utilisation de stratégies métacognitives favorise considérablement l'apprentissage, et constitue même l'une des caractéristiques essentielles, aussi bien des élèves performants que des experts (Tardif, 1992). Or cette capacité à guider et contrôler soi-même ses processus d'acquisition de connaissances et de résolution de problèmes repose dans une large mesure sur une compréhension approfondie du fonctionnement de la mémoire et des

autres processus cognitifs. De la même façon, il y a tout lieu de croire qu'une connaissance plus articulée des mécanismes qui sous-tendent la dynamique émotionnelle et les attitudes coopératives pourrait permettre la conception et l'enseignement de stratégies "méta-émotionnelles", dont l'application pourrait alors résulter dans l'atteinte d'une plus grande autonomie chez les élèves, en leur redonnant plus de contrôle sur leur propre motivation. Deux chercheurs en psychologie des émotions, Salovey et Mayer (1989; voir aussi Goleman, 1995), ont d'ailleurs proposé d'appeler "intelligence émotionnelle", cet ensemble d'habiletés qu'ils définissent comme suit:

> a set of skills hypothesized to contribute to the accurate appraisal and expression of emotion in oneself and others, the effective regulation of emotion in self and others, and the use of feelings to motivate, plan, and achieve in one's life. (Salovey et Mayer, 1989, p. 185)

La première composante identifiée par ces auteurs implique la capacité de reconnaître les expériences émotionnelles ressenties, aussi bien chez soi-même que chez les autres, ainsi que de les exprimer clairement, tant au plan verbal que non verbal. Ce type particulier d'intelligence inclut en outre la capacité d'agir sur les émotions, celles des autres comme les siennes propres, afin d'en assurer la régulation et le contrôle. Troisièmement, les émotions peuvent être utilisées de multiples façons pour favoriser et supporter le processus de résolution de problèmes. Toujours selon ces auteurs, les sautes d'humeurs pourraient mêmes être mises à profit pour la génération d'alternatives, induisant plus de flexibilité dans la planification. Elles aident également à rediriger l'attention vers des stratégies qui n'ont pas encore été considérées. Par ailleurs, comme l'ont révélé plusieurs résultats de recherche (Isen, 1993), les émotions positives favorisent généralement la réorganisation des informations en mémoire, permettant une pensée plus créative. En présence de tâches difficiles, il est finalement possible d'exploiter le pouvoir motivationnel des émotions afin de soutenir l'intérêt, maintenir la persistance et contrer l'anxiété.

Les stratégies "méta-émotionnelles" ainsi proposées, particulièrement dans le cas de la troisième composante, reposent cependant sur une conception plus conventionnelle des émotions, qui ne sont pas surtout envisagées sous l'angle interactif et social qui est privilégié ici. Ainsi que nous l'avons mentionné plus haut, une contribution significative et originale du modèle que nous avons élaboré est d'identifier les engagements comme "la" variable critique derrière le déclenchement et l'expression des émotions, et d'offrir ainsi, pour les praticiens et les praticiennes de l'éducation, un levier explicite d'interven-

tions sur les motivations des élèves, en mettant à profit la puissance considérable des mécanismes émotionnels inscrits en nous à travers des millénaires d'évolution.

5.5. Les avantages et les limites de l'approche

Au terme de l'entreprise qui a mené à la rédaction de cette thèse, il nous semble finalement utile et pertinent de formuler quelques commentaires sur l'approche utilisée. Deux avantages retiennent particulièrement l'attention. En premier lieu, bien que nous n'ayons pas effectivement réalisé d'implantation informatique du modèle, l'approche d'ingénierie qui nous a habité tout au long de cette recherche a largement contribué au résultat obtenu. Elle nous a d'abord imposé la contrainte de spécifier les comportements émotionnels comme résultant de mécanismes concrets, d'une façon suffisamment précise qu'il soit envisageable de les reproduire artificiellement. Mais elle nous a également incité à rechercher les avantages et les gains, en termes d'efficacité et de productivité au niveau de la gestion des ressources, qu'il pouvait y avoir, pour une application informatique, à disposer de tels mécanismes. C'est en particulier ce qui nous a poussé à examiner, du côté de la théorie de l'évolution, les contraintes d'adaptation qui avaient pu contribuer à favoriser la sélection de tels dispositifs, afin de réguler les tendances égoïstes risquant d'empêcher les membres des espèces sociales de tirer profit des bénéfices de la coopération.

Une autre caractéristique importante de notre approche, est de nous être imposé la contrainte de tenir simultanément compte d'une multiplicité de champs de recherche, apportant chacun, par des angles différents, leur contribution à la compréhension des phénomènes émotionnels. Nos arguments, nous les avons en effet puisé dans des domaines aussi diversifiés que la neurophysiologie, la théorie de l'évolution, l'éthologie, la psychologie des émotions, la psycholinguistique, les sciences de l'éducation, la sociologie et l'anthropologie, la théorie des jeux, la vie artificielle et l'intelligence artificielle distribuée. Malgré les risques de dispersion que comporte une telle diversité, nos contributions ont néanmoins été accueillies favorablement par la communauté scientifique internationale. Au congrès de Baltimore (voir Annexe B), notre présentation avait été insérée dans la section portant sur la théorie des actes de langage. Dans la publication des actes du congrès européen en IAD (chapitre 4, section 4.3), notre communication occupait la section sur les fondements épistémologiques des systèmes multi-agents. Quant congrès de Cape Cod où nous avons présenté notre quatrième article (chapitre 4, section 4.4):

> The objective or the biennal SAB conferences is to bring together researchers from a wide range of backgrounds including ethology, theoretical biology, psychology, artificial life, machine learning and robotics [...] The present proceedings are a comprehensive and up-to-date resource for the latest progess in this exciting field. (Maes et al., 1996, Preface, p. xi - reproduite à l'Annexe C)

La principale limite reste toutefois l'absence d'implantation, même si nous avons tenté, au chapitre 4, de formuler un certain nombre de spécifications, qui ont déjà reçu un accueil favorable dans différentes communautés de chercheurs en IA. Par delà le côté simplement spectaculaire d'une réalisation informatique, il y a également l'effet régulateur de ce processus qui révèle parfois l'incomplétude d'une spécification, son inefficacité d'opération ou l'apparition imprévue d'effets secondaires indésirables. Une implantation pourrait également générer plusieurs possibilités de simulation des comportements sociaux, qui seraient appréciables ici, notamment pour examiner plus en détails les rouages de la coopération, de la résolution de conflits, ou de la motivation scolaire. Nous restons donc persuadés qu'il faut envisager là un prolongement important de notre travail.

Un autre limite concerne la vérification empirique de notre modèle. Bien que le chapitre 3 ait rassemblé une multitude de données empiriques à l'appui de notre thèse, il reste indispensable que des confrontations plus systématiques soient mises en oeuvre. En postulant qu'il y a nécessairement un engagement en jeu derrière l'activation de n'importe quelle réaction émotionnelle, notre théorie offre à la vérification empirique des propositions falsifiables qu'il importe donc de soumettre à l'épreuve des faits. Par ailleurs, si nous avons raison de croire que les structures émotionnelles constituent des structures de contrôle des conduites aussi puissantes, il est également important d'en tirer des applications concrètes pour l'éducation, notamment au niveau de la gestion de la classe et des stratégies pour favoriser la motivation des élèves. Au terme de ce périple, nous voici donc relancé sur de multiples pistes d'action. Mais n'est-ce pas là ce qui rend la recherche aussi vivante et constitue son mérite premier?

Bibliographie

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist, 36(7),* 715-729.

Abelson, H., and diSessa, A. A. (1980). *Turtle geometry: The computer as a medium for exploring mathematics.* Cambridge: MIT Press.

Agha, G. (1986). *Actors.* Cambridge, MA : MIT Press.

Ainsworth, M. D. S. (1989). Attachments beyond infancy. *American Psychologist, 44,* 709-716.

Alain, M. (1993). Les théories sur les motivations sociales. In R. J. Vallerand and E. E. Thill (Eds.), *Introduction à la psychologie de la motivation* (pp. 465-507). Laval, Québec: Éditions Études Vivantes.

Anders Ericson, K. and Smith, J. (Eds.) (1991). *Toward a general theory of expertise. Propects and limits.* Cambridge: Cambridge University Press.

Anderson, J. R. (1980). *Cognitive psychology and its implications.* New York: W. H. Freeman.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge: Harvard University Press.

Anderson, J. R. and Bower, G. H. (1973). *Human associative memory.* Washington: Winston and Sons.

Anderson , J. R., Boyle, C. F., Farrell. R. and Reiser, B. J. (1987). Cognitive principles in the design of computer tutors. In P. Morris, (Ed.), *Modelling Cognition.* New York: John Wiley and Sons.

Angier, R. P. (1927). The conflict theory of emotion. *American Journal of Psychology, 39,* 390-401.

Archer, J. (1976). The organization of aggression and fear in vertebrates. In P. P. G. Bateson and P. H. Klopfer (Eds.), *Perspectives in ethology* (Vol. 2, pp. 231-298). New York: Plenum Press.

Arnold, M. B. (1960). *Emotion and personality* (2 volumes). New York: Columbia University Press.

Aubé, M. and Senteni, A. (1995). A foundation for commitments as resource management in multi-agents systems. In T. Finin and J. Mayfield, (Eds.), *Proceedings of the CIKM Workshop on Intelligent Information Agents.* Baltimore, Maryland, December 1-2 1995.

Aubé, M. and Senteni, A. (1996a). Emotions as commitments operators: A foundation for control structure in multi-agents systems. In W. Van de Velde and J. W. Perram, (Eds.), *Agents breaking away,* Proceedings of the 7th European Workshop on MAAMAW, Lecture Notes on Artificial Intelligence, No. 1038, (pp. 13-25). Berlin: Springer.

Aubé, M. and Senteni, A. (1996b). What are emotions for? Commitments management and regulation within animals/animats encounters. In P. Maes, M. Mataric, J.-A. Meyer, J. Pollack, and S. W. Wilson, (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* (pp. 264-271). Cambridge, MA: The MIT Press/Bradford Books.

Austin, J. L. (1962). *How to do things with words.* Cambridge: Harvard University Press.

Averill, J. R. (1979). Anger. In H. E. Howe and R. A. Dienstbier (Eds.), *Nebraska Symposium on Motivation 1978* (Vol. 26, pp. 1-80). Lincoln: University of Nebraska Press.

Averill, J. R. (1980). A constructivist view of emotion. In R. Plutchik and H. Kellerman (Eds.), *Emotion: Theory, research, and experience: Vol. 1. Theories of emotions* (pp. 305-339). New York: Academic Press.

Averill, J. R. (1983). Studies on anger and agression. Implications for theories of emotion. *American Psychologist, 38,* 1145-1160.

Axelrod, R. (1984). *The evolution of cooperation.* New York: Basic Books.

Axelrod, R. and Dion, D. (1988). The further evolution of cooperation. *Science, 242,* 1385-1390.

Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(27), 1390-1396.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

Barrett, K. C. (1995). A functionalist approach to shame and guilt. In J. P. Tangney and K. W. Fischer (Eds.), *Self-conscious emotions. The psychology of shame, guilt, embarrassment, and pride* (pp. 25-63). New York: The Guilford Press.

Bassili, J. N. (1976). Temporal and spatial contingencies in the perception of social events. *Journal of Personality and Social Psychology, 33*(6), 680-685.

Bates, J., Loyall, A. B. and Reilly, W. S. (1992). An architecture for action, emotion, and social behavior. In C. Castelfranchi and E. Werner, (Eds.), *Artificial social systems,* Proceedings of the 4th European Workshop on MAAMAW, Lecture Notes on Artificial Intelligence, No. 830, (pp. 55-68). Berlin: Springer-Verlag.

Bateson, G. (1955). A theory of play and phantasy. *Psychiatric Research Reports, 2,* 39-51.

Baumeister, R. F. and Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*(3), 497-529.

Baumeister, R. F., Stillwell, A. M. and Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin, 115*(2), 243-267.

Beaudoin, L. P. (1994). *Goal processing in autonomous agents.* Ph. D. Thesis, School of Computer Science, University of Birmingham, Edgbaston, Birmingham, U. K.

Beaudoin, L. P. and Sloman, A. (1993). A study of motive processing and attention. In A. Sloman, D. Hogg, G. Humphreys, A. Ramsay and D. Partridge (Eds.), *Prospects for artificial intelligence - Proceedings of AISB'93* (pp. 229-234). Oxford: IOS Press.

Becker, H. S. (1960). Notes on the concept of commitment. *American Journal of Sociology, 66,* 32-40.

Berkowitz, L. (1993). Pain and agression: Some findings and implications. *Motivation and Emotion, 17,* 277-293.

Bobrow, D. G. and Hayes, P. J. (1985). Artificial intelligence - Where are we?. *Artificial Intelligence, 25,* 375-415.

Bobrow, D. G. and Winograd, T. (1977). An overview of KRL, a knowledge representation language. *Cognitive Science, 1*(1), 3-46.

Bobrow, D. G. and Winograd, T. (1979). KRL: Another perspective. *Cognitive Science, 3*(1), 29-42.

Bonanno, G. A. and Stillings, N. A. (1986). Preference, familiarity, and recognition after repeated brief exposures to random geometric shapes. *American Journal of Psychology, 99,* 403-415.

Bond, A. H. (1990). A computational model for organizations of cooperating intelligent agents. *SIGOIS Bulletin, 11,* 21-30.

Bond, A. H. and Gasser, L. (Eds.) (1988). *Readings in distributed artificial intelligence.* San Mateo, Ca.: Morgan Kaufmann Publishers.

Boring, E. G. (1942). *Sensation and perception in the history of experimental psychology.* New York: Appleton-Century-Crofts.

Boucher, J. D. (1983). Antecedents to emotions across cultures. In S. H. Irvine and J. W. Berry (Eds.), *Human assessment and cultural factors* (pp. 407-420). New York: Plenum.

Bouron, T. (1992). *Structures de communication et d'organisation pour la coopération dans un univers multi-agents: Une contribution à la définition d'un modèle de programmation agent basé sur les concepts d'engagement et d'acte de langage.* Thèse de Doctorat, Université de Paris VIème, Paris.

Boyd, R. and Lorberbaum, J. P. (1987). No pure strategy is evolutionary stable in the repeated Prisoner's Dilemma game. *Nature, 327,* 58-59.

Bower, G. H., Black, J. B. and Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11,* 177-220.

Bowlby, J. (1969). *Attachment and loss. Vol. 1: Attachment.* London: Penguin Books.

Bowlby, J. (1973). *Attachment and loss. Vol. 2: Separation, anxiety, and anger.* London: Penguin Books.

Bowlby, J. (1980). *Attachment and loss. Vol. 3: Loss, sadness, and depression.* London: Penguin Books.

Bratman, M. E. (1990). What is intention? In P. R. Cohen, J. Morgan and M. E. Pollack (Eds.), *Intentions in communications* (pp. 15-31). Cambridge, MA: MIT Press.

Bretherton, I. and Beeghley, M. (1982). Talking about internal states: The acquisition of an explicit theory of mind. *Developmental Psychology, 18,* 906-921.

Briot, J.-P. (1994). Modélisation et classification de langages de programmation concurrente à objets : L'expérience Actalk, *Actes du colloque "Langages et Modèles à Objets"*. Grenoble, France.

Broadbent, D. E. and Broadbent, M. (1988). Anxiety and attentional bias: State and trait. *Cognition and Emotion, 2,* 165-183.

Brooke, R. (1985). What is guilt? *Journal of Phenomenological Psychology, 16*(2), 31-46.

Brooks, R. A. (1991a). Elephants don't play chess. In P. Maes (Ed.), *Designing autonomous agents: Theory and practice from biology to engineering and back* (pp. 3-15). Cambridge: MIT Press.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence, 47,* 139-159.

Brooks, R. A. (1995). Intelligence without reason. In L. Steels and R. Brooks, (Eds.), *The artificial life route to artificial intelligence,* (pp. 25-81). Hillsdale, NJ: Erlbaum.

Brophy, J. (1983). Conceptualizing student motivation. *Educational Psychologist, 18,* 200-215.

Brown, J. S. , Collins, A. and Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18,* 32-42.

Brown, A., Bransford, J., Ferrara, R., and Campione, J. (1983). Learning, remembering and understanding. In P. H. Mussen, (Ed.), *Handbook of Child Psychology (4th edition). Vol III: Cognitive Development,* (pp. 77-166). J. H. Flavel et E. M. Markman (Volume ed.). New York: John Wiley and Sons.

Bryson, J. B. (1991). Modes of response to jealousy-evoking situations. In P. Salovey (Ed.). *The psychology of jealousy and envy* (pp. 178-207). New York: The Guilford Press.

Buck, R. (1980). Nonverbal behavior and the theory of emotion: The facial feedback hypothesis. *Journal of Personality and Social Psychology, 38,* 811-824.

Buck, R. (1985). Prime theory: An integrated view of motivation and emotion. *Psychological Review, 92,* 389-413.

Butterfield, E. C. and Nelson, G. D. (1989). Theory and practice of teaching for transfer. *Educational Technology Research and Developpement, 37*(3), 5-38.

Cacioppo, J. T., Klein, D. J., Berntson, G. G. and Hatfield, E. (1993). The psychophysiology of emotion. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotionss*(pp. 119-142). New York: The Guilford Press.

Campbell, J. A. and d'Inverno, M. (1990). Knowledge interchange protocols. In Y. Demazeau and J.-P. Müller (Eds.), *Decentralized A.I.* (pp. 63-80). North-Holland: Elsevier Science Publishers.

Camras, L. A. (1992). Expressive development and basic emotions. *Cognition and Emotion, 6,* 269-283.

Carver, C. S. and Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological Review, 97*(1), 19-35.

Carver, S. M. (1987). Transfer of Logo debugging skill: analysis, instruction and assesment. *Computer System Group Bulletin, 14* (1), 4-6.

Castelfranchi, C. (1990). Social power. A point missed in multi-agent, DAI and HCI. In Y. Demazeau, and J.-P. Müller, (Eds.), *Decentralized A.I.,* (pp. 49-62), North-Holland: Elsevier Science Publishers.

Castelfranchi, C. (1994). Guarantees for autonomy in cognitive agent architecture. In M. J. Wooldridge and N. R. Jennings, (Eds.), *Intelligent Agents,* Proceedings of the ECAI-94 Workshop on Agent Theories, Architectures, and Languages, Lecture Notes on Artificial Intelligence, No. 890, (pp. 56-70), Berlin: Springer-Verlag.

Charlesworth, W. R. and Kreutzer, M. A. (1973). Facial expressions of infants and children. In P. Ekman (Ed.), *Darwin and facial expression: A century of research in review* (pp. 91-167). New York: Academic Press.

Charniak, E. (1972). *Toward a model of children's story comprehension.* Ph. D. Thesis, MIT.

Cheney, D., Seyfarth, R. and Smuts, B. (1986). Social relationships and social cognition in nonhuman primates. *Science, 234,* 1361-1366.

Chevalier-Skolnikoff, S. (1973). Facial expression of emotion in nonhuman primates. In P. Ekman (Ed.), *Darwin and facial expression: A century of research in review* (pp. 11-89). New York: Academic Press.

Chi, M. T. H., Glaser, R. and Farr, M. J. (Eds.) (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Chomsky, N. (1963). Formal properties of grammar. In R. D. Luce, R. R. Bush and E. Galanter, (Eds.), *Handbook of mathematical psychology, vol. II.* New York: John Wiley and Sons.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge: MIT Press.

Cialdini, R. B. (1984). *Influence.* New York: William Morrow and Company.

Clark, M. S. and Fiske, S. T. (Eds.) (1982). *Affect and cognition..* Hillsdale, N.J.: Erlbaum.

Clements, D.H. (1987). Longitudinal study of the effects of Logo programming on cognitive abilities and achievement. *Journal of Educational Computing Research, 3*(1), 73-94.

Clore, G. L. and Ortony, A. (1991). What more is there to emotion concepts than prototypes? *Journal of Personality and Social Psychology, 60,* 48-50.

Clore, G. L., Ortony, A. and Foss, M. A. (1987). The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology, 53,* 751-766.

Cofer, C. N. and Appley, M. H. (1964). *Motivation: Theory and research.* New York: John Wiley and Sons.

Cohen, P. R. and Levesque, H. J. (1990a). Intention is choice with commitment. *Artificial Intelligence, 42,* 213-261.

Cohen, P. R. and Levesque, H. J. (1990b). Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan and M. E. Pollack (Eds.), *Intentions in Communications* (pp. 221-255). Cambridge : MIT Press.

Cohen, P. R. and Perrault, C. R. (1979). Elements of a plan-based theory of speech acts. *Cognitive Science, 3,* 177-212.

Conway, M. A. (1990). Conceptual representation of emotions: The role of autobiographical memories. In K. J. Gilhooly, M. T. G. Keane, R. H. Logie and G. Erdos (Eds.), *Lines of thinking* (Vol. 2, pp. 133-143). New York: John Wiley and Sons.

Conway, M. A. and Bekerian, D. A. (1987). Situational knowledge and emotions. *Cognition and Emotion, 1,* 145-191.

Cormier, S. M. and Hagman, J. D. (Eds.) (1987). *Transfer of learning. Contemporary research and applications.* San Diego: Academic Press Inc.

Craik, F. I. M. and Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

d'Inverno, M. and Luck, M. (1996). Formalizing the contract net as a goal-directed system. In W. Van de Velde and J. W. Perram, (Eds.), *Agents breaking away,* Proceedings of the 7th European Workshop on MAAMAW, Lecture Notes on Artificial Intelligence, No. 1038, (pp. 72-85). Berlin: Springer.

Daly, E. M., Lancee, W. J. and Polivy, J. (1983) A conical model for the taxonomy of emotional experience. *Journal of Personality and Social Psychology, 45*(2), 443-457.

Damasio, A. R. (1994a). *Descarte's error: Emotion, reason and the human brain.* New York: Avon Books.

Damasio, A. R. (1994b). Descarte's error and the future of human life. *Scientific American, October 1994, 271*(4), 144.

Damasio, A. R., Tranel, D. and Damasio, H. (1990). Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli. *Behavioural Brain Research, 41,* 81-94.

Darwin, C. (1965). *The expression of the emotions in man and animals.* Chicago: The University of Chicago Press. (Original work published 1872)

Dautenhahn, K. (1995). Getting to know each other - Artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems, 16,* 333-356.

Davidson, R. J. (1984). Hemispheric assymetry and emotion. In K. R. Scherer, and P. Ekman (Eds.), *Approaches to emotion,* (pp. 39-37). Hillsdale, NJ: Erlbaum.

Davidson, R. J. (1992). Prolegomenon to the structure of emotion: Gleanings from neuropsychology. *Cognition and Emotion, 6,* 245-268.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31,* 169-193.

Dawkins, R. (1976). *The selfish gene.* Oxford: Oxford University Press.

Dawkins, R. (1986). *The blind watchmaker.* London: Longman.

De Wall, F. (1996). *Good natured. The origins of right and wrong in humans and other animals.* Cambridge, MA: Harvard University Press.

Deci, E. L. and Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York: Plenum.

Demazeau, Y. and Müller, J.-P. (Eds.) (1990). *Decentralized A.I.* Amsterdam: North-Holland Elsevier Science.

Demazeau, Y. and Müller, J.-P. (Eds.) (1991). *Decentralized A.I. 2.* Amsterdam: North-Holland Elsevier Science.

Detterman, D. K. and Sternberg, R. J. (Eds.) (1993). *Transfer on trial: Intelligence, cognition, and instruction.* Norwood, NJ: Ablex Publishing Corporation.

Dimberg, U. (1982). Facial reactions to facial expressions. *Psychophysiology, 19*(6), 643-647.

Dongha, P. (1994). Toward a formal model of commitment for resource bounded agents. In M. J. Wooldridge and N. R. Jennings, (Eds.), *Intelligent Agents,* Proceedings of the ECAI-94 Workshop on Agent Theories, Architectures, and Languages, Lecture Notes on Artificial Intelligence, No. 890, (pp. 86-101), Berlin: Springer-Verlag.

Dony, C. (1990). Exception handling and object-oriented programming : Towards a synthesis. In N. M. (Ed.), *OOPSLA ECOOP'90 Proceedings, Ottawa, October 21-25 1990, Sigplan Notices, 25* (10), 322-330.

Duffy, E. (1948). Leepers's "Motivational theory of emotion". *Psychological Review,* 55, 324-328.

Dunbar, R. I. M. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences, 16,* 681-735.

Durkheim, E. (1947). *Les règles de la méthode sociologique.* Paris: Presses Universitaires de France.

Dutton, D. and Painter, S. L. (1981). Traumatic bonding: The development of emotional attachments in battered women and other relationships of intermittent abuse. *Victimology, 6,* 139-155.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41,* 1040-1048

Dyer, M. G. (1987). Emotions and their computations: Three computer models. *Cognition and Emotion, 1,* 323-347.

Eibl-Eibesfeldt, I. (1973). The expressive behavior of the deaf-and-blind-born. In M. Von Cranach and I. Vine (Eds.), *Social communication and movement. Studies of interaction and expression in man and chimpanzee* (pp. 163-194). New York: Academic Press.

Eibl-Eibesfeldt, I. (1975). *Ethology: The biology of behavior.* (Second edition). New York: Holt, Rinehart and Winston.

Ekman, P. (1973). Cross-cultural studies of facial expressions. In P. Ekman (Ed.), *Darwin and facial expression: A century of research in review* (pp. 169-222). New York: Academic Press.

Ekman, P. (1977). Biological and cultural contributions to body and facial movement. In J. Blacking (Ed.), *Anthropology of the body* (pp. 34-84). London: Academic Press.

Ekman, P. (Ed.) (1982). *Emotion in the human face.* Cambridge, England: Cambridge University Press.

Ekman, P. (1984). Expression and the nature of emotion. In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 319-343). Hillsdale, N.J.: Erlbaum.

Ekman, P. (1992a). Are there basic emotions? *Psychological Review, 99,* 550-553.

Ekman, P. (1992b). An argument for basic emotions. *Cognition and Emotion, 6,* 169-200.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48,* 384-392.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin, 115,* 268-287.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17,* 124-129.

Ekman, P. and Friesen, W. V. (1986). A new pan-cultural facial expression of emotion. *Motivation and Emotion, 10,* 159-168.

Ekman, P. and Heider, K. G. (1988). The universality of a contempt expression: A replication. *Motivation and Emotion, 12,* 303-308.

Ekman, P. and Oster, H. (1982). Review of Research, 1970-1980. In P. Ekman (Ed.), *Emotion in the Human Face* (pp. 147-173). Cambridge, England: Cambridge University Press.

Ekman, P., Friesen, W. V. and Ellsworth, P. (1982a). What emotion categories or dimensions can observers judge from facial behavior? In P. Ekman (Ed.), *Emotion in the Human Face* (pp. 39-55). Cambridge, England: Cambridge University Press.

Ekman, P., Friesen, W. V. and Ellsworth, P. (1982b). What are the similarities and differences in facial behavior across cultures? In P. Ekman (Ed.), *Emotion in the Human Face* (pp. 128-143). Cambridge, England: Cambridge University Press.

Ekman, P., Friesen, W. V. and Simons, R. C. (1985). Is the startle reaction an emotion? *Journal of Personality and Social Psychology, 49,* 1416-1426.

Ekman, P., O'Sullivan, M. and Matsumoto, D. (1991). Confusions about context in the judgement of facial expression: A reply to "Contempt and the relativity thesis". *Motivation and Emotion, 15,* 169-176.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M. and Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology, 53,* 712-717.

Elliott, C. (1992). *The affective reasoner: A process model of emotions in a multi-agent system.* Ph. D. Thesis, Northwestern University, Evanston.

Ellsworth, P. C. (1991). Some implications of cognitive appraisal theories of emotion. In K. T. Strongman (Ed.), *International Review of Studies on Emotion,* (Vol. 1, pp. 143-161). New York: John Wiley and Sons.

Ellsworth, P. C. and Smith, C. A. (1988a). From appraisal to emotion: Differences among unpleasant feelings. *Motivation and Emotion, 12*(3), 271-302.

Ellsworth, P. C. and Smith, C. A. (1988b). Shades of joy: Patterns of appraisal differentiating pleasant emotions. *Cognition and Emotion, 2*(4), 301-331.

Ellsworth, P. C. and Tourangeau, R. (1981). On our failure to disconfirm what nobody ever said. *Journal of Personality and Social Psychology, 40,* 363-369.

Emde, R. N. (1984). Levels of meaning for infant emotions: A biosocial view. In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 77-107). Hillsdale, N.J.: Erlbaum.

Epstein, S. (1984). Controversial issues in emotion theory. In P. Shaver (Ed.), *Review of Personality and Social Psychology* (Vol. 5, pp. 64-88). Beverly Hills, Ca.: Sage.

Etcoff, N. L. and Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition, 44,* 227-240.

Evans, A. and Tomasello, M. (1986). Evidence for social referencing in young chimpanzees (Pan Troglodytes). *Folia Primatologica, 47,* 49-54.

Fehr, B. and Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General, 113,* 464-486.

Fehr, B., Russell, J. A. and Ward, L. M. (1982). Prototypicality of emotions: A reaction time study. *Bulletin of the Psychonomic Society, 20,* 253-254.

Ferber, J. (1996). *Cooperation strategies in collective intelligence.* Invited talk presented at the 7th European Workshop on MAAMAW, Eindhoven, The Netherlands, January 23 1996.

Feurzeig, W., Papert, S., Bloom, M. Grant, R. and Solomon, C. (1969). *Programming languages as a conceptual framework for teaching mathematics. Report No. 1889.* Cambridge: Bolt, Beranek and Newman.

Field, T. (1985). Attachment as psychobiological attunement: Being on the same wavelength. In M. Reite and T. Field (Eds.), *The psychobiology of attachment and separation* (pp. 415-454). New York: Academic Press.

Fikes, R. E. (1982). A commitment-based framework for describing informal cooperative work. *Cognitive Science, 6,* 331-347.

Fischer, K. W. and Tangney, J. P. (1995). Self-conscious emotions and the affect revolution: Framework and overview. In J. P. Tangney and K. W. Fischer (Eds.), *Self-conscious emotions. The psychology of shame, guilt, embarrassment, and pride* (pp. 3-22). New York: The Guilford Press.

Fischer, K. W., Shaver, P.R. and Carnochan, P. (1990). How emotions develop and how they organise development. *Cognition and Emotion, 4,* 81-127.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions.* New York: W. W. Norton and Company.

Frijda, N. H. (1986). *The emotions.* Cambridge, England: Cambridge University Press.

Frijda, N. H. (1987). Comment on Oatley and Johnson-Laird's "Towards a cognitive theory of emotions". *Cognition and Emotion, 1*(1), 51-38587.

Frijda, N. H. (1988). The laws of emotions. *American Psychologist, 43,* 343-358.

Frijda, N. H. (1989). The functions of emotional expression. In J. P. Forgas and J. M. Innes (Eds.), *Recent advances in social psychology: An international perspective* (pp. 205-217). Amsterdam: North-Holland.

Frijda, N. H. (1993a). The place of appraisal in emotion. *Cognition and Emotion, 7*(3/4), 357-387.

Frijda, N. H. (1993b). Moods, emotion episodes, and emotions. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 381-403). New York: The Guilford Press.

Frijda, N. H. and Swagerman, J. (1987). Can computers feel? Theory and design of an emotional system. *Cognition and Emotion, 1,* 235-257.

Frijda, N. H., Kuipers, P. and ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology, 57*(2), 212-228.

Fromme, D. K. and O'Brien, C. S. (1982). A dimensional approach to the circular ordering of the emotions. *Motivation and Emotion, 6,* 337-363.

Gagné, E., Yekovitch, C. W. and Yekovitch, F. R. (1993). *The cognitive psychology of school learning.* New York: Harper Collins.

Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution.* New York: Basic Books.

Gasser, L. (1991). Social conceptions of knowledge and action: DAI foundations and open systems semantics. *Artificial Intelligence, 47,* 107-138.

Gasser, L. and Huhns, M. N. (Eds.) (1989). *Distributed artificial intelligence 2.* London: Pitman/Morgan Kaufmann Publishers.

Gentner, D. and Stevens, A. L. (Eds.) (1983). *Mentals models.* Hillsdale, NJ: Erlbaum.

Gerson, E. H. (1976). On "quality of life". *American Sociological Review, 41,* 793-806.

Gibson, J. J. (1977). The theory of affordances. In R. Shaw and J. Bransford (Eds.), *Perceiving, acting and knowing. Toward an ecological psychology* (pp. 67-82). Hillsdale, N.J.: Erlbaum.

Gilovich, T. and Husted Medvec, V. (1995). The experience of regret: What, when, and why. *Psychological Review, 102*(2), 379-395.

Glance, N. S. and Huberman, B. A. (1994). The dynamics of social dilemmas. *Scientific American, March 1994, 270*(3), 76-81.

Goldberg, A. and Kay, A. (1977). *Methods for teaching the programming language Smalltalk. Technical report SSL-77-2.* Xerox Palo Alto Research Center.

Goldberg, A. and Robson, D. (1983). *Smalltalk-80: The language and its implementation.* Reading: Addison-Wesley.

Goleman, D. (1995). *Emotional intelligence.* New York: Bantam Books.

Gould, J. L. and Marler, P. (1987). Learning by instinct. *Scientific American, January 1987, 256*(1), 74-85.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review, 25*(2), 161-178.

Graham, D. L. R., Rawlings, E. and Rimini, N. (1988). Survivors of terror: Battered women, hostages, and the Stockholm syndrome. In K. Yllö and M. Bograd (Eds.), *Feminist perspectives on wife abuse* (pp. 217-233). London: Sage Publications.

Gray, J. A. (1982). Précis of Gray's "The neuropsychology of anxiety: An enquiry into the function of the septo-hippocampal system". *Behavioral and Brain Sciences, 5,* 469-525.

Gray, J. A. (1990). Brain systems that mediate both emotion and cognition. *Cognition and Emotion, 4,* 269-288.

Greeno, J. G., Smith, D. R. and Moore, J. L. (1993). Transfer of situated learning. In D. K. Detterman and R. J. Sternberg, (Eds.), *Transfer on trial: Intelligence, cognition, and instruction,* (pp. 99-167). Norwood, NJ: Ablex Publishing Corporation.

Guha, R. V. and Lenat, D. B. (1990). Cyc: A midterm report. *AI Magazine, 11*(3), 32-59.

Hager, J. C. and Ekman, P. (1981). Methodological problems in Tourangeau and Ellsworth's study of facial expression and experience of emotion. *Journal of Personality and Social Psychology, 40,* 358-362.

Hamilton, V., Bower, G. H. and Frijda, N. H. (Eds.) (1988). *Cognitive perspectives on emotion and motivation.* Dordrecht: Kluwer.

Hansen, C. H. and Hansen, R. D. (1988). Finding the face in the crowd: An anger superiority effect. *Journal of Personality and Social Psychology, 54,* 917-924.

Hardin, G. (1968). The tragedy of the commons. *Science, 162,* 1243-1248.

Harlow, H. F. and Harlow, M. K. (1965). The affectional systems. In A. M. Schrier, H. F. Harlow and F. Stollnitz (Eds.), *Behavior of nonhuman primates: Modern research trends* (pp. 287-334). New York: Academic Press.

Harré, R. (1986). *The social construction of emotions.* Oxford: Blackwell.

Harris, P. L. (1989). *Children and emotion: The development of psychological understanding.* Oxford: Basic Blackwell.

Hasselmo, M. E., Rolls, E. T. and Baylis, G. C. (1989). The role of expression and identity in the face-selective response of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research, 32,* 203-218.

Hatfield, E. and Rapson, R. (1993). Love and attachment processes. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 595-604). New York: The Guilford Press.

Hatfield, E., Brinton, C. and Cornelius, J. (1989). Passionate love and anxiety in young adolescents. *Motivation and Emotion, 13,* 271-289.

Hatfield, E., Cacioppo, J. T and Rapson, R. L. (1992). Primitive emotional contagion. In M. S. Clark (Ed.), *Review of Personality and Social Psychology* (Vol. 14, pp. 151-177). Newbury Park, Ca.: Sage.

Hatfield, E., Cacioppo, J. T and Rapson, R. L. (1994). *Emotional contagion.* Cambridge, England: Cambridge University Press.

Hewitt, C. E. (1971). *Description and theoretical analysis (using schemata) of PLANNER: A language for proving theorems and manipulating models in a robot.* Ph. D. Thesis, MIT.

Hewitt, C. E. (1985). The challenge of open systems. *Byte,* 10: 223-242, April 1985.

Hewitt, C. E. (1991). Open information systems semantics for distributed artificial intelligence. *Artificial Intelligence, 47,* 79-106.

Hillman, R. G. (1983) The psychopathology of being held hostage. In L. Z. Freedman and Y. Alexander (Eds.), *Perspectives on terrorism* (pp. 157-165). Wilmington, Delaware: Scholarly Resources.

Hoffman, H. S. (1974) Fear-mediated processes in the context of imprinting. In M. Lewis and L. A. Rosenblum (Eds.), *The origins of fear* (pp. 25-48). New York: John Wiley and Sons.

Hoffman, R. R. (Ed.) (1992). *The psychology of expertise. Cognitive research and empirical AI.* New York: Springer-Verlag.

Holyoak, K. J. and Thagard, P. (1995). *Mental leaps: Analogy in creative thought.* Cambridge: MIT Press.

Huhns, M. N. (Eds.) (1987). *Distributed artificial intelligence.* London: Pitman/Morgan Kaufmann Publishers.

Hupka, R. B. (1991). The motive for the arousal of romantic jealousy: Its cultural origin. In P. Salovey (Ed.). *The psychology of jealousy and envy* (pp. 252-270). New York: The Guilford Press.

Hutchinson, R. R. (1972). The environmental causes of agression. In J. K. Coles and D. D. Jenson (Eds.), *Nebraska Symposium on Motivation 1972* (Vol. 20, pp. 155-181). Lincoln: University of Nebraska Press.

Insel, T. R. (1992). Oxytocin and the neurobiology of attachment. *Behavioral and Brain Sciences, 15,* 515-516.

Isen, A. M. (1993). Positive affect and decision making. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 261-277). New York: The Guilford Press.

Izard, C. E. (1971). *The face of emotion.* New York: Appleton-Century-Crofts.

Izard, C. E. (1977). *Human emotions.* New York: Plenum.

Izard, C. E. (1981). Differential emotions theory and the facial feedback hypothesis of emotion activation: Comments on Tourangeau and Ellsworth's "The role of facial

response in the experience of emotion". *Journal of Personality and Social Psychology, 40,* 350-354.

Izard, C. E. (1984). Emotion-cognition relationships and human development. In C. E. Izard, J. Kagan and R. B. Zajonc (Eds.), *Emotions, cognition and behavior* (pp. 17-37). Cambridge, England: Cambridge University Press.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review, 99,* 561-565.

Izard, C. E. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review, 100,* 68-90.

Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin, 115,* 288-299.

Izard, C. E., Kagan, J. and Zajonc, R. B. (Eds.) (1983). *Emotions, cognition and behavior.* Cambridge, England: Cambridge University Press.

Jennings, N. R. (1993). Commitments and conventions: The foundation of coordination in multi-agent systems. *Knowledge Engineering Review, 8,* 223-250.

Johnson-Laird, P. N. (1988). *The computer and the mind: An introduction to cognitive science.* Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. and Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition and Emotion, 3,* 81-123.

Johnson-Laird, P. N. and Oatley, K. (1992). Basic emotions, rationality and folk theory. *Cognition and Emotion, 6,* 201-223.

Jolly, A. (1966). Lemur social behavior and primate intelligence. *Science, 153,* 501-506.

Kaplan, C. A. and Simon, H. A. (1990). In search of insight. *Cognitive psychology, 22,* 374-419.

Keltner, D. and Buswell, B. N. (1996). Evidence for the distinctness of embarrassment, shame, and guilt: A study of recalled antecedents and facial expressions of emotion. *Cognition and Emotion, 10*(2), 155-171.

Kemper, T. D. (1987). How many emotions are there? Wedding the social and the autonomic components. *American Journal of Sociology, 93,* 263-289.

Kerr, N. L. and Kaufman-Gilliland, C. M. (1994). Communication, commitment, and cooperation in social dilemmas. *Journal of Personality and Social Psychology, 66*(3), 513-529.

Kiesler, C. A. (1971). *The psychology of commitment: experiments linking behavior to belief.* New York : Academic Press.

Kihlstrom, J. F. (1987). The cognitive unconscious. *Science, 237,* 1445-1452.

Kirouac, G. (1989). *Les émotions. Monographies de psychologie, No. 8.* Québec: Presses de l'Université du Québec.

Kirouac, G. (1993). Les émotions. In R. J. Vallerand and E. E. Thill (Eds.), *Introduction à la psychologie de la motivation* (pp. 41-82). Laval, Québec: Éditions Études Vivantes.

Kirsh, D. (1991a). Foundations of AI: The big issues *Artificial Intelligence, 47,* 3-30.

Kirsh, D. (1991b). Today the earwig, tomorrow man? *Artificial Intelligence, 47,* 161-184.

Kitayama, S., Markus, H. R. and Matsumoto, H. (1995). Culture, self, and emotion: A cultural perspective on "self-conscious" emotions. In J. P. Tangney and K. W. Fischer (Eds.), *Self-conscious emotions. The psychology of shame, guilt, embarrassment, and pride* (pp. 439-464). New York: The Guilford Press.

Kleinginna, P. R. and Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion, 5,* 345-379

Kling, A. S. (1986). The anatomy of aggression and affiliation. In R. Plutchik and H. Kellerman, (Eds.), *Emotion: Theory, research, and experience. Vol. 3: Biological foundations of emotions,* (pp. 237-264). New York: Academic Press.

Klinnert, M. D. (1984). The regulation of infant behavior by maternal facial expression. *Infant Behavior and Development, 7,* 447-465.

Klinnert, M. D., Campos, J. J., Sorce, J. F., Emde, R. N. and Svejda, M. (1983). Emotions as behavior regulators: Social referencing in infancy. In R. Plutchik and H. Kellerman (Eds.), *Emotion: Theory, research, and experience: Vol. 2. Emotions in early development* (pp. 57-86). New York: Academic Press.

Kolb, B. and Taylor, L. (1990). Neocortical substrates of emotional behavior. In N. Stein, B. Leventhal and T. Trabasso, (Eds.), *Psychological and biological approaches to emotion,* (pp. 115-144). Hillsdale, NJ: Erlbaum.

Konner, M. (1986). The stranger. *The Sciences,* March/April 1986, 6-7.

Kornfield, W. A. and Hewitt, C. E. (1981). The scientific community metaphor. *IEEE Transactions on Systems, Man and Cybernetics, 11,* 24-33.

Kraemer, G. W. (1992). A psychobiological theory of attachment. *Behavioral and Brain Sciences, 15,* 493-541.

Kraut, R. E. and Johnston, R. E. (1979). Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology, 37*(9), 1539-1553.

Krebs, D. (1987). The challenge of altruism in biology and psychology. In C. Crawford, M. Smith and D. Krebs, (Eds.), *Sociobiology and psychology: Ideas, issues and applications,* (pp. 81-118). Hillsdale, NJ: Erlbaum.

Krebs, J. R. and Davies, N. B. (1987). *An introduction to behavioural ecology.* (Second edition). Oxford: Blackwell Scientific Publications.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* (Second edition). Chicago: University of Chicago Press.

Kulkarni, D. and Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science, 12,* 139-176.

Kunst-Wilson, W. R. and Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science, 207,* 557-558.

Kurland, D. M. and Kurland, L. C. (1987). Computer applications in education: A historical overview. *Annual Review of Computer Science, 2,* 317-358.

Laird, J. D. (1974). Self-attribution of emotion: The effects of expressive behavior on the quality of emotional experience. *Journal of Personality and Social Psychology, 29,* 475-486.

Laird, J. D. (1984). The real role of facial response in the experience of emotion: A reply to Tourangeau and Ellsworth, and others. *Journal of Personality and Social Psychology, 47,* 909-917.

Lange, C. G. and James, W. (1967). *The emotions.* (Collected papers from 1884, 1885 and 1890). New York: Hafner.

Lapierre, D. and Braun, M. J. (1993). La physiologie de la motivation. In R. J. Vallerand and E. E. Thill (Eds.), *Introduction à la psychologie de la motivation* (pp. 139-177). Laval, Québec: Éditions Études Vivantes.

Larsen, R. J. and Diener, W. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark (Ed.), *Review of Personality and Social Psychology* (Vol. 13, pp. 25-59). Newbury Park, Ca.: Sage.

Lazarus, R. S. (1981). A cognitivist reply to Zajonc on emotion and cognition. *American Psychologist, 36,* 222-223.

Lazarus, R. S. (1982). Thoughts on the relations between emotion and cognition. *American Psychologist, 37,* 1019-1024.

Lazarus, R. S. (1984). On the primacy of cognition. *American Psychologist, 39,* 124-129.

Lazarus, R. S. (1991a). *Emotion and adaptation.* New York: Oxford University Press.

Lazarus, R. S. (1991b). Cognition and motivation in emotion. *American Psychologist, 46,* 352-367.

Lazarus, R. S. and Smith, C. A. (1988). Knowledge and appraisal in the cognition-emotion relationship. *Cognition and Emotion, 2,* 281-300.

LeDoux, J. E. (1986). The neurobiology of emotion. In J. E. LeDoux and W. Hirst (Eds.), *Mind and brain* (pp. 301-354). New York: Cambridge University Press.

LeDoux, J. E. (1987). Emotion. In F. Plum, (Ed.), *Hanbook of physiology: Section I: The nervous system. Vol. 5: Higher functions of the brain,* (pp. 419-460). Bethesda, MD: American Physiological Society.

LeDoux, J. E. (1989). Cognitive-emotional interactions in the brain. *Cognition and Emotion, 3,* 267-289.

LeDoux, J. E. (1993). Emotional networks in the brain. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 109-118). New York: The Guilford Press.

Leeper, R. W. (1948). A motivational theory of emotion to replace "Emotion as disorganized response". *Psychological Review,* 55, 5-21.

Lehnert, W. G. and Vine, E. W. (1987). The role of affect in narrative structure. *Cognition and Emotion, 1,* 299-322.

Lemerise, E. A. and Dodge, A. D.(1993). The development of anger and hostile interactions. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 537-546). New York: The Guilford Press.

Lenat, D. B. (1975). BEINGs: Knowledge as interactings experts. *Proceedings of the 1975 International Joint Conference on Artificial Intelligence,* 126-133.

Lepper, M. R. (1983). Extrinsic reward and intrinsic motivation: Implications for the classroom. In J. M. Levine and M. C. Wang, (Eds.), *Teacher and student perceptions: Implications for learning,* (pp. 281-317). Hillsdale, NJ: Erlbaum.

Lepper, M. R. (1988). Motivational considerations in the study of instruction. *Cognition and Instruction, 5*(4), 289-309.

Leventhal, H. and Scherer, K. R. (1987). The relationship of emotion to cognition: A functional approach to a semantic controversy. *Cognition and Emotion, 1,* 3-28.

Levine, L. J. (1996). The anatomy of disappointment: A naturalistic test of appraisal models of sadness, anger, and hope. *Cognition and Emotion, 10*(4), 337-359.

Levy, R. I. (1984a). The emotions in comparative perspective. In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 397-412). Hillsdale, N.J.: Erlbaum.

Levy, R. I. (1984b). Emotion, knowing, and culture. In R. A. Shweder and R. A. LeVine (Eds.), *Culture theory: Essays on mind, self, and emotion* (pp. 214-237). Cambridge, England: Cambridge University Press.

Lewis, M. (1993a). The emergence of human emotions. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 223-235). New York: The Guilford Press.

Lewis, M. (1993b). Self-conscious emotions: Embarrassment, pride, shame, and guilt. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 563-573). New York: The Guilford Press.

Lewis, M. (1995a). Self-conscious emotions. *American Scientist, 83*(1), 68-78.

Lewis, M. (1995b). Embarrassment: The emotion of self-exposure and evaluation. In J. P. Tangney and K. W. Fischer (Eds.), *Self-conscious emotions. The psychology of shame, guilt, embarrassment, and pride* (pp. 198-218). New York: The Guilford Press.

Lewis, M. and Rosenblum, L. A. (Eds.) (1974). *The effect of the infant on its caregiver.* New York: John Wiley and Sons.

Lindsay, P. H. and Norman, D. A. (1972). *Human information processing: An introduction to psychology.* New York: Academic Press.

Lindsay-Hartz, J. (1984). Contrasting experiences of shame and guilt. *American Behavioral Scientist, 27*(6), 689-704.

Lomborg, B. (1992). Game theory vs. multiple agents: The iterated prisoner's dilemma. In C. Castelfranchi and E. Werner, (Eds.), *Artificial social systems,* Proceedings of the 4th European Workshop on MAAMAW, Lecture Notes on Artificial Intelligence, No. 830, (pp. 69-93). Berlin: Springer-Verlag.

Lorberbaum, J. (1994). No strategy is evolutionary stable in the repeated Prisoner's Dilemma. *Journal of Theoretical Biology, 168,* 117-130.

Lorenz, K. (1970). *Tous les chiens, tous les chats.* Paris: Flammarion.

Luck, M. and d'Inverno, M. (1995). A formal framework for agency and autonomy. In *Proceedings of the First International Conference on Multi-Agent Systems,* (pp. 254-260). Cambridge: AAAI Press/MIT Press.

Lutz, C. (1986). The domain of emotion words on Ifaluk. In R. Harré (Ed.), *The social construction of emotions* (pp. 267-288). Oxford: Blackwell.

Lutz, C. (1988). Ethnographic perspectives on the emotion lexicon. In V. Hamilton, G. H. Bower and N. H. Frijda (Eds.), *Cognitive perspectives on emotion and motivation* (pp. 399-419). Dordrecht: Kluwer.

Lyman, B. and Waters, J. C. E. (1986). The experiential loci and sensory qualities of various emotions. *Motivation and Emotion, 10*(1), 25-37.

MacLean, P. D. (1993). Cerebral evolution of emotion. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 67-83). New York: The Guilford Press.

MacLean, P. D. and Guyot, R. (1990). *Les trois cerveaux de l'homme.* Paris: Robert Laffont.

Maes, P. (Ed.) (1991a). *Designing autonomous agents: Theory and practice from biology to engineering and back.* Cambridge: MIT Press.

Maes, P. (1991b). Situated agents can have goals. In P. Maes (Ed.), *Designing autonomous agents: Theory and practice from biology to engineering and* back (pp. 49-70). Cambridge, MA: MIT Press.

Maes, P., Mataric, M., Meyer, J.-A., Pollack, J. and Wilson, S. W. (Eds.) (1996). *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior.* Cambridge, MA: The MIT Press/Bradford Books.

Malatesta, C. Z. (1981). Infant emotion and the vocal affect lexicon. *Motivation and Emotion, 5,* 1-23.

Malatesta, C. Z. and Haviland, J. M. (1982). Learning display rules: The socialization of emotion expression in infancy. *Child Development, 53,* 991-1003.

Mandler, G. (1964). The interruption of behavior. In D. Levine, (Ed.), *Nebraska Symposium on Motivation,* (pp. 163-219). Lincoln: University of Nebraska Press.

Mandler, G. (1975). *Mind and emotion.* New York: John Wiley and Sons.

Mandler, G. (1984). *Mind and body.* New York: Norton.

Mandler, G., Nakamura, Y. and Van Zandt, B. J. S. (1987). Nonspecific effects of exposure on stimuli that cannot be recognized. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 646-648.

Marler, P. (1984). Animal communication: Affect or cognition? In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 345-365). Hillsdale, N.J.: Erlbaum.

Mascolo, M. F. and Fischer, K. W. (1995). Developmental transformations in appraisals for pride, shame, and guilt. In J. P. Tangney and K. W. Fischer (Eds.), *Self-conscious emotions. The psychology of shame, guilt, embarrassment, and pride* (pp. 64-113). New York: The Guilford Press.

Mathes, E. W., Adams, H. E. and Davies, R. M. (1985). Jealousy: Loss of relationship rewards, loss of self-esteem, depression, anxiety, and anger. *Journal of Personality and Social Psychology, 48*(6), 1552-1561.

Matsumoto, D. (1987). The role of facial response in the experience of emotion: More methodological and a meta-analysis. *Journal of Personality and Social Psychology, 52,* 769-774.

Mauro, R. (1992). Affective dynamics: Opponent processes and excitation transfer. In M. S. Clark (Ed.), *Review of Personality and Social Psychology* (Vol. 13, pp. 150-174). Newbury Park, Ca.: Sage.

Mauss, M. (1950). Essai sur le don. Forme et raison de l'échange dans les sociétés archaïques. In M. Mauss, *Sociologie et anthropologie* (pp. 143-279). Paris: Presses Universitaires de France.

McCarthy, J. (1990). *Formalizing common sense.* Norwood, NJ: Ablex.

McFarland, D. (1995). Autonomy and self-sufficiency in robots. In L. Steels and R. Brooks, (Eds.), *The artificial life route to artificial intelligence,* (pp. 187-213). Hillsdale, NJ: Erlbaum.

McGraw, K. M. (1987). Guilt following transgression: An attribution of responsibility approach. *Journal of Personality and Social Psychology, 53*(2), 247-256.

McGuire, T. R. (1993). Emotion and behavior genetics in vertebrates and invertebrates. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 155-166). New York: The Guilford Press.

Mead, G. H. (1934). *Mind, self, and society from the standpoint of a social behaviorist.* Chicago: The University of Chicago Press.

Mealey, L. (1995). The sociobiology of sociopathy: An integrated evolutionary model. *Behavioral and Brain Sciences, 18,* 523-599.

Meltzoff, A. N. and Gopnik, A. (1993). The role of imitation in understanding and developping a theory of mind. In S. Baron-Cohen, H. Tager-Flusberg and D. J. Cohen, (Eds.), *Understanding other minds. Perspectives from autism.* New York: Oxford University Press.

Meltzoff, A. N. and Moore, K. M. (1994). Imitation, memory, and the representation of persons. *Infant Behavior and Development, 17,* 83-99.

Michotte, A. E. (1950). The emotions regarded as functional connections. In M. Reymert (Ed.), *Feelings and emotions. The Mooseheart symposium* (pp. 114-126). New York: McGraw-Hill.

Mineka, S., Davidson, M., Cook, M. and Keir, R. (1984). Observational conditioning of snake fear in rhesus monkeys. *Journal of Abnormal Psychology, 93,* 355-372.

Minsky, M. (Ed.) (1968). *Semantic information processing.* Cambridge: MIT Press.

Minsky, M. (1970). Form and content in computer science. *Journal of the Association for Computing Machinery, 17,* 197-205.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), *The Psychology of Computer Vision* (pp. 211-281). New York: McGraw-Hill.

Minsky, M. (1985). *The society of mind.* New York: Simon and Schuster.

Minsky, M. and Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge: MIT Press.

Mitchell, R. W. (1987). A comparative-developmental approach to understanding imitation. In P. P. G. Bateson and P. H. Klopfer, (Eds.), *Perspective in ethology, Volume 7,* (pp. 183-215). New York: Plenum Press.

Moreland, R. L. and Zajonc, R. B. (1977). Is stimulus recognition a necessary condition for the occurrence of exposure effects? *Journal of Personality and Social Psychology, 35,* 191-199.

Moreland, R. L. and Zajonc, R. B. (1979). Exposure effects may not depend on stimulus recognition. *Journal of Personality and Social Psychology, 37,* 1085-1089.

Murray, A. D. (1979). Infant crying as an elicitor of parental behavior: An examination of two models. *Psychological Bulletin, 86,* 191-215.

Neisser, U. (1963). The imitation of man by machine. *Science, 139,* 193-197.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Newell, A. (1990). *Unified theories of cognition.* Cambridge: Harvard University Press.

Newell, A. and Simon, H. A. (1963). GPS, a program that simulates human thought. In E. A. Feigenbaum and J. Feldman, (Eds.), *Computers and thought,* (pp. 279-293). New York: McGraw-Hill.

Newell, A. and Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Norman, D. A. (1980). Twelve issues for cognitive science. *Cognitive Science, 4,* 1-32.

Norman, D. A. (Ed.) (1981). *Perspectives on cognitive science.* Norwood, NJ: Ablex.

Norman, T. J. and Long, D. (1995). *Alarms: An implementation of motivated agency.* Research Note RN/95/24, Department of Computer Science, University College London, London, UK, April 6, 1995.

Nowak, M. and Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature, 364,* 56-58.

Nowak, M. A., May, R. M. and Sigmund, K. (1995). The arithmetics of mutual help. *Scientific American, June 1995, 272*(6), 76-81.

Oatley, K. (1992). *Best laid schemes: The psychology of emotions.* Cambridge, England: Cambridge University Press.

Oatley, K. and Duncan, E. (1992). Incidents of emotion in daily life. In K. T. Strongman (Ed.), *International Review of Studies on Emotion,* (Vol. 2, pp. 249-293). New York: John Wiley and Sons.

Oatley, K. and Duncan, E. (1994). The experience of emotions in everyday life. *Cognition and Emotion, 8*(4), 369-381.

Oatley, K. and Johnson-Laird, P. N. (1987). Towards a cognitive theory of emotions. *Cognition and Emotion, 1*(1), 29-50.

Oatley, K. and Johnson-Laird, P. N. (1990). Semantic primitives for emotions: A reply to Ortony and Clore. *Cognition and Emotion, 4,* 129-143.

Oatley, K. and Johnson-Laird, P. N. (1996). The communicative theory of emotions: Empirical tests, mental models, and implications for social interaction. In L. L. Martin and A. Tesser (Eds.), *Striving and feeling. Interactions among goals, affect and self-regulation.* Hillsdale, N.J.: Erlbaum.

Öhman, A. (1986). Face the beast and fear the face: Animal and social fears as prototypes for evolutionary analyses of emotion. *Psychophysiology, 23,* 123-145.

Öhman, A. (1993). Fear and anxiety as emotional phenomena: Clinical phenomenology, evolutionary perspectives, and information-processing mechanisms. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 511-536). New York: The Guilford Press.

Öhman, A. and Dimberg, U. (1978). Facial expressions as conditioned stimuli for electrodermal responses: A case of "preparedness"? *Journal of Personality and Social Psychology, 36*(11), 1251-1258.

Orbell, J. M., van de Kragt, A. J. C. and Dawes, R. M. (1988). Explaining discussion-induced cooperation. *Journal of Personality and Social Psychology, 54*(5), 811-819.

Ortony, A. (1987). Is guilt an emotion?. *Cognition and Emotion, 1*(3), 283-298.

Ortony, A. and Clore, G. L. (1989). Emotions, moods, and conscious awareness. Comments on Johnson-Laird and Oatley's "The language of emotions: An analysis of a semantic field". *Cognition and Emotion, 3,* 125-137.

Ortony, A. and Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review, 97,* 315-331.

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

Ortony, A., Clore, G. L. and Foss, M. A. (1987). The referential structure of the affective lexicon. *Cognitive Science, 11,* 341-364.

Osgood, C. E. (1966). Dimensionality of the semantic space for communication via facial expressions. *Scandinavian Journal of Psychology, 7,* 1-30.

Oster, H., Hegley, D. and Nagel, L. (1992). Adult judgments and fine-grained analysis of infant facial expressions: Testing the validity of a priori coding formulas. *Developmental Psychology, 28,* 1115-1131.

Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences, 5,* 407-467.

Panksepp, J. (1986). The anatomy of emotions. In R. Plutchik and H. Kellerman (Eds.), *Emotion: Theory, research, and experience: Vol. 3. Biological foundations of emotions* (pp. 91-124). New York: Academic Press.

Panksepp, J. (1989). Les circuits des émotions. *Science et Vie, Hors série, 168,* 58-67.

Panksepp, J. (1991). Affective neuroscience: A conceptual framework for the neurobiological study of emotions. In K. T. Strongman (Ed.), *International Review of Studies on Emotion,* (Vol. 1, pp. 59-99). New York: John Wiley and Sons.

Panksepp, J. (1992). A critical role for "affective neuroscience" in resolving what is basic about basic emotions. *Psychological Review, 99,* 554-560.

Panksepp, J. (1993). Neurochemical control of moods and emotions: Amino acids to neuropeptides. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 87-107). New York: The Guilford Press.

Papert, S. (1973). Theory of knowledge and complexity. In G. J. Dalenoort (Ed.), *Process models for psychology. Lecture notes of the NUFFIC international summer course, 1972* (pp. 1-49). Rotterdam: Rotterdam University Press.

Papert, S. (1980). *Mindstorms.* New York: Basic Books.

Paquette, G., Aubin, C. and Crevier, F. (1994). An intelligent support system for course design. *Educational Technology,* November-December 1994, 50-57.

Parkinson, B. and Manstead, A. S. R. (1992). Appraisal as a cause of emotion. In M. S. Clark (Ed.), *Review of Personality and Social Psychology* (Vol. 13, pp. 122-149). Newbury Park, Ca.: Sage.

Parrot, W. G. and Smith, S. F. (1991). Embarrassment: Actual vs. typical cases, classical vs. prototypical representations. *Cognition and Emotion, 5,* 467-488.

Patrick, C. J. (1994). Emotion and psychopathy: Starting new insights. *Psychophysiology, 31,* 319-330.

Pea, R. D. and Kurland, D. M. (1987). Logo programming and the development of planning skills. In R. D. Pea and K. Sheingold, (Eds.), *Mirrors of minds: Patterns of experience in educational computing.* Norwood, NJ: Ablex Publishing Corporation.

Pedersen, C. A., Ascher, J. A., Monroe, Y. L. and Prange, A. J., Jr. (1982). Oxytocin induces maternal behavior in virgin female rats. *Science, 216,* 648-650

Piaget, J. (1927). *La représentation du monde chez l'enfant.* Paris: Presses Universitaires de France.

Piaget, J. (1936). *La naissance de l'intelligence chez l'enfant.* Neuchâtel: Delachaux et Niestlé.

Piaget, J. (1947). *La psychologie de l'intelligence.* Paris: Armand Colin.

Piaget, J. (1967). *Biologie et connaissance.* Paris: Gallimard.

Pintrich, P. R. and De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational psychology, 82*(10), 33-40.

Pintrich, P. R. and Garcia, T. (1994). Self-regulated learning in college students: Knowledge, strategies and motivation. In P. R. Pintrich, D. R. Brown and C. E. Weinstein (Eds.), *Student motivation, cognition, and learning. Essays in honor of Wilbert J. McKeachie,* (p.113-133). Hillsdale, NJ: Erlbaum..

Pintrich, P R., Brown, D. R. and Weinstein, C. E. (Eds.) (1994). *Student motivation, cognition, and learning. Essays in honor of Wilbert J. McKeachie.* Hillsdale, NJ: Erlbaum..

Pittam, J. and Scherer, K. R. (1993). Vocal expression and communication of emotion. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 185-197). New York: The Guilford Press.

Ploog, D. (1986). Biological foundations of the vocal expressions of emotions. In R. Plutchik and H. Kellerman (Eds.), *Emotion: Theory, research, and experience: Vol. 3. Biological foundations of emotions* (pp. 173-197). New York: Academic Press.

Plutchik, R. (1980). *Emotion: A general psychoevolutionary synthesis.* New York: Harper and Row.

Plutchik, R. (1984). Emotion: A general psychoevolutionary theory. In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 197-219). Hillsdale, N.J.: Erlbaum.

Plutchik, R. and Kellerman, H. (Eds.) (1980). *Emotion: Theory, research, and experience: Vol. 1. Theories of emotions.* New York: Academic Press.

Plutchik, R. and Kellerman, H. (Eds.) (1983). *Emotion: Theory, research, and experience: Vol. 2. Emotions in early development.* New York: Academic Press.

Plutchik, R. and Kellerman, H. (Eds.) (1986). *Emotion: Theory, research, and experience: Vol. 3: Biological foundations of emotions.* New York: Academic Press.

Pope, L. K. and Smith, C. A. (1994). On the distinct meaning of smiles and frowns. *Cognition and Emotion, 8*(1), 65-72.

Popper, K. R. (1979). *Objective knowledge.* Oxford, U. K.: Clarendon Press.

Povinelli, D. J. and Eddy, T. J. (1996). Chimpanzees: Joint visual attention. *Psychological Science, 7*(3), 129-135.

Provine, R. R. (1992). Contagious laughter: Laughter is a sufficient stimulus for laughs and smiles. *Bulletin of the Psychonomic Society, 30,* 1-4.

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation of cognitive science.* Cambridge: MIT Press.

Redican, W. K. (1982). An evolutionary perspective on human facial displays. In P. Ekman (Ed.), *Emotion in the human face* (pp. 212-280). Cambridge, England: Cambridge University Press.

Reite, M. and Field, T. (Eds.) (1985). *The psychobiology of attachment and separation.* New York: Academic Press.

Rogoff, B. and Lave, J. (Eds.) (1984). *Everyday cognition: Its development in social context.* Cambridge: Harvard University Press.

Rolls, E. T. (1990). A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition and Emotion, 4,* 161-190.

Rosaldo, M. Z. (1984). Toward an anthropology of self and feeling. In R. A. Shweder and R. A. LeVine (Eds.), *Culture theory: Essays on mind, self, and emotion* (pp. 137-157). Cambridge, England: Cambridge University Press.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. In P. Shaver (Ed.), *Review of Personality and Social Psychology* (Vol. 5, pp. 11-36). Beverly Hills, Ca.: Sage.

Roseman, I. J. (1991). Appraisal determinants of discrete emotions. *Cognition and Emotion, 5,* 161-200.

Roseman, I. J., Aliki Antoniou, A. and Jose, P. E. (1996). Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion, 10*(3), 241-277.

Roseman, I. J., Wiest, C. and Swartz, T. S. (1994). Phenomenology, behaviors, and goals differentiate discrete emotions. *Journal of Personality and Social Psychology, 67(2),* 206-221.

Rosenblum, L. A. and Alpert, S. (1974) Fear of strangers and specificity of attachment in monkeys. In M. Lewis and L. A. Rosenblum (Eds.), *The origins of fear* (pp. 165-193). New York: John Wiley and Sons.

Rozin, P. and Fallon, A. E. (1987). A perspective on disgust. *Psychological Review, 94,* 23-41.

Rozin, P., Haidt, J. and McCauley, C. R. (1993). Disgust. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 575-594). New York: The Guilford Press.

Rozin, P., Lowery, L. and Ebert, R. (1994). Varieties of disgust faces and the structure of disgust. *Journal of Personality and Social Psychology, 66*(5), 870-881.

Ruch, W. (1993). Exhilaration and humor. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 605-616). New York: The Guilford Press.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Bobrow and A. M. Collins, (Eds.), *Representation and understanding: Studies in cognitive science.* New York: Academic Press.

Rumelhart, D. E. (1981). Schemata: The building blocks of cognition. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 3-26). Newark: International Reading Association.

Russell, J. A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology, 37,* 345-356.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39,* 1161-1178.

Russell, J. A. (1989). Culture, scripts, and children's understanding of emotion. In C. Saarni and P. L. Harris (Eds.), *Children's understanding of emotion* (pp. 293-318). Cambridge, England: Cambridge University Press.

Russell, J. A. (1991a). In defense of a prototype approach to emotion concepts. *Journal of Personality and Social Psychology, 60,* 37-47.

Russell, J. A. (1991b). The contempt expression and the relativity thesis. *Motivation and Emotion, 15,* 149-168.

Russell, J. A. (1991c). Culture and the categorization of emotion. *Psychological Bulletin, 110,* 426-450.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin, 115,* 102-141.

Russell, J. A. and Bullock, M. (1986). Fuzzy concepts and the perception of emotion in facial expressions. *Social Cognition, 4,* 309-341.

Russell, J. A., Lewicka, M. and Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology, 57,* 848-856.

Ryan, B. (1991). Dynabook revisited with Alan Kay. *Byte,* 16(2): 203-208, February 1991.

Sackett, G. P. (1966). Monkeys reared in isolation with pictures as visual input: Evidence for an innate releasing mechanism. *Science, 154,* 1468-1473.

Salovey, P. and Mayer, J. D. (1989). Emotional intelligence. *Imagination, Cognition and Personality, 9*(3), 185-211.

Salovey, P. and Rodin, J. (1986). The differentiation of social-comparison jealousy and romantic jealousy. *Journal of Personality and Social Psychology, 50*(6), 1100-1112.

Schachter, S. and Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69,* 379-399.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, goals, plans and understanding. An inquiry into human knowledge structures.* Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 293-317). Hillsdale, N.J.: Erlbaum.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99,* 143-165.

Scherer, K. R. (1988a). Criteria for emotion-antecedent appraisal: A review. In V. Hamilton, G. H. Bower and N. H. Frijda (Eds.), *Cognitive perspectives on emotion and motivation* (pp. 89-126). Dordrecht: Kluwer.

Scherer, K. R. (Ed.) (1988b). *Facets of emotion: Recent research.* Hillsdale, N.J.: Erlbaum.

Scherer, K. R. (1992). What does facial expression express? In K. T. Strongman (Ed.), *International Review of Studies on Emotion,* (Vol. 2, pp. 139-165). New York: John Wiley and Sons.

Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition and Emotion, 7,* 325-355.

Scherer, K. R. and Ekman, P. (Eds.) (1984). *Approaches to emotion.* Hillsdale, N.J.: Erlbaum.

Scherer, K. R. and Tannenbaum, P. H. (1986). Emotional experiences in everyday life: A survey approach. *Motivation and Emotion, 10*(4), 295-314.

Scherer, K. R. and Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology, 66*(2), 310-328.

Scherer, K. R., Wallbott, H. G. and Summerfield, A. B. (1986). *Experiencing emotion: A cross-cultural study.* Cambridge, England: Cambridge University Press.

Scherer, K. R., Banse, R., Wallbott, H. G. and Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion, 15,* 123-148.

Scherer, K. R., Wallbott, H. G., Matsumoto, D. and Kudoh, T. (1988). Emotional experience in cultural context: A comparison between, Europe, Japan, and the United States. In K. R. Scherer (Ed.), *Facets of emotion: Recent research* (pp. 5-30). Hillsdale, N.J.: Erlbaum.

Schlosberg, H. (1941). A scale for the judgment of facial expressions. *Journal of Experimental Psychology, 29,* 497-510.

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology, 44,* 229-237.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review, 61,* 81-88.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26,* 207-231.

Schunk, D. H. and Meece, J. L. (Eds.) (1992). *Student perceptions in the classroom.* Hillsdale, NJ: Erlbaum.

Scott, J. P. (1980). The function of emotions in behavioral systems: A systems theory analysis. In R. Plutchik and H. Kellerman (Eds.), *Emotion: Theory, research, and experience: Vol. 1. Theories of emotions* (pp. 35-56). New York: Academic Press.

Seamon, J. G., Brody, N. and Kauff, D. M. (1983a). Affective discrimination of stimuli that are not recognized: Effects of shadowing, masking, and cerebral laterality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 544-555.

Seamon, J. G., Brody, N. and Kauff, D. M. (1983b). Affective discrimination of stimuli that are not recognized: II. Effect of delay between study and test. *Bulletin of the Psychonomic Society, 21,* 187-189.

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

Searle, J. R. (1979). *Expression and meaning: Studies in the theory of speech acts.* Cambridge: Cambridge University Press.

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

Searle, J. R. (1990). Collective intentions and actions. In P. R. Cohen, J. Morgan and M. E. Pollack (Eds.), *Intentions in communications* (pp. 401-415). Cambridge, MA: MIT Press.

Sharpsteen, D. J. (1991). The organization of jealousy knowledge: Romantic jealousy as a blended emotion. In P. Salovey (Ed.). *The psychology of jealousy and envy* (pp. 31-51). New York: The Guilford Press.

Shaver, P. R., Wu, S. and Schwartz, J. (1992). Cross-cultural similarities and differences in emotion and its representation: A prototype approach. In M. S. Clark (Ed.), *Review of Personality and Social Psychology* (Vol. 13, pp. 175-212). Newbury Park, Ca.: Sage.

Shaver, P.R., Schwartz, J., Kirson, D. and O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. In M. S. Clark (Ed.), *Journal of Personality and Social Psychology, 52,* 1061-1086.

Sherry, D. F. and Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review, 94,* 439-454.

Shoham, Y. (1993). Agent-Oriented Programming. *Artificial Intelligence, 60,* 51-92.

Shoham, Y. and Tennenholtz, M. (1995). On social laws for artificial agent societies: Off-line design. *Artificial Intelligence, 73,* 231-252.

Shultz, T. R. (1991). From agency to intention : A rule-based, computational approach. In A. Whiten (Ed.), *Natural theories of mind. Evolution, development and simulation of everyday mindreading* (pp. 79-95). Oxford: Blackwell.

Shweder, R. A. and LeVine, R. A. (Eds.) (1984). *Culture theory: Essays on mind, self, and emotion.* Cambridge, England: Cambridge University Press.

Simner, M. L. (1971). Newborn's response to the cry of another infant. *Developmental Psychology, 5,* 136-150.

Simon, H. (1967). Motivational and emotional controls of cognition. *Psychological Review, 74,* 29-39.

Simon, H. (1980). The behavioral and social sciences. *Science, 209,* 72-78.

Simon, H. (1990a). Invariants of human behavior. *Annual Review of Psychology, 41,* 1-19.

Simon, H. (1990b). A mechanism for social selection and successful altruism. *Science, 250,* 1665-1668.

Simon, H. A. (1995). Explaining the ineffable: AI on the topics of intuition, insight and inspiration. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence,* (pp. 939-948). Montreal, Quebec, Canada.

Singh, M. P. (1994). *Multiagent systems: A theoretical framework for intentions, know-how, and communications.* Lecture Notes on Artificial Intelligence, No. 799. Berlin: Springer-Verlag.

Singley, M. K. and Anderson, J. R. (1989). *The transfer of cognitive skills.* Cambridge: Harvard University Press.

Sloman, A. (1987). Motives, mechanisms, and emotions. *Cognition and Emotion, 1,* 217-233.

Sloman, A. (1992). Prolegomena to a theory of communication and affect. In A. Ortony, J. Slack and O. Stock, (Eds.), *Communication from an artificial intelligence perspective. Theoretical and applied issues,* (pp.229-259). Berlin: Springer-Verlag.

Sloman, A. (1995). Architectures for emotional agents. Conference presented at the *Geneva Emotion Week,* Université de Genève, Geneva, April 8-13, 1995.

Sloman, A. and Croucher, M. (1981). Why robots will have emotions. *Proceedings of the Seventh International Joint Conference on Artificial Conference* (pp. 197-202). Vancouver, British Columbia.

Small, M. H. (1990). Political animal. Social intelligence and the growth of the primate brain. *The Sciences,* March/April 1990, 36-42.

Smith, C. A. (1989). Dimensions of appraisal and physiological response in emotion. *Journal of Personality and Social Psychology, 56,* 339-353.

Smith, C. A. and Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology, 48,* 813-838.

Solomon, R. C. (1984). Getting angry: The Jamesian theory of emotion in anthropology. In R. A. Shweder and R. A. LeVine (Eds.), *Culture theory: Essays on mind, self, and emotion* (pp. 238-254). Cambridge, England: Cambridge University Press.

Solomon, R. L. (1980). The opponent-process theory of acquired motivation: The costs of pleasure and the benefits of pain. *American Psychologist, 35,* 691-712.

Solomon, R. L. and Corbit, J. D. (1974). An opponent-process theory of motivation: Temporal dynamics of affect. *Psychological Review, 81,* 119-145.

Sorce, J. F., Emde, R. N., Campos, J. and Klinnert, M. D. (1985). Maternal emotional signaling: Its effect on the visual cliff behavior of 1-year-olds. *Developmental Psychology, 21,* 195-200.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195-231.

Sroufe, L. A. (1979). Socioemotional development. In J. D. Osofsky (Ed.), *Hanbook of infant development* (pp. 462-516). New York: John Wiley and Sons.

Sroufe, L. A. (1984). The organization of emotional development. In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 109-128). Hillsdale, N.J.: Erlbaum.

Sroufe, L. A., Waters, E. and Matas, L. (1974). Contextual determinants of infant affective response. In M. Lewis and L. A. Rosenblum (Eds.), *The origins of fear* (pp. 49-72). New York: John Wiley and Sons.

Stearns, C. Z. (1993). Sadness. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 547-561). New York: The Guilford Press.

Steck, L. Levitan, D., McLane, D. and Kelley, H. H. (1982). Care, need, and conceptions of love. *Journal of Personality and Social Psychology, 43*(3), 481-491.

Steels, L. (1995). Building agents out of autonomous behavior systems. In L. Steels and R. Brooks, (Eds.), *The artificial life route to artificial intelligence,* (pp. 83-121). Hillsdale, NJ: Erlbaum.

Steels, L. and Brooks, R. (Eds.) (1995). *The artificial life route to artificial intelligence.* Hillsdale, NJ: Erlbaum.

Stefik, M. J. and Bobrow, D. G. (1986). Object-oriented programming: Themes and variations. *The AI Magazine, 6,* 40-62.

Stein, N. L. and Oatley, K. (1992). Basic emotions: Theory and measurement. *Cognition and Emotion, 6,* 161-168.

Stein, N. L. and Trabasso, T. (1992). The organisation of emotional experience: Creating links among emotion, thinking, language, and intentional action. *Cognition and Emotion, 6,* 225-244.

Stein, N. L., Leventhal, B. and Trabasso, T. (Eds.) (1990). *Psychological and biological approaches to emotion.* Hillsdale, NJ: Erlbaum.

Stein, N. L., Trabasso, T. and Liwag, M. (1993). The representation and organization of emotional experience: Unfolding the emotion episode. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 279-300). New York: The Guilford Press.

Sternberg, R. J. (1986). A triangular theory of love. *Psychological Review, 93*(2), 119-135.

Stipek, D. (1995). The development of pride and shame in toddlers. In J. P. Tangney and K. W. Fischer (Eds.), *Self-conscious emotions. The psychology of shame, guilt, embarrassment, and pride* (pp. 237-252). New York: The Guilford Press.

Storm, C. and Storm, T. (1987). A taxonomic study of the vocabulary of emotions. *Journal of Personality and Social Psychology, 53,* 805-816.

Strongman, K. T. (1987). *The psychology of emotion.* New York: John Wiley and Sons.

Swan, K. (1989). Logo programming and the teaching and learning of problem solving. *Journal of Artificial Intelligence in Education, 1*(1), 73-92.

Tangney, J. P. (1990). Assessing individual differences in proneness to shame and guilt: Development of self-conscious affect and attribution inventory. *Journal of Personality and Social Psychology, 59,* 102-111.

Tangney, J. P. (1991). Moral affect: The good, the bad, and the ugly. *Journal of Personality and Social Psychology, 61*(4), 598-607.

Tangney, J. P. (1995). Shame and guilt in interpersonal relationships. In J. P. Tangney and K. W. Fischer (Eds.), *Self-conscious emotions. The psychology of shame, guilt, embarrassment, and pride* (pp. 114-139). New York: The Guilford Press.

Tangney, J. P. and Fischer, K. W. (Eds.) (1995). *Self-conscious emotions. The psychology of shame, guilt, embarrassment, and pride.* New York: The Guilford Press.

Tangney, J. P., Burggraf, S. A. and Wagner, P. E. (1995). Shame-proneness, guilt-proneness, and psychological symptoms. In J. P. Tangney and K. W. Fischer (Eds.), *Self-conscious emotions. The psychology of shame, guilt, embarrassment, and pride* (pp. 343-367). New York: The Guilford Press.

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

Tardif, J. (1992). *Pour un enseignement stratégique: L'apport de la psychologie cognitive.* Montréal: Les Éditions Logiques.

Tardif, J. (1995). Les influences de la psychologie cognitive sur les pratiques d'enseignement et d'évaluation. *Revue québécoise de psychologie, 16*(2), 175-207.

Tesser, A., Gatewood, R. and Driver, M. (1968). Some determinants of gratitude. *Journal of Personality and Social Psychology, 9*(3), 233-236.

Tinbergen, N. (1980). *L'étude de l'instinct.* Paris: Petite Bibliothèque Payot.

Toates, F. (1986). *Motivational systems.* Cambridge: Cambridge University Press.

Toates, F. (1988). Motivation and emotion from a biological perspective. In V. Hamilton, G. H. Bower and N. H. Frijda (Eds.), *Cognitive perspectives on emotion and motivation* (pp. 3-35). Dordrecht: Kluwer.

Tomkins, S. S. (1962). *Affect, imagery and consciousness: Vol. 1.* New York: Springer.

Tomkins, S. S. (1963). *Affect, imagery and consciousness: Vol. 2.* New York: Springer.

Tomkins, S. S. (1981). The role of facial response in the experience of emotion: A reply to Tourangeau and Ellsworth. *Journal of Personality and Social Psychology, 40,* 355-357.

Tomkins, S. S. (1984). Affect theory. In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 163-195). Hillsdale, N.J.: Erlbaum.

Toobey, J. and Cosmides, L. (1990). The past explains the present: Emotion adaptations and the structure of ancestral environment. *Ethology and Sociobiology, 11,* 375-424.

Tourangeau, R. and Ellsworth, P. C. (1979). The role of facial response in the experience of emotion. *Journal of Personality and Social Psychology, 37,* 1519-1531.

Trevarthen, C. (1984). Emotions in infancy: Regulators of contact and relationships with persons. In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 129-157). Hillsdale, N.J.: Erlbaum.

Trevarthen, C., Hubley, P. and Sheeran, L. (1979). Les activités innées du nourrisson. In J.-P. Desportes and A. Vloebergh (Eds.), *La Recherche en éthologie: Les comportements animaux et humains* (pp. 25-53). Paris: Seuil, Collection Points.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quaterly Review of Biology, 46,* 35-57.

Tulving, E. (1985). How many memory systems are there? *American Psychologist, 40,* 385-398.

Turner, J. T. (1985). Factors influencing the development of the hostage identification syndrome. *Political Psychology, 6,* 705-711.

Turner, T. J. and Ortony, A. (1992). Basic emotions: Can conflicting criteria diverge? *Psychological Review, 99,* 566-571.

Vallerand, R. J. and Thill, E. E. (1993). Introduction au concept de motivation. In R. J. Vallerand and E. E. Thill (Eds.), *Introduction à la psychologie de la motivation* (pp. 3-39). Laval, Québec: Éditions Études Vivantes.

Van der Maren, J.-M. (1993). *Méthodes de recherche pour l'éducation.* Montréal: Université de Montréal, Faculté des Sciences de l'Éducation.

Vanderveken, D. (1990). On the unification of speech act theory and formal semantics. In P. R. Cohen, J. Morgan and M. E. Pollack (Eds.), *Intentions in communications* (pp. 401-415). Cambridge, MA: MIT Press.

Van de Velde, W. and Perram, J. W. (Eds.) (1996). *Agents breaking away,* Proceedings of the 7th European Workshop on MAAMAW, Lecture Notes on Artificial Intelligence, No. 1038. Berlin: Springer.

Van Hooff, J. A. R. A. M. (1972). A comparative approach to the phylogeny of laughter and smiling. In R. A. Hinde (Ed.), *Non-verbal communication* (pp. 209-241). Cambridge: Cambridge University Press.

Viau, R. (1994). *La motivation en contexte scolaire.* Saint-Laurent: ERPI.

Vollmer, P. J. (1977). Do mischievous dogs reveal their "guilt"? *Veterinary Medicine Small Animal Clinician, 72,* 1002-1005.

Von Uexküll, J. (1956). *Mondes animaux et monde humain.* Hambourg: Gonthier.

Vosniadou, S. and Ortony, A. (Eds.) (1989). *Similarity and analogical reasoning.* Cambridge: Cambridge University Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge: Harvard University Press

Wallbott, H. G. and Scherer, K. R. (1986). The antecedents of emotional experiences. In K. R. Scherer, H. G. Wallbott and A. B. Summerfield (Eds.), *Experiencing emotion: A cross-cultural study* (pp. 69-83). Cambridge, England: Cambridge University Press.

Wallbott, H. G. and Scherer, K. R. (1995). Cultural determinants in experiencing shame and guilt. In J. P. Tangney and K. W. Fischer (Eds.), *Self-conscious emotions. The psychology of shame, guilt, embarrassment, and pride* (pp. 465-487). New York: The Guilford Press.

Weiner, B. (1982). The emotional consequences of causal attributions. In M. S. Clark and S. T. Fiske (Eds.), *Affect and cognition* (pp. 185-209). Hillsdale, N.J.: Erlbaum.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92,* 548-573.

Weiner, B. and Graham, S. (1984). An attributional approach to emotional development. In C. E. Izard, J. Kagan and R. B. Zajonc (Eds.), *Emotions, cognition and behavior* (pp. 167-191). Cambridge, England: Cambridge University Press.

Weisfeld, G. E. and Beresford, J. M. (1982). Erectness of posture as an indicator of dominance or success in humans. *Motivation and Emotion, 6*(2), 113-131.

White, G. L. (1981). A model of romantic jealousy. *Motivation and Emotion, 5*(4), 295-310.

Wicker, F. W., Payne, G. C. and Morgan, R. D. (1983). Participant descriptions of guilt and shame. *Motivation and Emotion, 7*(1), 25-39.

Wierzbicka, A. (1992). Talking about emotions: Semantics, culture, and cognition. *Cognition and Emotion, 6,* 245-268.

Wilson, E. O. (1975). *Sociobiology: The new synthesis.* Cambridge: Harvard University Press.

Winograd, T. (1970). *Procedures as a representation for data in a computer program for understanding natural language.* Ph. D. Thesis, MIT.

Winograd, T. (1975). Frame representations and the declarative/procedural controversy. In D. G. Bobrow and A. M. Collins, (Eds.), *Representation and understanding: Studies in cognitive science,* (pp. 185-210). New York: Academic Press.

Winograd, T. (1988). Where the action is. *Byte,* December 1988.

Winograd, T. and Flores, F. (1986). *Understanding computers and cognition: A new foundation for design.* New York: Addison Wesley.

Winston, P. H. (1970). *Learning structural descriptions from examples.* Ph. D. Thesis, MIT.

Winston, P. H. (Ed.) (1975). *The psychology of computer vision.* New York: McGraw-Hill.

Winston, P. H. (1978). Learning by creating and justifying transfer frames. *Artificial Intelligence, 10*(2), 147-172.

Winston, P. H. (1984). *Artificial intelligence (Second edition),* Chap. 5: Control metaphors, (pp. 133-157). Reading, Mass.: Addison-Wesley.

Winton, W. M. (1986). The role of facial response in self-reports of emotion: A critique of Laird. *Journal of Personality and Social Psychology, 50,* 808-812.

Wong, P. T. P. and Weiner, B. (1981). When people ask "why" questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology, 40,* 650-663.

Wright, I. (1996). Reinforcement learning and animat emotions. In P. Maes, M. Mataric, J.-A. Meyer, J. Pollack, and S. W. Wilson, (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* (pp. 272-281). Cambridge, MA: MIT Press/Bradford Books.

Young, P. T. (1943). *Emotion in man and animal, its nature and relation to attitude and motive.* New York: John Wiley and Sons.

Young, P. T. (1949). Emotion as disorganized response - A reply to Professor Leeper. *Psychological Review,* 56, 184-191.

Zahn-Waxler, C. and Kochanska, G. (1990). The origins of guilt. In R. Thompson (Ed.), *Nebraska Symposium on Motivation 1990* (Vol. 36, pp. 183-258). Lincoln: University of Nebraska Press.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35,* 151-175.

Zajonc, R. B. (1984a). On the primacy of affect. *American Psychologist, 39,* 117-123.

Zajonc, R. B. (1984b). The interaction of affect and cognition. In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 239-246). Hillsdale, N.J.: Erlbaum.

Zajonc, R. B. and Markus, H. (1984). Affect and cognition: The hard interface. In C. E. Izard, J. Kagan and R. B. Zajonc (Eds.), *Emotions, cognition and behavior* (pp. 73-102). Cambridge, England: Cambridge University Press.

Annexe A

Lettre d'acceptation de l'article présenté dans le *Journal of Artificial Intelligence in Education*

AACE Association for the Advancement of Computing in Education

PO Box 2966 ■ Charlottesville, VA 22902 USA
Phone: 804-973-3987 ■ Fax: 804-978-7449
E-mail: aace@virginia.edu ■ http://aace.virginia.edu/aace

January 7, 1997

Michel Aubé
Université de Sherbrooke
Sherbrooke, Qc, Canada, J1K 2R1

Dear Michel Aubé:

I am pleased to inform you that your article, "Toward Computational Models of Motivation: A Much Needed Foundation for Social Sciences and Education," has been accepted for publication in the *Journal of Artificial Intelligence in Education (JAIE)* and will appear as soon as the publication schedule permits. You will receive an author's copy of the issue upon publication and an article reprint order form.

If your manuscript was created with a microcomputer word processing program, please supply an IBM/MS DOS or Macintosh disk containing the file in your word processor format, preferably *WordPerfect* or *Microsoft Word*, and in an ASCII text file format. This will reduce both publication time and chance for errors. Please inform us if a disk cannot be supplied.

Also, please complete and return the enclosed copyright transfer agreement and notify us if you have a change of address.

Thank you for your interest in and contribution to the *Journal*. We look forward to receiving future article submissions from you.

Sincerely,

Karen L. Whitlock

Karen L. Whitlock
Publications Coordinator

Annexe B

Article soumis et accepté au
CIKM Workshop on Intelligent Information Agents
et message de Tim Finin au sujet de l'article

CIKM Workshop on Intelligent Information Agents

held in conjunction with the

Fourth International Conference on Information and Knowledge Management (CIKM'95)

Friday, December 1, 1995

Baltimore, Maryland

Edited by: Tim Finin and James Mayfield

Sponsored by: ACM SIGART and ACM SIGIR, in cooperation with: NASA, NSF, AAAI, IEEE Computer Society, SIGLINK, CACS/USL and University of Maryland Baltimore County.

CIKM'95 Intelligent Information Agents Workshop

Omni Inner Harbor Hotel, Baltimore MD -- December 1-2, 1995

Organizers' Welcome and Overview

·[CIKM] [IIAW] [Schedule] [papers]

We are happy to welcome you to the second CIKM Workshop on Intelligent Information Agents. The workshop is intended to bring together a group of researchers and developers who are exploring the use of agent-oriented paradigms in information systems. Approximately 50 papers, position paper and extended abstracts were submitted to the program committee for review. Of these, fourteen were selected for presentation and discussion. These papers plus an additional 25 papers are assembled into an on-line proceedings which is available at <http://www.cs.umbc.edu/~cikm/iia/1995/>

We would like to thank the many people who helped plan and organize the workshop. We could not have done this with out their help.

- The program committee consisted of Yigal Arens, Edmund Durfee, Benjamin Grosof, Michael Huhns, Alon Levy, Pattie Maes, Yannis Labrou, Don McKay, Daniella Rus, Ellen Voorhees, and Michael Wooldridge. All we very helpful in the initial planning for the workshop and in reviewing submitted papers.
- Dr. Charles Nicholas, the general chair of CIKM'95 was very supportive and helped solve many of the logistical problems that arose.
- Dr. Carolyn Harriger and her group at the UMBC Office of Continuing Education have done a wonderful job of supporting the entire CIKM conference including this workshop.
- A number of UMBC graduate students helped with organization, reviewing, proceedings and web pages and many other tasks. We would like to especially think Yannis Labrou, R. Scott Cost and Anupama Potluri.
- The workshop, as part of CIKM'95, is sponsored by ACM special interest groups SIGART and SIGIR and is held in cooperation with NASA, Bellcore, NSF, AAAI, IEEE Computer Society, ACM SIGLINK, CACS/USL, and UMBC.

We hope that everyone finds the workshop to be both interesting and worthwhile. If you have any comments or suggestions for future workshops, please let us know.

James Mayfield, (mayfield@umbc.edu)
Tim Finin, (finin@umbc.edu)

CIKM'95 Intelligent Information Agents Workshop

Table of Contents

A FOUNDATION FOR COMMITMENTS AS RESOURCE MANAGEMENT IN MULTI-AGENTS SYSTEMS

Michel Aubé
Université de Sherbrooke,
2500 Blvd Université,
Sherbrooke, Qc, CANADA, J1K 2R1
maube@jrv.qc.ca

Alain Senteni
Université de Montréal,
C. P. 6128, succursale A,
Montréal, Qc, CANADA, H3C 3J7
senteni@iro.umontreal.ca

ABSTRACT

From their earliest definitions, systems of distributed intelligent agents, so-called multi-agent systems (MAS) by the DAI community, relie on sociological theory and use centrally the concept of commitment. A definition of this concept as well as a tentative computational model of commitments, emerge from research in sociology such as Elihu Gerson's and research in computer science such as Alan Bond's. Both threads lead naturally to the study of resource management. We claim that effective resource management in a distributed system requires powerful control structure yet to be founded. We propose a model of control that stems from the psychology of human emotions, introducing "emotion-like control structures" that should be part of the fundamental definition of more autonomous, reliable and adaptive multi-agent systems. We illustrate how the introduction of "emotion-like control structures" could improve a great deal the architecture of a computational model of distributed agendas, using a model of communication rooted in speech act theory.

• INTRODUCTION

Communities of cooperating autonomous agents as well as human societies use resources whose management is at the core of any organized activity. Resources are defined more or less as energy is in physics, whatever is required for producing work : it basically amounts to the cost of work or actions. Within animals or animats, needs, such as hunger and thirst, are the motivational processes that insure management and regulation of directly consumable resources, be it available food or water for animals or processing time for a computer. These are first-order resources. Besides, there exist second-order resources, which one gets only indirectly, through having others committed to provide them : newborns could get access to food only through having their parents provide the feeding resources for them.

Commitments, essentially measured in terms of resource allocation and constraints upon resources appear to be one critical reason for the setting up of multi-agent societies. By regulating exchange, they enable using each other as a way to access additional resources otherwise unaccessible or too costly to obtain. From their earliest definitions, systems of distributed intelligent agents, so-called multi-agent systems (MAS) by the DAI community, relie on sociological theory [Gasser 91] and use centrally the concept of commitments [Fikes 82]. Building upon the seminal work of symbolic interactionists such as [Gerson 76], [Bond 90] and [Dongha 94] propose a definition and a computational model of the concept of commitment that leads naturally to the study of resource management. We think that effective resource management in a distributed system requires powerful control structures yet to be founded.

The main purpose of this paper is to sketch the design for such a model of control that stems from the psychology of human emotions [Oatley 92]. The first section defines first-order and second-order resources, and revisit the concept of commitments in this context. The second section proposes a model of control that stems from the psychology of human emotions. It introduces emotions as powerful control structures that act so as to protect agents - and societies of agents - from running out of resources, typically second-order-ones. The third section takes a closer look at the utilisation of speech acts as communication primitives between agents and introduces the new concept of emotional acts, as a way to resolve ambiguities inherent to the expressive and declarative illocutionary categories. We claim that "emotion-like control structures" should become part of the fundamental definition of more autonomous, reliable and adaptive multi-agent systems.

I • FROM RESOURCES TO COMMITMENTS

I.1 • First-order and second-order resources

If agents are ever to get autonomous, it will only be through getting control over their motivations, by having ways of setting their own goals. The idea of providing agents with the capability to generate their own top level goals is not alien to the DAI community : it can be found in Maes' earlier work [Maes 91] while the concept of *motivated agency* is introduced by [Norman & Long 95]. Motivations - needs and emotions - essentially have to do with managing resources : this is

what is most crucial and vital for the agents, and what becomes their ultimate goals. In the agent context, resources are defined more or less as energy is in physics : whatever enables - and is required for - producing work. It could be physical energy like heat or electricity or food, but also time, money, affiliation, power, knowledge... Basically, the concept seems tantamount to the cost of work or actions, but it is however essential to refine it and distinguish at least two kinds of resources : first-order and second-order.

> • First-order resources are directly consumable ones that an agent already has at disposal (e.g. available food or water for animals, processing time or memory for computer). Needs - such as hunger, thirst, fatigue, shelter, sex... - are the motivational processes [Toates 86] that insure management and regulation of first-order resources.
>
> • Second-order resources are those that you get only indirectly, through having other agents committed to give them to you : newborns could get access to food only through having their parents provide the feeding resources for them. Money as well should be seen as a second-order resource : any piece of money always amounts to a commitment to exchange some goods for it. And one should keep in mind that it is always revocable, be it only through the variation of currency!

I.2 • Commitments as second-order resources

Commitments are essentially measured in terms of allocation of resources and constraints upon resources [Gerson, 76] [Bond 90] [Dongha 94]. They appear as one critical reason for the setting up of multi-agent societies : by regulating exchange, they enable using each other as a way to access additional resources that would remain otherwise unaccessible, or would be too costly to obtain :

> • An agent A commits certain of its resources to another agent B (commits stands for "does not give right away, but promises to give eventually"). As a consequence, B is allocated new resources, while A is constrained as to the usage of its own resources. For instance, A might be constrained not to consumate all of them and to take into account some amount that has to remain available to fulfill the commitment. On the other hand, such commitment could also involve for A setting the goal of acquiring for B some resources that A does not already possess.

One important corollary is that, by virtue of making resources available, commitments themselves are to be counted as resources. Within complex species that pursue multiple goals, and who

absolutely require as such the cooperation of other agents, commitments might well become the most important of all resources! Hence, it is very likely that, in social species like ours, evolutionary forces would have favored and selected out precisely those organisms that would have become equipped with powerful mechanisms to insure control over commitments!

II • EMOTIONS AS A METAPHOR FOR CONTROL STRUCTURES

II.1 • The appraisal structure of emotions : Valence, Certainty, Agency

• Valence : positive and negative emotions In higher (social or at least nurturing) animals, emotions - such as anger, fear, joy, sadness, guilt... - are the motivational processes that insure management and regulation of second-order resources, namely commitments. Very much like needs, emotions are powerful control structures that act so as to protect agents - and societies of agents - from running out of resources. Hence they require being allocated a higher level of processing priority than most other mechanisms. These control processes are triggered by significant variations in the flow of resources available or required for an agent (or a community of agents). There indeed has to be regular variations in the daily consumption of resources, without these overpowerful processes intervening within the ongoing computation. Significant variations here means that the amount of increase or decrease of resources should exceed a certain predefined (although eventually ajustable) threshold.

There obviously has to be two main classes of emotions, positive and negative, that deal respectively with significative gain or loss in resources (leading respectively to joy or sadness).

> • Positive emotions act so as to reinvest the gain (that triggered them in the first place) in order to "buy" access to further resources : they operate as accumulators or amplifiers [Isen 93].
>
> • Negative emotions, on the other hand, are used so as to patch the breach, slow down the running waste, and eventually to call for help so as to replace the incurred loss : they operate as regulators.

Of course, these processes are themselves costly and it might not seem too adaptive to spend so much resources when one is already losing some. Yet, even if one has to pay for the plumber, it

might well be worth to do so before completely running out of water! Sometimes, this protection function of negative emotions against the loss would rather be met by setting the agent in economy mode, i.e. by isolating it from the others and reducing its activity level, as it often happens with people suffering from grief or severe distress.

• Certainty, uncertainty and temporal dynamics On the other hand, emotions being (as motivations) intrinsically related to the goal structure of an agent, they forcefully have to be related also to planning and expectations. That is to say, emotions are necessarily subject to a temporal dynamics, and could then be triggered by anticipated as well as by actual gain or loss (hence leading to hope or fear). While actual gains would be reinvested to obtain further resources, anticipated gains would rather insure goal persistence, by acting so as to justify sustaining the actual investment, or even spending more resources in the pursuit of the current goal. While actual loss would trigger calling for extra resources from other agents, anticipated loss would rather induce significative change in the current planning strategy. In order to do so, the required resources will generally have to be substracted from other current (goal-pursuing) processes that would then get temporarily (and even sometimes definitively) suspended.

• Agency : causal attribution of events to other agents Moreover, in multiagent societies, gains and losses are most frequently caused by the action of agents (self or others). In other words, due to the very structure of commitments as second-order resources, agency is, in those societies, the general cause that a resource management system should"naturally" (i.e. "evolutionarily") get a grip upon!

Before classifying emotions according to their appraisal structure, let's first recall what is agency in the psychology of emotions :

> • Agency, is understood as causal attribution [Weiner 85] : if something important happens to agent A, this one tries to attribute the responsibility to some other agent B. This is a way, for A, to get a grip on who can be held responsible for the considered upcoming event in order to react efficiently.

Anger is the process that gets control over situations when a significant loss of resources is attributable to the action of another agent [Hutchinson 72] [Averill 83] . On the other hand, if the

loss is caused by self, it is the guilt emotion that steps in. [McGraw 87] . When someone else is seen as responsible for some noticeable gain, gratitude is triggered. Finally, pride takes charge when it is self that appears clearly responsible for the gain.

Joy, sadness, hope or fear are triggered either when no attributable agency could be found as a cause for the loss or gain, or when there appears to be no possible control over the corresponding event. From the "point of view" of control structures over commitments, non-controllable agency is tantamount to no agency at all, and there is no point spending resources against overpowerful agents : better call for help then or fly away (and alert relatives to flee as well)! For instance, in cases of anticipated loss attributable to others, anger would be triggered only if there is something to be done against the agents to prevent their wrongdoing, while fear would be triggered if nothing could be envisioned to prevent it.

II.2 • A Taxonomy Of Emotions

Figure 1 summarizes the preceding categorization of emotions. Although much refinement could be introduced within each category (e.g. distinguishing guit from shame), the taxonomy is pretty much congruent with the intersecting aspects of the current "appraisal theories" in the psychology of emotion [Ortony, Clore & Collins 88] [Roseman 91] [Scherer 88] [Smith & Ellsworth 85] .

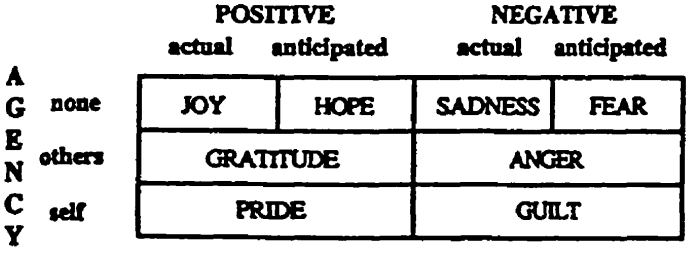


Fig. 1 The appraisal structure for the basic categories of emotions.

The first four emotions (joy, sadness, hope and fear) act so as to set up or generate new commitments (through giving and sharing resources, strengthening affiliation, building up trust and confidence, calling for help and support...) between agents.

It is perhaps mandatory to stress here some theoretical consequence of having restricted emotions to being commitments operators. This means that we would not count as "fear proper" the escaping reactions of lower organisms such as fish, amphibians and most of the reptiles (that happen not to have evolved limbic structures), but as being part of a simpler "need structure". Our basic criterion for counting a motivational process as an emotion, is that it should provide for some communicative means in the service of interagent binding (such as alarm or distress calls, warning growls, tail waving...).

The last four emotions (pride, guilt, gratitude and anger) act so as to handle or prevent the breaking of current commitments between agents (through praising or threatening...) or to strengthen and sustain them. Figure 2 illustrates the model we have sketched so far for the flow of control relative to the management of resources.

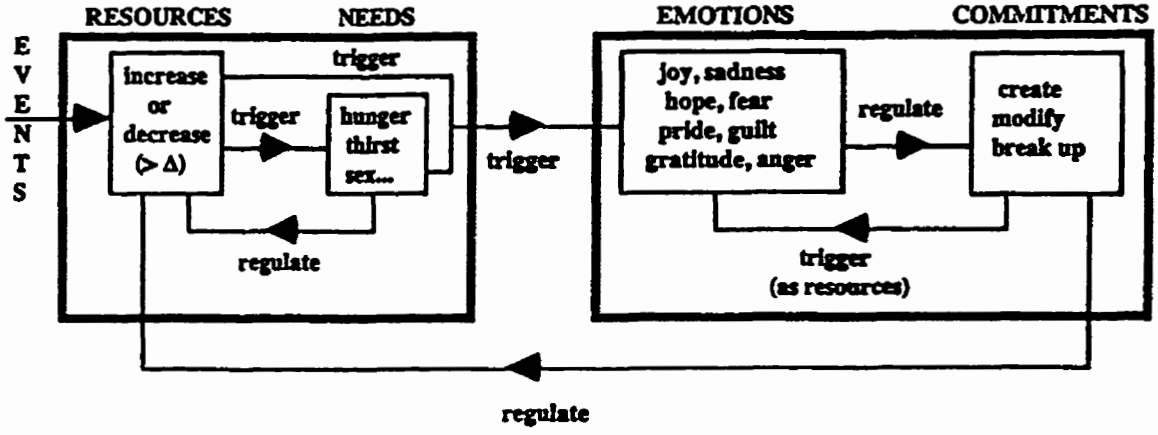


Fig. 2. The flow of control in the management of resources.

It could be seen from this diagram that needs could themselves activate emotions in their attempt to regulate a significant variation of resources, for instance when a hungry baby gets distressed or even angered at not being fed. It also seems to be the case that first-order resources could activate emotions directly, when the variation is large enough to cross the critical threshold so as to "launch" emotional expression "evolutionarily" designed to commit others to set in and react (so they would help or flee or repent...) : states of exhaustion often lead to distress, and sudden intense pain is a well documented releaser of anger [Berkowitz 93] .

III • EMOTIONS IN THE CONTEXT OF AGENT COMMUNICATION

III.1 • Agent communication deals with commitments

In the context of MAS, communication has to be about interagent binding, resource sharing, cooperation or conflict, and then has to rest upon establishing and managing commitments. In other words, communication has to be, in some way, closely related to the emotional control mechanisms. Obviously, all communications will not be emotional : variations in commitments (as resources) do not always exceed the critical threshold (as stated above) beyond which emotions are triggered. But all emotions - as commitment operators - sure have to be communicative (i.e. they have to make use of communication methods)!

III.2 • Speech acts are communication primitives

Both linguistic and DAI communities agree that, in human languages, the basic units of communication are speech acts, i.e. various ways of specifying intentions - and, as a consequence, commitments [Cohen & Perrault 79] [Winograd & Flores 86]. Classically [Searle 79] and later [Vanderveken 90], reformulating [Austin 62] both provide a taxonomy for these communication primitives.

III.3 • A taxonomy of illocutionary speech acts

Speech acts capture categories of statements (typically centered around verbs) as various ways of specifying intentions and commitments [Winograd & Flores 86] among communicating agents, and thereby introduce a taxonomy of communication primitives :

- Assertive commit the speaker to some information being reliable.
- Directive request from someone to commit to accomplish some action.
- Promissive commit oneself to accomplish some action for someone else.
- Declarative commit, and request from others to do as well, to some change in status.
- Expressive express some feelings or emotions.

• Expressive speech acts However there are a few well documented problems with this taxonomy. On the one hand, [Austin 62] and [Searle 79] themselves argue that the expressive

category has indeed a special character ("a very funny one"), and most further formalisms of speech acts bluntly ignored it [Cohen & Perrault 79] [Cohen & Levesque 90] [Singh 94] . We claim that expressive speech acts cluster intentions that should not belong to the same processing level as those from other categories, in particular because they correspond to different processing priorities. We would rather consider the expressive category as emotional acts (vs speech acts). Our contention is that, although this category also has to do with managing and conveying intentions, it simply does not stand on the same level as the others, and rather corresponds, on the verbal level, to a "vestigial" form of a more primitive underneath layer for the handling of commitments in social animals! It amounts to the verbal aspect of a much wider channel of emotional expression (emotional acts) that already flows through various facial, postural, or sound (v.g. crying, yelling, growling, laughing...) behaviors.

> • From speech acts to emotional acts : we contend that agent communication should shift back from speech acts to emotional acts whenever the threshold in the variation of second-order resources is exceeded or crossed over.

Hence, communication between agents should non only reflect the verbal aspect of typical human communication but most importantly emotional non-verbal aspects as well. The choice of a platform for a testbed should take this into account : each script used by an agent for communicating with others would have to embody the specific way - in terms of speech acts or emotional acts - through which it operates upon commitments. One critical point here about the ontology of communicating agents is that they should be hardly able to ignore (or escape from responding to) emotional signals emitted by other agents, so that those signals could, for instance, generate interrupts or reset computation.

• Declarative speech acts On the other hand, the case for declaratives is also peculiar : it looks like a very special device of using some authority (i.e. agents with a high status - like judges in our societies) as a third party or witness to guarantee that a whole community of agents would at once subscribe to (i.e. get bound to) a new commitment (which for instance happens when a new status is ascribed to some agent - "I name you President" - or when some object of collective usage is given a new name - "From now on, this street will be called Kennedy Avenue"). This peculiar kind of commitment, shared by a whole community at once, is closely related to what [Shoham & Tennenholtz 95] termed a useful social law in computational environments.

• Assertive, directive and promissive speech acts It then looks as if classical categories of speech acts did in fact encompass three different layers of communicating tools for setting up and managing commitments between interacting agents. In the middle lay the categories most frequently referred to and formalized within DAI literature and practice, namely assertives, directives and promissives, that deal with common ongoing transactions between agents : exchanging information, requesting and promissing, bidding and contracting...

Underneath lies the more primitive (but also more powerful, controlwise) layer of emotional acts, that sometimes emerge as expressive speech acts when they reach the verbal level. Emotional acts are triggered when there is serious risk of breaking ongoing commitments or unforeseen opportunities of establishing new ones. Serious risk here means that the variation in second-order resources is exceeding some predefined threshold.

Finally, the top layer has to do with declaratives, that are used to set up new social laws : they hence provide a mechanism for establishing and distributing new commitments at once across a whole community of agents. They rest on the essential condition that the emitter's status should be high enough to enforce the new regulation, so that receivers belonging to the same community would as such be constrained to respect it.

IV • THE DISTRIBUTED AGENDAS TESTBED

The choice of a proper testbed for the kind of control structure we have in mind is not obvious. First, the problem has to be complex enough so as to require distributing computation over a small community of agents, yet it has to remain manageable enough so we could draw useful conclusions from its implementation. It also seems interesting that it would be resonant or consonant with the kind of problems inhabiting the DAI culture. Finally, we feel that the nature of the problem should - when raised in human or animal societies - frequently, if not inherently, call for "emotional means of resolution". Distributed meeting scheduling emerges as a rare case satisfying enough with respect to all such criteria.

In the context of meeting scheduling (agendas),

- relevant first-order resources are :
  - periods of time (to be allocated for the various meetings);
  - processing priority;
  - knowledge (about self or others' available periods of time, or constraints, or preferences or commitments...).
- relevant second-order resources are :
  - commitments over time periods reserved for the meetings;
  - reliability of the knowledge exchanged.
- special kinds of commitments have to do with :
  - alliances (or affiliation between agents that would then feel committed to cooperate with each other);
  - status (i.e. allocated roles and power, that could sometimes enable obtaining, or even imposing, consensus among subordinates).

From what we said in the beginning of this article regarding motivations, it should be clear that goals (for instance, setting some specific meeting) are basically about getting resources (and from then on, preserving them) or about setting commitments (and from then on, keeping them). But this does not preclude that they could also bear upon exchanging resources (i.e. give away some resources in order to "buy" other ones that are required for the current activity), especially when this leads to creating new commitments. Plans are means to achieve goals, i.e. sequences of primitives or procedures (methods) owned by the agent, that could be assembled and executed so as to get to the expected results. As such, plans could be counted as part of the knowledge resource. From the above, it is easy to see that an agent's emotions are to be very closely related to its goal structure. For instance, blocking access to a required resource, would generally amount to goal blocking, which sure is one "universal" way (i.e. wide across cultures - and even across species!) of triggering anger [Averill 83] [Hutchinson 72] [Oatley 92] .

In the context of meeting scheduling, there forcefully has to be many agents involved at the same time, each pursuing many different goals, having to keep and fulfill various commitments, sometimes requiring help and additionnal resources from a small set of intervening emotions. In terms of implementation, this means that not only agents should be represented as dynamic entities and operate in parallel, but this should be the case as well for goals, commitments and emotions! That is, all four concepts have to be represented as actors operating concurrently by sending and

receiving messages asynchronously [Agha 86]. An agent does set a goal when asked for a new meeting, or as a subgoal of a current goal (while following a plan). Achieving goals could mainly be done through communicating with other agents and making commitments with them. Of course, some checks and bookkeeping also have to be performed within the agent's internal state, be it only to make sure that the target period is free, or that the request does not interfere with other commitments. Sometimes, meetings could be settled quite easily, when there is no conflict as to the allotted period of time. But with many people from many different groups, there are problems to be solved (and plans to be unfolded) often enough, especially as time goes and as more periods get allocated (hence becoming unavailable for further meetings within the critical time schedule). Then an agent might think of reconsidering a commitment concerning some "minor" meeting, or try make use of alliances and status as resources in convincing others to revoke theirs. This is typically on such occasions that emotions do step in!

In the spirit of this testbed, one should envision that some requests have high enough priority (as when the meeting is about getting a doctor for an urgency" or if the requester owns a much higher status than self) so that reactions such as fear or distress do appear "natural" and practically unavoidable. In order to be able to interrupt ongoing activity, emotions actors should be allocated a higher processing priority than goals actors or commitments actors. Emotions are to be triggered when current goals or commitments go awry.

IV.1 • Implementation Issues

The implementation uses the Actalk actor platform built on top of ST-80 [Briot 94]. It takes profit of the Smalltalk exception handling system in which an exception is defined as a situation when computation could not be completed [Dony 90]. When such a situation can be clearly described, it could also be foreseen and prevented by having a signal watch for its occurrence. Again, as was stated before, proper signals to raise for the triggering of emotions have to do with significant variations in the stock of resources, more precisely those involved within the critical goal or commitment (the one which caused problem). Once a signal is caught, control is transfered to a well-tailored handler that could execute some appropriate piece of code before resuming or halting computation.

V •CONCLUSION

Effective resource management in a distributed system requires powerful control structure yet to be founded. We proposed a model of control, rooted in the psychology of human emotions, that introduces "emotion-like control structures" in the context of multi-agent systems. We pointed out how such control structures fit naturally into an inter-agent communication scheme based on speech act theory. Their introduction improves the potential autonomy of agents by taking explicitly into account the management of the resources that are the core of their activity and that should hence be part of the fundamental definition of more autonomous, reliable and adaptive multi-agent systems.

VI • REFERENCES

[Agha 86] Gul Agha, *Actors*, MIT Press, 1986.

[Austin 62] J.L. Austin. *How to do things with words*, Cambridge, MA: Harvard University Press, 1962.

[Averill 83] James R. Averill. Studies in Anger and Agression: Implications for Theories of Emotion. *American Psychologist*, 38, 1145-1160, 1983.

[Berkowitz 93] Leonard Berkowitz. Pain and Agression: Some Findings and Implications. *Motivation and Emotion*, 17, 277-293, 1993.

[Bond 90] Alan H. Bond. A Computational Model for Organizations of Cooperating Intelligent Agents. *SIGOIS Bulletin*, 11, 21-30, 1990.

[Briot 94] Jean-Pierre Briot. Modélisation et Classification de Langages de Programmation Concurrente à Objets : L'Expérience Actalk, *Actes du colloque "Langages et Modèles à Objets"*, Grenoble, France, 1994.

[Cohen & Levesque 90] Philip R. Cohen and Hector. J. Levesque. Rational Interaction as the Basis for Communication. In Philip R. Cohen, Jerry Morgan & Martha E. Pollack, editors, *Intentions in Communications*, pages 221-255, Cambridge: MIT Press, 1990.

[Cohen & Perrault 79] Philip R. Cohen and C. Raymond Perrault. Elements of a Plan-Based Theory of Speech Acts. *Cognitive Science*, 3, 177-212, 1979.

[Dongha 94] Paul Dongha. Toward a Formal Model of Commitment for Resource Bounded Agents. In Michael J. Wooldridge and Nicholas R. Jennings, editors, *Intelligent Agents, ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, pages 86-101, New York : Springer-Verlag, 1994.

[Dony 90] Christophe Dony. Exception Handling and Object-Oriented Programming: Towards a Synthesis. In Norman Meyrowitz, editor, *OOPSLA ECOOP'90 Proceedings, Ottawa, October 21-25 1990, Sigplan Notices*, 25 (10), 322-330, 1990.

[Fikes 82] Richard E. Fikes. A Commitment-Based Framework for Describing Informal Cooperative Work. *Cognitive Science*, 6, 331-347, 1982.

[Gasser 91] Les Gasser. Social Conceptions of Knowledge and Action: DAI Foundations and Open Systems Semantics. *Artificial Intelligence*, 47, 107-138, 1991.

[Gerson 76] Elihu M. Gerson. On "Quality Of Life". *American Sociological Review*, 41, 793-806, 1976.

[Hutchinson 72] Ronald R. Hutchinson. The Environmental Causes of Aggression. In James K. Coles & Donald D. Jensen, editors, *Nebraska Symposium on Motivation 1972*, pages 155-181, Lincoln: University of Nebraska Press, 1972.

[Isen 93] Alice M. Isen. Positive Affect and Decision Making. In M. Lewis & J. M. Haviland, editors, *Hanbook of Emotion*, pages 261-277, New York: The Guilford Press, 1993.

[Maes 91] Pattie Maes. Situated Agents Can Have Goals. In Pattie Maes, editor, *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*, pages 49-70, Cambridge: MIT Press, 1991.

[McGraw 87] Kathleen M. McGraw. Guilt Following Transgression: An Attribution of Responsibility Approach. *Journal of Personality and Social Psychology*, 53, 247-256, 1987.

[Norman & Long 95] Timothy J. Norman and Derek Long. *Alarms: An Implementation of Motivated Agency*. Research Note RN/95/24, Department of Computer Science, University College London, London, UK, April 6, 1995.

[Oatley 92] Keith Oatley. *Best Laid Schemes: The Psychology of Emotions*. Cambridge, England: Cambridge University Press, 1992.

[Ortony, Clore & Collins 88] Andrew Ortony, G. L. Clore & Allan Collins. *The Cognitive Structure of Emotions*. Cambridge, England: Cambridge University Press, 1988.

[Roseman 91] Ira J. Roseman. Appraisal Determinants of Discrete Emotions. *Cognition and Emotion*, 5, 161-200, 1991.

[Scherer 88] Klaus R. Scherer. Criteria for Emotion-Antecedent Appraisal: A Review. In V. Hamilton, G. H. Bower & N. H. Frijda, editors, *Cognitive Perspectives on Emotion and Motivation*, pages 89-126, Dordrecht: Kluwer, 1988.

[Searle 79] John R. Searle. *Speech Acts*. Cambridge, England: Cambridge University Press, 1969.

[Shoham & Tennenholtz 95] Yoav Shoham & Moshe Tennenholtz. On Social Laws for Artificial Agent Societies : Off-Line Design. *Artificial Intelligence*, 73, 231-252, 1995.

[Singh 94] Munindar P. *Singh. Multiagent Systems. A Theoretical Framework for Intentions, Know-How, and Communications*. New York : Springer-Verlag, 1994.

[Smith & Ellsworth 85] Craig A. Smith & Phoebe C. Ellsworth. Patterns of Cognitive Appraisal in Emotion. *Journal of Personality and Social Psychology*, 48, 813-838, 1985.

[Toates 86] Frederick Toates. *Motivational Systems*. Cambridge, Cambridge University Press, 1986.

[Vanderveken 90] Daniel Vanderveken. On the Unification of Speech Act Theory and Formal Semantics. In Philip R. Cohen, Jerry Morgan & Martha E. Pollack, editors, *Intentions in Communications*, pages195-220, Cambridge: MIT Press, 1990.

[Weiner 85] Bernard Weiner. An Attributional Theory of Achievement Motivation and Emotion. *Psychological Review*, 92, 548-573, 1985.

[Winograd & Flores 86] Terry Winograd & Fernando Flores. *Understanding Computers and Cognition*, New York: Addison Wesley, 1986.

Re: Request for information : Emotions and Agents

Timothy Finin (*finin@cs.umbc.edu*)
Wed, 8 May 1996 19:19:19 -0700

- **Messages sorted by:** [date][thread][subject][author]
- **Next message:** Marc Eisenstadt: "ANNOUNCE: Don Norman (Apple) LIVE! 15-May 5PM UK = noon EDT"
- **Previous message:** eneeter@mic.atr.co.jp: "Request for information : Emotions and Agents"
- **Maybe in reply to:** eneeter@mic.atr.co.jp: "Request for information : Emotions and Agents"

eneeter@mic.atr.co.jp writes:
> *I would like to ask if anyone have information (ideas, papers,*
> *examples, WWW/FTP sites) about the role of the emotions in intelligent*
> *agents. I am focusing in how could the agents handle different muds.*

See "A Foundation for Commitments as Resource Management in Multi-Agent Systems". Michel Aube, Universite de Sherbrooke, Alain Senteni, Universite de Montreal. PostScript version available as "http://www.cs.umbc.edu/~cikm/iia/submitted/viewing/aube.ps".

Professor Tim Finin, Computer Science & Electrical Engineering, University of Maryland Baltimore County, Baltimore MD 21228. http://umbc.edu/~finin/
finin@umbc.edu, VOICE: 410-455-3522, FAX: -3969, SEC:Angie Silanskis -2732

- **Next message:** Marc Eisenstadt: "ANNOUNCE: Don Norman (Apple) LIVE! 15-May 5PM UK = noon EDT"
- **Previous message:** eneeter@mic.atr.co.jp: "Request for information : Emotions and Agents"
- **Maybe in reply to:** eneeter@mic.atr.co.jp: "Request for information : Emotions and Agents"

Annexe C

Conditions de sélection des articles soumis à la *Fourth International Conference on Simulation of Adaptive Behavior (SAB'96)*

FROM ANIMALS TO ANIMATS 4

Proceedings of the Fourth International Conference
on Simulation of Adaptive Behavior

edited by
Pattie Maes, Maja J. Mataric, Jean-Arcady Meyer, Jordan Pollack,
and Stewart W. Wilson

A Bradford Book

The MIT Press
Cambridge, Massachusetts
London, England

This work relates to Department of Navy grant N00014-96-1-0475 issued by the Office of Naval Research. The United States government has a royalty-free license throughout the world in all copyrightable material contained within.

Printed and bound in the United States of America.

This publication may be ordered from The MIT Press, 55 Hayward Street, Cambridge, Massachusetts 02142 USA; by telephone: 1-800-356-0343 (toll-free) or 617-625-8569; by fax: 617-625-6660; by e-mail: mitpress-orders@mit.edu or www-mitpress.mit.edu.

ISSN 1089-4365
ISBN 0-262-63178-4

PREFACE

The papers in this book were presented at the Fourth International Conference on Simulation of Adaptive Behavior (SAB96), held at North Falmouth, Cape Cod, Massachusetts, USA, on September 9-13, 1996. The objective of the biennial SAB conferences is to bring together researchers from a wide range of backgrounds including ethology, theoretical biology, psychology, artificial life, machine learning and robotics, to further our understanding of the behaviors and underlying mechanisms that allow natural and artificial animals ('animats') to adapt and survive in uncertain environments.

Adaptive Behavior research is distinguished by its focus on the modeling and creation of complete animal-like systems, which—however simple at the moment—may be one of the best routes to understanding intelligence in natural and artificial systems. The field received initial recognition on the occasion of the first SAB conference, which was held in Paris in September, 1990. Subsequent SAB conferences (Hawaii, December 1992 and Brighton, England, August 1994) drew increasing numbers of papers and participants. In 1992, The MIT Press introduced the quarterly journal *Adaptive Behavior*, and The International Society for Adaptive Behavior (ISAB) was established in 1995—both events further marking the emergence of Adaptive Behavior as a full-fledged scientific discipline. The present proceedings are a comprehensive and up-to-date resource for the latest progress in this exciting field.

The 66 papers presented at the conference and published here were selected from 150 submissions after a two-pass review process designed to ensure high and consistent overall quality. The authors focus on well-defined models, computer simulations, and robotics demonstrations that help characterize and compare various organizational principles or architectures capable of inducing adaptive behavior in real animals or synthetic agents. The papers are ordered according to the scale at which adaptive behavior takes place, ranging from immediate adaptation in sensorimotor control, to learning within an animat's lifetime, to adaptive behavior exhibited by successive generations of animats, and finally to adaptive collective behavior of animats in groups.

In addition to the presentations of accepted papers, a number of distinguished researchers gave keynote lectures. The SAB96 Keynote Lecturers were:

James S. Albus, Head of the Intelligent Systems Division at the National Institute of Standards and Technology, pioneer in robotics and neural networks, and author of *Brains, Behavior, and Robotics*, whose paper "The Engineering of Mind" is included in this volume;

Jelle Atema, Professor of Biology at Boston University and the director of the Boston University Marine Program at Woods Hole Marine Biology Laboratory;

Daniel Dennett, Distinguished Arts and Sciences Professor and Director of the Center for Cognitive Studies at Tufts University, the author of *Consciousness Explained* and *Darwin's Dangerous Idea*;

C. R. Gallistel, Professor of Psychology at UCLA and the author of *The Organization of Learning* and *Animal Cognition*;

J. A. Scott Kelso, Director of the Center for Complex Systems at Florida Atlantic University and author of *Dynamic Patterns: The Self-Organization of Brain and Behavior*;

David Touretzky, Senior Research Scientist in the Computer Science Department and the Center for the Neural Basis of Cognition at Carnegie Mellon University, author of *The Mathematics of Inheritance Systems*.

The conference and its proceedings could not have happened without the substantial help of a wide range of people. First and foremost, we would like to thank the members of the Program Committee. They thoughtfully reviewed all the submissions and worked with many of the authors to ensure a high standard in the accepted papers. The Committee members were:

Peter Angeline, USA
Randall Beer, USA
Lashon Booker, USA
Thomas Collett, UK
Jacques Ferber, France
Simon Giszter, USA
Inman Harvey, UK
Phil Husbands, UK
Harry Klopf, USA
David McFarland, UK
Geoffrey Miller, Germany
Alan Schultz, USA
Tim Smithers, Spain
Luc Steels, Belgium
Frederick Toates, UK
Saburo Tsuji, Japan
David Waltz, USA
Ronald Arkin, USA
Bruce Blumberg, USA
Dave Cliff, UK
Holk Cruse, Germany
Dario Floreano, Italy
John Hallam, UK
Ian Horswill, USA
Leslie Pack Kaelbing, USA
Michael Littman, USA
José del R. Millán, Italy
Rolf Pfeifer, Switzerland
Jean-Jacques Slotine, USA
Emmet Spier, UK
Lynn Andrea Stein, USA
Peter Todd, Germany
Patrick Tufts, USA

Professor Herbert Roitblat of the University of Hawaii contributed his wisdom and experience with previous SAB conferences throughout the planning and preparation for SAB96. We are very grateful to him.

We are indebted to our sponsors,

The Office of Naval Research for generous financial support;
The M. R. Bauer Foundation for its support of academic activities at the Volen Center for Complex Systems of Brandeis University;
The International Society for Adaptive Behavior for early support and sponsorship.

We are grateful to Dr. Thomas McKenna of ONR for his advice and guidance.

The enthusiasm and hard work of numerous individuals was essential to the conference's success. We thank Christie Davidson of the MIT Media Lab for administration of the program and review process. We are grateful to the Brandeis Computer Science Department, especially Myrna Fox, Jeanne DeBaie, and the student staff of the Department office, for their work in the local, promotional, and financial aspects of the meeting. Alan Danziger, Jeremy Gilbert, and Aryeh Primus provided indispensable World-Wide-Web and computer support.

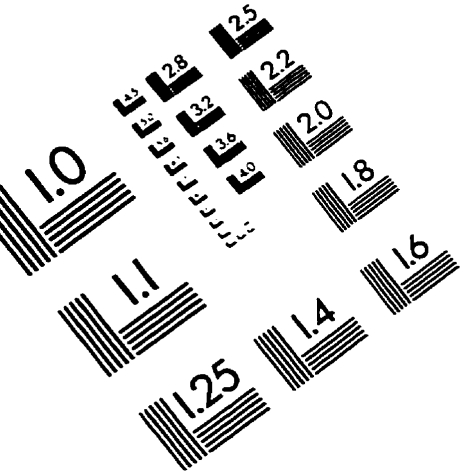
Finally, we are once again especially indebted to Jean Solé for the artistic conception of the SAB96 poster and the proceedings cover.

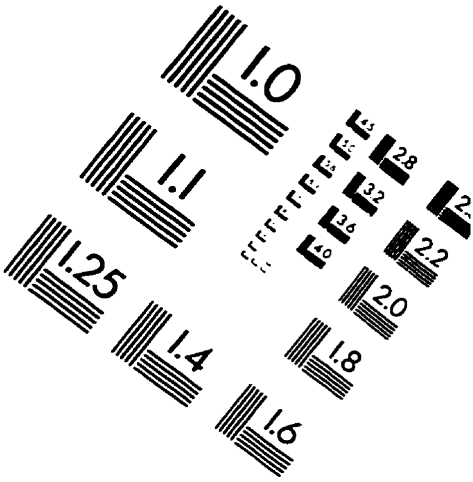
We invite readers to enjoy and profit from the papers in this book, and look forward to the next conference, SAB98.

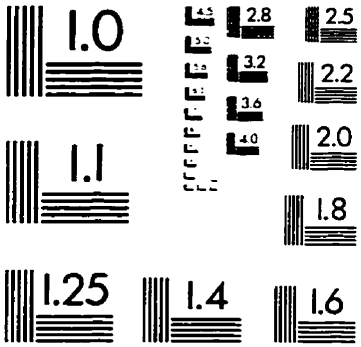
Pattie Maes, Program Chair

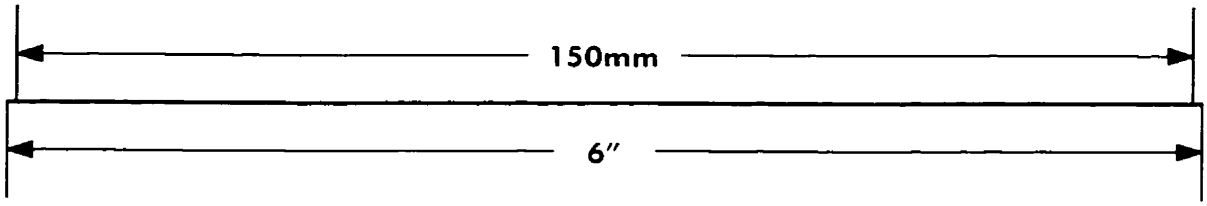
Maja Mataric Jean-Arcady Meyer Jordan Pollack Stewart W. Wilson

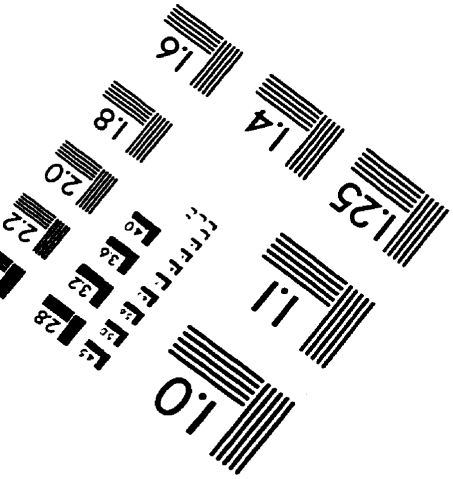
IMAGE EVALUATION TEST TARGET (QA-3)

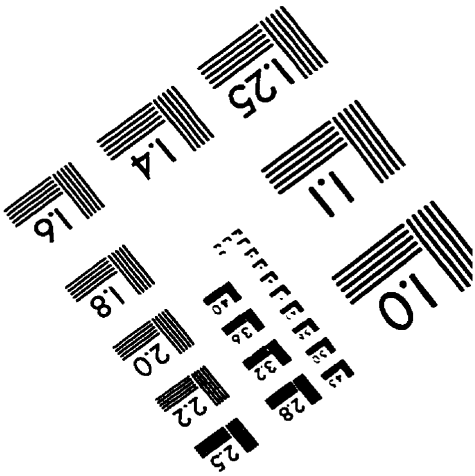












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989