In Proceedings IJCAI 1981, Vancouver
(Terminology now out of date: "intensity" replaced by "insistence")
(Cognitive Science Research Paper 176, Sussex University)

WHY ROBOTS WILL HAVE EMOTIONS

Aaron Sloman and Monica Croucher
Cognitive Studies Programme
(Now School of Cognitive and Computing Sciences)
University of Sussex

NOTE: Aaron Sloman is now at:
School of Computer Science, The University of Birmingham,
Edgbaston, Birmingham, B15 2TT, England
Email: A.Sloman@cs.bham.ac.uk

ABSTRACT

Emotions involve complex processes produced by interactions between motives, beliefs, percepts, etc. E.g. real or imagined fulfilment or violation of a motive, or triggering of a 'motive-generator', can disturb processes produced by other motives. To understand emotions, therefore, we need to understand motives and the types of processes they can produce. This leads to a study of the global architecture of a mind. Some constraints on the evolution of minds are disussed. Types of motives and the processes they generate are sketched.

INTRODUCTION

We all know a lot about the differences and similarities between states like anger, embarrassment, elation, dismay, etc., but it is not easy to *articulate* this knowledge. The task of making such knowledge explicit could be called (following Heider [2]) *Naive Psychology.* This paper makes some steps towards Naive Psychology, and partly goes beyond it, showing that:

> the need to cope with a changing and partly unpredictable world makes it very likely that any intelligent system with multiple motives and limited powers will have emotions.

So the belief that emotions and intellect are somehow quite separate is mistaken.

To constrain our search for a model of mental processes we recommend a multidisciplinary approach combining (a) conceptual analysis, used by philosophers to articulate common implicit knowledge (see [3] ch 4), (b) analysis of *constraints* within which organisms have to operate, and which determine what would make them well adapted, (c) a survey of abilities of different animals and (d) the design of possible systems. This should yield a *grammar* of possible types of minds, natural and artificial. We need to explore types of environmental constraints, types of needs or motives organisms or robots may have, types of information-processing mechanisms and strategies relevant to achieving these motives within different sorts of constraints.

There is a huge array of possible cases, from systems with very simple feedback loops to those containing all the complexity sketched below. Arguing about where to draw the line between cases of real intelligence or mentality and the rest is quite pointless, like arguing over whether it is still 'really' chess if one player accepts the handicap of playing without a queen.

(In the end, some of the decisions are ethical.)

CONSTRAINTS ON INTELLIGENT SYSTEMS

Here are some "constraints" on the design of intelligent systems. Some involve physical needs, some mental needs (the need for powerful inference strategies, etc), some the needs of a social system, and some the needs of the comparatively helpless young.

(1) Goals, obstacles, opportunities, friends and foes do not come with simple physical patterns to identify them. Recognition will often require the use of structural descriptions. Motives will need to incorporate structural descriptions of states to be achieved, preserved, avoided, etc.

(2) The collection of motives will not be static. The physical needs of an organism can vary. Old motives can generate new ones as subgoals for achieving some task. Mechanisms for changing the current set of motives and their priorities are required. The revision of goals makes it necessary for the system to be able to interrupt and re-direct processing.

(3) The environment is not static: opportunities and dangers, may all vary from time to time. Changes will need to be perceived or predicted. Predictions will not always be reliable. This implies a need for constant monitoring, and ability to notice and deal with the unexpected.

(4) Speed of computation may often be important. Events may happen quickly. The organism may have to decide and move quickly. Often there is no time for normal processes of inference, and deliberation.

(5) Some opportunities exist only during a relatively short time interval, often long before they are needed. The ability to store things is therefore important. Some subgoals may have to be achieved long in advance of the main goal. There may be intense, though less important, motives which conflict with a long term goal. It is necessary to be able to *interleave* pursuit of different intentions. This implies an ability to notice conditions for continuation.

(6) The complexity of the environment will often lead to mistaken beliefs, plans, and actions. Later the mistaken belief may be corrected. It is then necessary to modify plans or actions, accordingly. This requires storage of *reasons* for motives and decisions, so that when the new information turns up, chains of relevance can be identified.

(7) If the organism or machine has a complex body, it will need to have many monitors capable of detecting shortages, disturbances, etc., and either causing automatic corrective action or causing a new motive to be created, possibly with a high priority.

(8) Making full use of different parts of the body requires facilities for different processes to exist in parallel. This considerably complicates the problems of interrupting and re-directing thoughts and actions.

(9) Different motives in the same individual may be inconsistent. Mechanisms and strategies for dealing with inconsistencies will be needed.

(10) Where the actions or attitudes of other agents matter it will be necessary to have the ability to represent the mental states of others. This plays an important role in many human emotions, such as embarrassment.

(11) Individuals may perish, and new ones be produced. Transmission of information to offspring requires mechanisms and strategies for such communication and motives which ensure that they are used. The task of communicating skills and knowledge to the young may be a considerable burden to the older individuals, and conflict with many of their motives.

(12) In co-operative communities, individuals should develop motives which do not necessarily maximise their own advantage, but which enable the community as a whole to function well. This can, of course, lead to conflicting motives within and between individuals.

The need to learn motives (including tastes, aesthetic and ethical principles, standards of behaviour, etc.) could apply to machines capable of functioning, in collaboration with human beings, in a wide variety of cultures.

(13) The young will need motives concerned with exploring, practising, noting information, and stimulating older individuals to communicate with them.

(14) Choosing between alternatives will not be simple. The notion of an *optimal choice* will not necessarily even be well-defined. Achieving a long term balance between different needs of the individual or the community can be a major problem. Decision-making processes will have to be capable of coping with such conflicts.

Such constraints define a set of questions to be asked about organisms, and a set of possible tasks for the robot designer.

If individuals contain such complex collections of motives and abilities, with delicately balanced conflicts between motives, they are likely to be very unstable, with relatively small external changes producing new motives decisions and actions, which in turn can lead to further internal changes. Thus initially similar individuals can develop in very different ways.

THE COMPUTATIONAL ARCHITECTURE OF A MIND

Chapters 6 to 10 of [3] sketched some of the computational architecture of the human mind. It turned out useful to distinguish at least the following:

* A store of 'springs of action' (motives).
* A store of resources (including facts, procedures, etc.).
* Indexes to resources.
* A collection of concurrent on-going goal-directed processes.
* Many stores of temporary information, local to the various processes.
* Indexes to current processes and their generating purposes.
* Many kinds of monitoring processes, perceiving external and internal objects and events, and able to cause interruptions.
* A 'central' administrative process concerned with deciding: forming intentions, making or selecting plans and resolving major conflicts in the light of motives (tastes, principles, etc.). Since not all decisions can be taken independently, and sub-processes will sometimes generate incompatible goals, it will be necessary for conflicts to be resolved at a high level in the light of major goals, policies, etc.

These ideas need to be refined. The rest of this paper fills in some details.

PROCESSING MOTIVES

Motives include representations of states of affairs or events to be achieved, preserved, prevented, etc. There are very many different sorts of motives, and the next section will enlarge on this. Similarly, there are many sorts of internal *processes* involving motives, such as:

(1) Adding motives to and removing them from the store.

(2) Selecting motives to act on (forming intentions):- Not all motives can be acted on. Those which are selected will be described as *operative.* Those which are rejected, are *inoperative.* Some may be left for further consideration later, i.e. *suspended.* The word 'intention' covers operative motives. Desires may be operative or inoperative. Even strong desires can be over-ridden by other motives. Motives may be suspended because there is not yet enough information, or because measures of importance, or rules for comparing, fail to yield a clear cut decision.

(3) Drawing attention to suspended or inoperative motives:- Suspended and inoperative motives may have to be reconsidered. Some may be assigned a monitoring process which decides when to interrupt the administrator. Others may constantly request attention. The power to request attention we call the *intensity* of a motive.

(4) Changing the intensity of motives:- Intensity of a motive is a measure of how hard it attempts to get itself selected. We contrast intensity with importance. Intensity is the power to be considered for selection, whereas *importance,* or strength, is the power to be selected for action, and to over-ride alternative motives. Sometimes the importance as well as intensity will change, in which case an inoperative motive may be reconsidered and selected for action.

(5) Abandoning an intention:- An operative motive may revert to an inoperative state. Goals which have been achieved can be removed from the store.

(6) Deciding when to act on operative motives:- Motives will need to be temporally ordered, or even interleaved. Operative motives not currently being acted on we call *dormant,* the others *active.* Some dormant motives are very general, and when the relevant circumstances turn up produce more specific intentions. Hence a motive may be a *motive generator.* (Most of the time you are not in the process of trying to do anything to help people in distress, yet if you observe someone have an accident, a dormant but operative general motive may generate a new specific active and operative intention to help.) Sometimes a motive will be rendered dormant because some other motive has been assigned higher priority. Dormant motives may require some kind of monitoring process.

(7) Waking up dormant motives.

(8) Making or choosing plans:- Sometimes alternative plans will be available, and relevant motives may have to be invoked for choosing between them. So some motives need to take the form of *preferences.* (The planning process can be made more efficient by letting pre-compiled preferences control the process of construction so that instead of being rejected after construction bad alternatives are not even constructed.)

(9) Executing plans:- This may involve adding new subgoals to the store, removing subgoals when they have been achieved, monitoring progress, etc. In some cases the plan can be executed by a lower-level process, which allows the central administrator to carry on with more difficult tasks. This includes both automatic processes like breathing and processes for which fully debugged and reliable strategies are available, like walking.

(10) Retrospective analysis:- Failures sometimes provide information from which the organism can learn. So can unexpected successes. Even actions which go as expected can provide new information revealing redundancy, missed opportunities, risks, etc.

(11) Inductive learning:- If doing something is consistently followed by satisfaction of existing motives, then it might be useful to develop a new motive to do that thing whenever possible, or increase the intensity or strength of an existing motive (or motive-generator). Similarly, if a type of action or situation is frequently followed by the frustration of other motives. This sort of mechanism would explain some results of behaviourist research.

(12) Interruptions:- Inoperative or suspended motives can interrupt the process which selects motives for action, if something makes their intensity high enough. A new motive becoming operative, or detection of conditions relevant to operative but dormant motives can interrupt the processes of planning or execution, as can detection of new circumstances relevant to the current action. Some interrupts may be capable of being handled by a lower-level specialist subsystem, e.g. reflexes. Other interrupts require global re-organisation. Some interrupts may involve matters of such urgency and importance that the new motive bypasses normal selection and planning processes and directly causes a strategy to be invoked.

(13) Suppressing interruptions:- When something which is very important and requires full attention is in progress, then it may be necessary to prevent interruptions. This could be handled by priority mechanisms. For instance, people who are injured in battle sometimes don't notice the injury till long afterwards. The inability of relatively intense desires, pains, etc. to get any attention at all during the performance of some actions, might seem to imply that not all the processes run in parallel on independent processors. Alternatively, it may be the case that what we are conscious of is restricted to information accessed by some one process, all of whose resources are reserved for certain tasks.

We now have a framework to account for different sorts of motives. We see a need for many parallel processes, monitoring internal and external events for significant occurrences. These may trigger (or re-awaken) processes, which may involve interrupting others.

TYPES OF MOTIVES

Motives are representations used in deciding what to do, i.e. desires, wishes, tastes, preferences, ideals, and so on. This includes 'motive generators', and 'motive comparators'. What makes a representation function as a motive is not where it is stored or what its structure is, but its role in processing. There are many different sorts of motives:

(1) Motives may have any of the logical complexity of factual statements. including negation, disjunction, conditionals, universal quantification, and so on.

(2) Some motives constitute desires or preferences, unlike mere sub-goals of other motives. Not all goals are valued intrinsically. Sometimes, in people, a mere subgoal comes to be valued as an end, perhaps because 'reason' information gets lost somehow?

(3) First-order motives directly specify goals, while second-order motives are concerned with generating new motives or deciding relative priorities of conflicting motives: *motive generators* and *motive comparators.* A motive produced by a motive generator may have the status of a desire.

(4) Motives may vary in *intensity.* Intensity can be thought of as an interrupt priority level. The process interrupted is the administrative process concerned with choosing motives to act on. This may also cause other processes to be disturbed.

(5) Some motives may be genetically programmed, such as a motive generator which produces desires to find things out: a 'curiosity instinct'. Others are generated by cognitive processes, such as the interaction of existing motive-generators with new perceptions. Others are generated by bodily processes.

(6) Some motives, e.g. some pains, do not contain specific goals to be achieved, yet indicate a need for something to be done, leaving it to other processes to determine what the problem is. The motive may include information about the *location* or *seriousness,* even if the precise nature or cure is left unspecified.

(7) Some motives are concerned with preserving a state or process. These may involve physical sensations, physical or mental activities (e.g. sex, eating or drinking, developing some skill). 'Pleasure' and 'enjoyment' characterise this sort of thing.

(8) Pleasure and displeasure may be associated with success and failure of actions. The desire for the situation to be different from what it is, is at least part of what is involved in finding the failure unpleasant. The desire to dwell on a success is part of gaining pleasure from success.

(9) Motivation to persevere in the face of failure is often important since repeated actions don't always produce the same result. However, sometimes flexibility is preferable.

(10) Inability to decide between alternatives can lead to disasters. It may be important to be able to detect when deciding is taking too long and generate a motive to speed up the decision-making process.

(11) Some kind of priority system will be needed. Different motives will have different degrees of *urgency* associated with them, concerned with how much time is left. This is not the same as intensity: something not wanted very much may be urgent. In addition there may be a measure of *importance* of the motive. This is different from both urgency and intensity: e.g. important but unattractive duties. Importance may be linked to effects of not achieving the motive.

(12) Besides *measures,* there may be a collection of heuristics for choosing when conflicts arise: motive-comparators

It is difficult to specify what the task of a decision-maker is, when so many complex motives interact without any well defined concept of what constitutes the best decision. Some cases may be dealt with by very general strategies, e.g. choosing at random, or choosing the urgent one. Even where importance is represented by a measure, this may be very coarsely quantised, leaving a need for additional choice heuristics. There may be several incommensurable measures. The notion of an optimal decision, and therefore of an optimal decision-making system is probably not well-defined. An entirely rational decision making process is therefore not to be expected.

Many people deny that machines could ever be said to have their own motives. Machines hitherto familiar to us either are not goal-directed at all (clocks, etc.) or else, like current game-playing computer programs, have a simple hierarchical set of goals, with the highest-level goal put there by a programmer. If machines were designed with a system of motives and motive generators as complex as that described above, then the machine could develop and change over time in such a way that it would be misleading to say that the machine was pursuing the goals of its designer. Ultimately the decision whether to say such machines have motives is a *moral* decision, concerned with how we ought to treat them.

ANGER

Consider a familiar emotion, anger. Typically if some individual X is angry, there is another individual Y with whom X is angry, and something X believes Y to have done or failed to do. Y is believed to be responsible for violating one of X's motives, and X wants to hurt or harm Y: a new motive in X directed against Y. This motive need not be selected for action: it may remain intense yet inoperative, e.g. for fear of consequences. The existence of the desire is still not sufficient for X's state to be one of anger. He may have the desire, but put it out of mind, and calmly get on with something else, and in that case he would not be angry. Anger involves an *intense* desire to do something to Y, that is, the desire should frequently 'request attention' from X's decision-making processes. So unless acted on, the desire will frequently come back into X's thoughts, making it hard for him to concentrate on other activities. (Unconscious anger is possible too.)

It is possible, in human beings, for the anger to produce physical disturbances. However, if X satisfied enough cognitive conditions he could rightly describe himself as being very angry, despite not having the physical symptoms. The anger could be strong, insofar as it constantly intruded into his thoughts and decisions, and insofar as he strongly desired to make Y suffer, and suffer a great deal.

Strength of an emotion can vary along different dimensions: anger can vary according to how much X minds what he thinks Y has done, which will depend on how important the violated motive was. It can vary according to how much harm X wants to do to Y. It can vary also with how important the wish to harm Y is: the desire may or may not be hard to override. Finally, the strength of the anger can vary according to how much general disturbance it produces, or tends to produce in X. Emotion need not actually interfere with other motives: for instance if the new motive to punish Y is acted on, there need be no disturbance. But the anger has the *potential* to disturb other activities if they are attempted.

When there is no desire to cause harm to Y, the emotion is more like exasperation than anger. If there is no attribution of responsibility, then the emotion is some form of annoyance, and if the motive that is violated is very important, and cannot readily be satisfied by some alternative, then the emotion is dismay.

TOWARDS A GRAMMAR OF EMOTIONS

We can now generalise towards a list of components of a grammar of emotional states.

(1) An emotional state normally involves having at least one fairly strong motive. Real or imagined or expected satisfaction or violation of this motive produces the emotion. This generates several sorts of cases, depending on whether the motive is concerned with something strongly desired, or something strongly disliked, whether the desire is thought to be satisfied or violated, or whether there is uncertainty about which is the case.

(2) The combination of motive and belief (or uncertainty) must be capable of producing a *disturbance,* i.e. continually interrupting thinking and deciding, and influencing one's decision-making criteria and perceptions.

(3) The disturbance may or may not involve specific new motives, for instance a desire to right what has gone wrong, or a desire to inform other people. Where there is no new motive: there may simply be a constant dwelling on what has happened.

(4) The new motives need not be selected for action. Some emotional states such as fright may involve the direct production of actions, by-passing the processes of deliberation and planning. A violated motive may be able directly to activate some existing strategy, interrupting other actions. This would include cases of 'impulsive' action. This could make it possible to take very rapid remedial action in times of great danger, or when sudden opportunities are recognized.

(5) Some emotional states arise out of actions performed by the individual, for instance fear about possible errors, and secondary motives may be generated to take extra care, etc. These secondary motives may generate so much disturbance that they lead to disaster.

(6) In some cases, an emotion involves interrupting and redirecting a large number of ongoing processes, for instance processes controlling different parts of the body. Sensory detectors may record local changes produced by the interruptions, and the system's perception of its own state will be changed. However, this kind of experience is not a necessary part of all emotions. Internal monitoring need not produce recognition. The ability to discriminate and recognise complex internal states may have to be learnt, and may involve perceptual processes no less complex than recognising a face or a typewriter.

(7) One need not be conscious of, or feel, the emotion: the disturbance, or tendency to disturb need not be recognised, e.g. because relevant schemata have not been learnt (see [3] ch 10). If it is recognised, this may activate further dormant motives or motive-generators, and possibly lead to a second-order emotion (recursive escalation).

Our conjecture is that the interruptions, disturbances and departures from rationality which characterise emotions are a natural consequence of the sorts of mechanisms required by the constraints on the design of intelligent systems listed earlier. Our theory does nothing, as yet, to characterise the detailed phenomenology of the experiences involved in emotions. These are as rich and varied as other perceptions, e.g. vision, where there is also much work to be done yet. The fine structure of human emotional experience often includes awareness and interpretation of bodily changes. A full account of what it is typically like to *feel* anger, elation, fear, etc. would have to include descriptions of this fine structure. Yet what makes many emotions important in our lives is not this sort of detail, but the more global structure. And that is what we have tried to describe. This is what would make it possible for us to use terms like 'angry', 'afraid', 'disappointed', 'embarrased', 'ecstatic' to describe the state of mind of an alien being, or possibly a sufficiently sophisticated robot.

MOODS AND ATTITUDES ARE NOT EMOTIONS

A *mood* is partly like an emotion in that it involves some kind of global disturbance of, or disposition to disturb, mental processes. But it need not be the intrusion of specific thoughts, desires and inclinations to act, etc. In humans, moods can be induced by chemical processes, or by cognitive processes, for instance hearing good or bad news. A mood can colour the way one perceives things, interprets the actions of others, predicts the outcome of actions, makes plans, etc. As with an emotional state, a mood may or may not be perceived and classified by the individual concerned.

Attitudes are often confused with emotions. It is possible to love, pity, admire, or hate someone without being at all emotional about it. The attitude will be expressed in tendencies to take certain decisions rather than others *when the opportunity arises,* but there need not be any continual disturbance of thoughts and decisions. A mother may love her children without their being constantly in her mind, though a specific occurrence, such as news of danger to them may well interact with this attitude to produce an emotion, such as anxiety. There is no sharp dividing line between attitudes or moods and emotions. The space of possible mental states and processes is too rich and complex for simple divisions to be useful.

There are many kinds of experiences which can be deep and in some sense moving, and which we may describe as emotions, for lack of a richer, more fine-grained vocabulary: for instance delight in a landscape, reading poetry, hearing music, being absorbed in a film or a problem. These seem to involve processes in which what is currently perceived interacts powerfully with a large number of processes, sometimes physical as well as mental. For instance, listening to music can produce a tendency to move physically in time to the music and also a great deal of mental 'movement': memories, perceptions, ripples of association all under the control of the music. These sorts of processes are *not* explained by the present theory, but they might be accounted for in terms of some aspects of the design of intelligent systems not discussed here, such as the need for subtle forms of integration and synchronisation of many processes in controlling physical movement. The synchronisation is needed both within an individual and between individuals engaged in co-operative tasks. Music seems to be able to take control of some such processes.

CONCLUSION

This sketch of some aspects of the architecture of the human mind, provides a framework for thinking about a range of possible types of intelligent systems, natural and artificial, though the analysis still has gaps. In particular, our account of pleasure and pain requires development, and we are not yet able to give an acceptable analysis of what it is to find something funny!

We have outlined a number of mechanisms and processes, which probably don't occur in all animals. In some less flexible forms of intelligence, the process of selection of a motive could be inseparable from the process of initiating action: operative motives could not be dormant. So not every intelligent system will necessarily have emotions, though it is very likely for any machine or animal whose collection of motives and motive generators is comparable in variety to those of human beings, and whose perceptual and reasoning abilities are similar. The mechanisms producing emotions are also the mechanisms required for great flexibility in a complex environment.

The model is relevant to a number of old philosophical problems, for instance the relationship between mind and body, and the question whether free will is possible. It can be argued that the only significant form of free will is that which involves taking decisions on the basis of one's own motives, beliefs, etc., as opposed to being externally constrained. We have begun to sketch an explanation of the possibility of this sort of freedom. In this sense a robot might also be free. (Whether we would be wise to make such robots is another question.)

The model may be important for psychotherapy and education, since it reveals enormous scope for 'bugs'. For instance, recursive escalation of emotions might account for some catatonic states. Moreover, there are many ways in which the processes by which motives are generated, compared, selected for action, related to planning, triggered when dormant, etc. may go wrong. The theory also implies that processes of learning and cognitive development, which are often studied as if they were autonomous, will occur within the framework of a complex and frequently changing collection of motives and motive-generators. These, and the emotional and other processes they generate must have a profound influence on what is learnt when, and it is to be expected that there will be enormous variation between individuals. A deep understanding of the full range of computational possibilities is surely a prerequisite for any systematic attempt to deal with the human problems.

ACKNOWLEDGEMENTS

This paper was written by the first author, using some ideas of the second author. Margaret Boden also helped(See [1]). Discussions in an SRC-funded distributed computing project, including Jim Hunter, Keith Baker, Paul Bennett, David Owen and Allan Ramsay provided some of the background. Sue Adams and Judith Dennison helped with production. [4] is a longer fuller version of this paper.

REFERENCES
(Truncated for lack of space)

[1] Boden, Margaret *Purposive Explanation in Psychology* Harvard University Press 1972, Harvester Press 1978.

[2] Heider, Fritz, *The Psychology of Interpersonal Relations,* Wiley 1958.

[3] Sloman, Aaron, *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind,* Harvester Press, 1978.

[4] Sloman, Aaron, and Monica Croucher "You don't need a soft skin to have a warm heart", CSRP No 4, School of Cognitive and Computing Sciences, University of Sussex, 1981.