A. Sloman, ‘Motives Mechanisms Emotions’ in *Emotion and Cognition* 1,3, pp.217-234 1987, reprinted in M.A. Boden (ed) *The Philosophy of Artificial Intelligence* "Oxford Readings in Philosophy" Series Oxford University Press, pp 231-247 1990. (Also available as Cognitive Science Research Paper No 62, Sussex University.)

MOTIVES MECHANISMS AND EMOTIONS

Aaron Sloman
Cognitive Studies Programme
University of Sussex
Brighton BN1 9QN
England
(At the University of Birmingham, School of Computer Science, since 1991)

Introduction

Ordinary language makes rich and subtle distinctions between different sorts of mental states and processes such as mood, emotion, attitude, motive, character, personality, and so on. Our words and concepts have been honed for centuries against the intricacies of real life under pressure of real needs and therefore give deep hints about the human mind.

Yet actual usage is inconsistent, and our ability to articulate the distinctions we grasp and use intuitively is as limited as our ability to recite rules of English syntax. Words like "motive" and "emotion" are used in ambiguous and inconsistent ways. The same person will tell you that love is an emotion, that she loves her children deeply, and that she is not in an emotional state. Many inconsistencies can be explained away if we rephrase the claims using carefully defined terms. As scientists we need to extend colloquial language with theoretically grounded terminology that can be used to mark distinctions and describe possibilities not normally discerned by the populace. For instance, we’ll see that love is an attitude, not an emotion, though deep love can easily trigger emotional states. In the jargon of philosophers (Ryle 1949), attitudes are dispositions, emotions are episodes, though with dispositional elements.

For a full account of these episodes and dispositions we require a theory about how mental states are generated and controlled and how they lead to action -- a theory about the mechanisms of mind. The theory should explain how internal representations are built up, stored, compared, and used to make inferences, formulate plans or control actions. Outlines of a theory are given below. Design constraints for intelligent animals or machines are sketched, then design solutions are related to the structure of human motivation and to computational mechanisms underlying familiar emotional states.

Emotions are analysed as states in which powerful motives respond to relevant beliefs by triggering mechanisms required by resource-limited intelligent systems. New thoughts and motives get through various filters and tend to disturb other ongoing activities. The effects may interfere with or modify the operation of other mental and physical processes, sometimes fruitfully sometimes not. These are states of being "moved". Physiological changes need not be involved. Emotions contrast subtly with related states and processes such as feeling, impulse, mood, attitude, temperament; but there is no space for a full discussion here.

On this view we need posit no special subsystem to account for emotions since mechanisms underlying intelligence suffice. (Compare Oatley and Johnson-Laird 1985). If emotional states arise from mechanisms required for coping intelligently in a complex and rapidly changing world, this challenges the common separation of emotion and cognition. This applies equally to human beings, other animals, or intelligent machines to come.

Design constraints for a mind

The enormous variety of animal behaviours indicates that there are different ways of designing agents that take in information from the environment and are able to act on it individually or co-operatively. Human beings merely occupy one corner of this "space of possible minds". Elsewhere I have sketched constraints determining the design solutions embodied in the human mind. I have space only to summarise the key results relevant to emotions.

Constraints include: a multiplicity of internal and external sources of motivation (often inconsistent), speed limitations, inevitable gaps and errors in beliefs about the environment and varying degrees of urgency associated with motives. Resource limits and urgency render inevitable the use of potentially unreliable ‘rule-of-thumb’ strategies. Unpredictability of new information and new goals implies a need to be able to interrupt, modify, suspend or abort ongoing activities, whether external or internal. This includes such things as hardware and software "reflex" actions, some of which should be modifiable in the light of experience.

Reflexes are inherently fast-acting but stupid. They may be partly controlled by context-sensitive filters using rules-of-thumb to assess priorities rapidly and allow extremely important, urgent, or dangerous ongoing activities to proceed without disturbance while allowing new, specially important or urgent, motives to interrupt them. (Below I define the "insistence" of a motive in terms of its ability to get past such filters.) A major conclusion is that intelligent systems will have fast though stupid sub-systems, including filters for new motives. Fast dumb filters will sometimes let in undesirables.

Incomplete information and the need to cope with long term change in the social or physical environment require higher order sources of action that provide learning: not only generators and comparators of motives, but generators and comparators for the generators and comparators themselves.

Although several independent sub-systems can execute plans in parallel, like eating and walking, conflicts among requirements can generate incompatible goals, necessitating a decision-making mechanism. The two main options are a "democratic" voting scheme, and a centralised decision-maker. If subsystems do not all have access to the full store of available information or not all have equal reasoning powers, a "democratic" organisation may be dangerous. Instead a specialised central mechanism is required for major decisions (Sloman 1978 chapters 6 and 10). This seems to be how normal human minds are organised,

Similar constraints determine the design of intelligent artefacts. Physical limitations of biological or artificial computing equipment necessitate major divisions of functions, including the allocation of the highest level control to a part with access to most information and the most powerful inference mechanisms. However, an occasional urgent need for drastic action requires overriding hardware or software reflexes that operate independently of higher level control -- a mechanism enabling emotional processes described below.

Goal generators

Many different sorts of motivators are pointed to by ordinary words and phrases such as:

> aims, attitudes, desires, dislikes, goals, hates, hopes, ideals, impulses, likes, loves, preferences, principles, alluring, amusing, bitter, boring, charming, cheering, depressing, distressing,

and many, many more. They mark subtle distinctions between different springs of action and the various ways things affect us. Conceptual analysis (Sloman 1978 Ch 4) brings out their presuppositions. A key concept is having a goal.

To a first approximation, to have a goal is to use a symbolic structure, represented in some formalism, to describe a state of affairs to be produced, preserved, or prevented. The symbols need not be physical structures: virtual formalisms will do as well (Sloman 1984). Goals can use exactly the same descriptive formalism as beliefs and hypotheses. The difference is solely in the roles they play.

> A representation of a state of affairs functions as a goal if it tends (subject to many qualifications) to produce behaviour that changes reality to conform to the representation.
>
> A representation functions as a belief if it is produced or modified by perceptual and reasoning processes which tend (subject also to many qualifications) to alter the representations to conform to reality.

(What "conform" means here can’t be explained without a lengthy digression.) The same representations may also be used in other roles, as instructions, hypothesised situations, rules, etc.

Some new goals subserve a prior goal and are generated by planning processes. Some are responses to new information, such as wanting to know what caused the loud noise around the corner. Goals are not triggered only by external events: a thought, inference or recollection may have the same effect.

How can a goal be a produced by a belief or thought? If goals involve symbolic structures, a computational explanation might be that a ‘goal-generating’ condition-action rule is used. For example, a benevolence rule might be: "if X is in distress generate the goal [X is not distressed]". A retribution goal-generator underlying anger might be "If X harms me generate the goal [X suffers]". A full analysis would describe various "goal generators", "goal generator generators", and so on. A learning system would produce new goal generators in the light of experience, using generator generators.

Goal comparators

Generators do not always produce consistent goals. Different design constraints lead to different co-existing goal generators. Social animals or machines need goal generators that produce goals for the benefit of others, and these can conflict with the individual’s own goals and needs. Goal comparators are therefore needed for selection beween different ends.

Some comparators apply constraint goals in planning, for instance using a "Minimise cost" rule to select the cheaper of two sub-goals. Others directly order ends, like a rule that saving life is always more important than any other goal, but not because of some common measure applicable to both. Since there are different incommensurable sources of motivation and different bases for comparison, there need not be any *optimal* resolution of a conflict.

Higher order motivators

Despite possibly confusing colloquial connotations, I use the general term ‘motivator’ to refer to mechanisms and representations that tend to produce or modify or select between actions, in the light of beliefs. Motivators recursively include generators and comparators of motivators. Some motivators are transient, like the goal of picking up a particular piece of cake, while others are long term, like an ambition to be slim.

Motivators should not be static -- motivator generators are required for flexible production of new goals. Still higher intelligence involves the ability to learn from experience and modify the generators. Thus the requirement for generators is recursive. The same applies to comparators: if two generators, regularly come into conflict by generating conflicting goals, then it may be necessary to suppress or modify one of them. This requires a generator comparator. Comparator generators and comparator comparators are also needed. Higher order generators and comparators account for some personality differences. Their effects account for some of the subtleties of emotional states.

Theoretical research is needed to design generally useful higher-level generators and comparators. Empirical research is needed to establish what the mechanisms are in people. Do we have a bound on levels of generators and comparators or can new levels be recursively generated indefinitely?

Varieties of motivators

‘Derivative’ and ‘non-derivative’ motivators can be distinguished. Roughly, a motivator is derivative if it is explicitly derived from another motivator by means-ends analysis and this origin is recorded and plays a role in subsequent processing. A desire to drink when thirsty would be non-derivative, whereas a desire for money to buy the drink would be derivative. A motive can be partly derivative, partly non-derivative, like a desire to quench one’s thirst with whisky to impress others. I shall try to show how non-derivative motivators are central to emotional states.

The distinction has behavioural implications. Derivative goals are more readily abandoned and their abandonment has fewer side-effects, e.g. if they appear to be unattainable or if the goals from which they are derived are satisfied or abandoned. An unpromising derivative goal can easily be replaced by another if it serves the supergoal as well. If a non-derivative goal is abandoned because it conflicts with something regarded as more important it may continue to demand attention - one source of emotions. Abandonment will produce regret and a disposition to revive the goal if the inconsistency can be removed.

Human non-derivative goals include bodily needs, desire for approval, curiosity, aesthetic wishes and the desire to succeed in tasks undertaken. Because these goals serve more general biological purposes some theorists regard them as derivative. However, the mechanisms that create a goal need not explicitly associate it with higher level goals, but simply give it the causal power to produce planning and action, for instance by simply inserting a representation of the goal in a database whose contents constantly drive the system. Despite its implicit function such a goal is non-derivative for the individual.

Quantitative dimensions of variation

Motives can be compared on different dimensions, definable in terms of the mechanism sketched above. *Insistence* of a motive is its interrupt priority level. Insistent desires, pains, fears, etc. are those that more easily get through interrupt filters, depending on the threshold set in relation to current activities.

Goals that get through filters need to be compared to assess their relative *importance.* This (sometimes partial) ordering is determined by beliefs and comparators, and may change if they do. Complex inferences may be required. Importance of a derivative goal is linked to beliefs about effects of achieving or not achieving it. Insistence concerns how likely a goal is to get through the interrupt filter in order to be considered, whereas importance concerns how likely it is to be adopted as something to be achieved if considered. Insistence has to be assessed very quickly, and should correlate with importance but sometimes will not. A bad filter will assign low priorities to important goals, and vice versa. A desire to sneeze doesn’t go away just because silence is essential for survival. (Not all animals have such complex motivational systems.)

Urgency. is a measure of how much time is left before it is too late. This is not the same as insistence or importance: something not wanted very much may be urgent, and vice versa.

Intensity of a goal determines how actively or vigorously it is pursued if adopted. It is partly related to urgency and importance, and partly independent of them. Obstacles to an intense goal tend to be treated as a challenge rather than a reason to abandon the goal. Often a long term important goal will lose out to something much less important but more intense - the age-old conflict between desire and duty. Ideally insistence, intensity and importance should be correlated, but the relationship can be upset by interactions with urgency and the way reflex assignments of insistence or intensity derive from prior experience or evolutionary origins.

Another measure of a motive is how distressing or disruptive failure to achieve it is. Different again is how much pleasure is derived from fulfilment. This can be assessed by how much effort tends to go into preserving the state of fulfilment, or achieving a repetition at a later time. Both are normally expressed as how much someone "cares", and relate to the potential to generate emotional states as described below.

These different kinds of ‘strength’ of motives all play a role in cognitive functioning and may be needed in sophisticated robots. They can have subjective correlates in a system with self-monitoring, though self-monitoring is not always totally reliable. Objectively they are defined in terms of dispositions to produce effects or resist changes of various sorts, internal or external. Different combinations of strengths will affect what happens at various stages in the evolution of a goal, from initial conception to achievement, abandonment, or failure.

Summary of processes involving motives

So far the following intermediate processes through which motives may go have been sketched:

* Initiation - by a body monitor, motive generator, or planner creating a new subgoal

* Reflex prioritisation of a new goal - assigning insistence

* Suppression or transmission by the interrupt filter

* Triggering a reflex action (internal or external, hardware or software)

* Evaluation of relative importance, using comparators.

* Adoption, rejection, or deferred consideration - adopted motives are generally called ‘intentions’. Desires may persist as desires, though not adopted for action.

* Planning - ‘intrinsic’ planning is concerned with how to achieve the goal, ‘extrinsic’ planning with when, and how to relate it to other activities - e.g. should it be postponed?

* Activation - starting to achieve the motive, or re-activating temporarily suspended motives.

* Plan execution

* Interruptions - abandonment or suspension

* Comparison with new goals

* Plan or action modification in the light of new information or goals, including changes of speed, style, or subgoals.

* Satisfaction (complete or partial)

* Frustration or violation

* Internal monitoring (self awareness)

* Learning - modification of generators and comparators in the light of experience.

These are all computational processes, capable of being expressed in terms of rule-govered manipulation of representations of various sorts, though filling out the details is not a trivial task. I’ll now try to indicate how they relate to emotions. The full story is very complicated.

An example: anger.

What is meant by "X is angry with Y". This implies that X believes that there is something Y did or failed to do and as a result one of X’s motives has been violated. This combination of belief and motive does not suffice for anger, since X might merely regret what happened or be disappointed in Y, without being angry. Anger also requires a new motive in X: a desire to hurt or harm Y. Most people and many animals seem to have retributive motive generators that react like this, alas. The new motive is not necessarily selected for action however intense it may be: fear of consequences and appropriate comparators may keep it inoperative.

Production of the new desire is still not sufficient for anger. X may have the desire, yet put it out of mind and calmly get on with something else: in that case he is not angry. Alternatively, X’s desire to do something unpleasant to Y may be entirely derivative: purely a practical measure to reduce the likelihood of future occurrence, without any ill-will felt towards Y. Then if X can be assured somehow that there will be no recurrence, he will drop the motive. That is not anger.

Anger involves an *insistent* and *intense* non-derivative desire to do something to make Y suffer. High insistence means the desire frequently gets through X's filters to 'request attention' from X's decision-making processes. So even after rejection by comparators, the desire frequently comes back into X's thoughts, making it hard for him to concentrate on other activities. Filters designed for speed can be too stupid to reject motives already ruled out by higher levels. Moreover, the desire must not be derivative, that is a subgoal that will disappear if a supergoal is removed. In socially sophisticated agents, anger may include a belief that Y's action had no social or ethical justification.

So emotions are states produced by motivators, and involve production of new motivators.

The violation of the original motive, and the insistence of the new motive, may be associated with additional secondary effects. For example, if X becomes aware of his anger this can make him annoyed with himself. If other people perceive his state, this can also affect the nature of the emotion. The episode can revive memories of other situations which enhance the anger.

Sometimes, in human beings, emotional states produce physiological disturbances too, probably as a result of the operation of physical and chemical reflexes driven by 'rule-of-thumb' strategies, as suggested above. However, if X satisfied enough of the other conditions he could rightly be described as very angry, even without any physical symptoms. Strong anger can exist without any physical side effects insofar as it constantly intrudes into X's thoughts and decisions, and insofar as he strongly desires to make Y suffer, and suffer a great deal. Though non-physical anger might be called 'cold', it would still have all the socially significant aspects of anger.

Anger is partly dispositional in that it need not *actually* interfere with other motives: for instance if the new motive to punish Y is acted on, there need be no further disturbance. But the anger has the *potential* to disturb other activities if the new motive has high insistence.

Anger is sometimes *felt*, as a result of self monitoring. But it is possible to be angry, or in other emotional states, without being aware of the fact. For example, I suspect that dogs and very young children are unaware of their anger (though very much aware of whatever provoked it).

Emotions like anger can vary along different quantitative and qualitative dimensions, such as: how certain X is about what Y has done, how much X cares about it (i.e. how important and intense the violated motive is); how much harm X wishes to do to Y; how important this new desire is, how intense it is, how insistent it is, how long lasting it is; how much mental disturbance is produced in X; how much physiological disturbance there is; which aspects of the state X is aware of; how many secondary motives and actions are generated. Different dimensions will be appropriate to different emotions.

Variations at different stages of the scenario correspond to different states, some not emotional. 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 simply 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 involves dismay. Because arbitrarily many motives, beliefs and motive generators can be involved, with new reactions triggered by the effects of old ones, the range of variation covered by this theory is bound to be richer than the set of labels in ordinary language. It will also be richer than the range of physiological responses.

Towards a generative grammar for emotions

Analysing anger and other other emotions in the light of the mechanisms sketched above, suggests the following components of emotional states:

* There is at least one initiating motive M1 with a high level of importance and intensity.

* A belief B1 about real or imagined or expected satisfaction or violation of M1 triggers generators of various kinds, often producing new motives.

* Different sorts of cases depend on (a) whether M1 is concerned with something desired or disliked, (b) whether B1 is a belief about M1 being satisfied or violated (c) whether B1 concerns past, present or future (d) whether B1 involves uncertainty or not (e) whether the agent is aware of his emotion or not (f) whether other agents are thought to be involved or not, (g) whether M1 is concerned with how other agents view one. (Cf. Roseman 1979).

* In more complex situations several motives simultaneously interact with beliefs, e.g. a situation where B1 implies that important motives M1a and M1b are inconsistent - e.g. in dilemmas.

* Sometimes M1 and B1 trigger a generator that produces a secondary motive, M2, for instance a desire to put things right, preserve a delight, punish a perpetrator or inform others. This in turn can interact with other beliefs, to disturb, interrupt, or otherwise affect cognitive processes. This would be a ‘two level’ emotional state. Several levels are possible.

* Sometimes M1 and B1 trigger several motive generators simultaneously. The resulting interactions can be very complex especially when new motives are in conflict, e.g. a desire to undo the damage and to catch the culprit.

* Sometimes the newly generated motives conflict with previously existing motives.

* New motives with high insistence get through interrupt filters and tend to produce (though they need not actually produce) a *disturbance,* i.e. continually interrupting thinking and deciding, and influencing decision-making criteria and perceptions.

* Thoughts as well as motives can interrupt. Even with no new motive there may simply be a constant dwelling on M1 and B1. This is especially true of emotions like grief, involving what can’t be undone. Such compulsive dwelling might might derive from triggering of automatic learning mechanisms concerned with re-programming generators.

* New motives need not be selected for action. M2 may be considered and rejected as unimportant, yet continue to get through interrupt filters if its insistence is high.

* In some emotional states, like fright, M2 triggers reflex action, by-passing deliberation and planning, and interrupting other actions (Sloman 1978 ch 6). ‘Software reflexes’ are called ‘impulsive’ actions. Reflexes make it possible to take very rapid remedial action or grasp sudden opportunities. Sometimes they are disastrous, however. Some reflexes are purely mental: a whole barrage of thoughts and feelings may be triggered.

* Some emotional states arise out of the individual’s own thoughts or actions, for instance fear generated by contemplating possible errors. Secondary motives may be generated to take extra care, etc. These secondary motives may generate so much disturbance that they lead to disaster.

* Some emotions involve interrupting and redirecting many ongoing processes, for instance processes controlling different parts of the body in restoring balance. If sensory detectors record local changes, the system’s perception of its own state will be changed.

* Self-monitoring processes may or may not detect the new internal state. If not, X will not be conscious of, or feel, the emotion. Internal monitoring need not produce recognition: e.g. relevant schemata might not have been learnt (Sloman 1978, ch 10). People have to learn to discriminate and recognise complex internal states, using perceptual processes no less complex than recognising a face or a typewriter.

* Recognition of an emotion can produce further effects, e.g. if the internal state fulfils or violates some motive. It may activate dormant motives or motive-generators and possibly lead to successively higher-order emotions (recursive escalation).

The interruptions, disturbances and departures from rationality that characterise some emotions are a natural consequence of the sorts of mechanisms arising from constraints on the design of intelligent systems, especially the inevitable stupidity of resource-limited interrupt filters that have to act quickly. A robot with an infinitely fast computer and perfect knowledge and predictive power would not need such mechanisms. However, not all emotions are dysfunctional: when walking on a narrow ledge it is important that you don’t forget the risks.

These mechanisms allow so many different sub-processes in different situations that no simple table of types of emotions can do justice to the variety. The same rich variation could characterise the detailed phenomenology of emotions in clever robots with self-monitoring abilities.

A full account of how people typically *feel* anger, elation, fear, etc. would have to include bodily awareness. Yet what makes many emotions important in our lives is not this sort of detail, but the more global cognitive structure. Fury matters because it can produce actions causing harm to the hater and hated, not because there is physical tension and sweating. Grief matters because the beloved child is lost, not because there’s a new feeling in the belly. So it would be reasonable for us to use terms like ‘afraid’, ‘disappointed’, ‘ecstatic’ ‘furious’ or ‘grief-stricken’ to describe the state of mind of an alien being, or even a sufficiently sophisticated robot, without the physiological responses. (Contrast Lyons 1980).

Moods, attitudes and personality

A *mood* is partly like an emotion: it involves some kind of global disturbance of, or disposition to disturb, mental processes. But it need not include any specific beliefs, desires, inclinations to act, etc. In humans, moods can be induced by chemical or by cognitive factors, for instance drinking or 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. A more detailed theory would have to distinguish different mechanisms, for instance global ‘hardware-induced’ speed changes of certain sub-processes and global ‘software-induced’ changes in relative priorities of motives or inference strategies.

An *attitude*, such as love or admiration, is a collection of beliefs, motives, motive generators and comparators focussed on some individual, object, or idea. One who loves his children will acquire new goals when he detects dangers or opportunities that might affect their well-being. The strength of the love determines the importance and interrupt priority levels assigned to such goals. Selfishness is a similar attitude to oneself. In communities of intelligent systems able to think and care about the mental states of others, the richness and variety of attitudes makes them an inexhaustible topic for study by poets, novelists and social scientists. Attitudes are often confused with emotions. It is possible to love, pity, admire, or hate someone without being at all emotional about it. Attitudes are expressed in tendencies to make certain choices *when the opportunity arises*, but need not include continual disturbance of thoughts and decisions. One can love one’s children without having them constantly in mind, though news of danger to loved ones may trigger emotions.

Character and personality include long term attitudes. Generosity for example is not a goal but a cluster of goal generators that produce new goals in response to information about another’s needs and comparators that select them over more self-centred goals. Hypocrites produce similar goals but never adopt them for action. A personality or character is a vast collection of unfocussed general dispositions to produce certain goals in specific situations. The set of such collections is too rich for ordinary adjectives. A whole novel may be required to portray a complex personality. More generally the space of possible mental states and processes is too rich and complex for colloquial labels like "attitude", "emotion", "mood" to survive in an adequate scientific theory.

There are many kinds of deep and moving experiences that we 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 involve powerful interactions between perception and a large number of additional processes, some physical as well as mental. Listening to music can produce a tendency to move physically and also a great deal of mental ‘movement’: memories, perceptions, ripples of association all controlled by the music. Such processes might be accounted for in terms of aspects of the design of intelligent systems not discussed here, such as the need for associative memories and subtle forms of integration and synchronisation in controlling physical movements. The synchronisation is needed both within an individual and between individuals engaged in co-operative tasks. Music seems to take control of some such processes.

I conjecture that the mechanisms sketched here are capable of generating states we ordinarily describe as emotional - fear, anger, frustration, excitement, dismay, grief, joy, etc. The mechanisms are generative in the sense that the relevant motives beliefs, plans and social contexts can be indefinitely complex and the emotional processes they generate can be correspondingly complex and varied (Abelson 1973, Dyer 1981, Lehnert 1981). This means that no simple bounded taxonomy of emotional states can begin to capture the variety, any more than a taxonomy can capture the variety of sentences of English. (Cf. Roseman 1979.)

Does a scientific theory of mind need such concepts?

It is sometimes suggested that although concepts like "belief", "desire", "emotion" play an important role in individual thoughts about other people, they are not required for a fully developed scientific theory of the mind. In its extreme form this is materialist reductionism, but that is as implausible for psychology as the suggestion that concepts of software design can be replaced by concepts relating only to computing hardware.

A more subtle suggestion (S.Rosenschein SRI, personal communication) is that an entirely new collection of "intermediate level" concepts, unrelated to beliefs, desires, intentions, etc., will suffice for a predictively and explanatorily successful scientific theory of how people and other intelligent organisms work. Because it is unlikely that ordinary concepts can be dispensed with entirely in expressing significant generalisations about human behaviour (Pylyshyn 1986) I have taken a weaker stand: instead of totally replacing ordinary concepts we need to extend and refine them, showing how they relate to a working design specification.

Even if this sort of theory is wrong, it may be deeply implicated in semantics of natural language concepts concerning human mental states and actions. If so, a machine able to understand ordinary language and simulate human communication will require at least an implicit grasp of the theory.

Implications

Not all these mechanisms can be found in all animals. In some less intelligent creatures, selection of a motive might be inseparable from the process of initiating action: operative motives could not be dormant. In such animals or machines lacking the mechanisms required for flexibility in a complex environment, emotions in the sense described here would be impossible.

It is also unlikely that all of this richness exists in young children. By investigating the development of the cognitive and computational mechanisms in children, including the motivational mechanisms, we can hope to understand more about their emotional states. In particular, it seems that many higher order generators and comparators are not available to infants, and that interrupt filters are far less selective than in most adults, which is not surprising if software filters are the result of learning.

The very complexity of the mechanisms described reveals enormous scope for ‘bugs’. Motive generators and comparators could produce unfortunate desires and preferences. Interrupt priorities may be assigned in a way that doesn’t correlate well with reflective judgements of importance. Thresholds for interrupts may be set too high or too low. Learning processes that modify generators and comparators may be too quick to change things on flimsy evidence. Given the inevitable stupidity of some of the faster reflexes and filters, we can expect some kinds of malfunctions of generators and comparators to lead to intense emotions that interfere with normal cognitive or social functioning. Reactions to unfulfilled motives may be too strong, or too weak for the long term good of the individual or his associates. The relative importance assigned to different sorts of motives by the goal assessment procedures may produce a tendency to select goals that are unachievable or achievable only at enormous cost. Dormant, temporarily suspended motives may too often go unattended because the monitoring process fails to detect opportunities, perhaps because of inadequate indexing. The pervasiveness of ‘rules-of-thumb’ for coping with inadequate information, limited resources, and the need for speed, provides enormous scope for systematic malfunction. Recursive escalation of emotions might account for some catatonic states.

The inevitability of familiar types of fallibility should be a matter of concern to those who hope that important decisions can be taken very rapidly by machines in the not too distant future.

In fact, if people are as complex and intricate as we have suggested, it is amazing that so many are stable and civilised. Perhaps this theory will reveal types of disturbance we previously could not recognize.

The theory implies that processes of learning and cognitive development, occur in a framework of a complex and frequently changing collection of motivators. These and the 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. The implications for educators have yet to be explored.

Conclusion

A theory of this general sort is a *computational* theory of mind. The computations may occur in a *virtual* machine implemented in lower level machines, brain-like or computer-like: they need not be implemented directly in physical processes. So the theory is neutral between physically explicit representations as found at low levels in conventional computers and implicit or distributed representations studied in neural-net models.

The test of this approach will be the explanatory power of the theories based on it. We need both a systematic explanation of the whole range of possibilities we find in human behaviour and an account of how people differ from one another and from other actual and possible behaving systems. (Concerning explanations of possibilities see Sloman 1978, chapter 2.)

Understanding computational mechanisms behind familiar mental processes may enable us to reduce suffering from emotional disturbances, learning disabilities, and a range of social inadequacies. Some problems may be due to brain damage or neural malfunction. Other problems seem more like software faults in a computer. I conjecture that many emotionally disturbed people are experiencing such software bugs.

The analysis still has many gaps. In particular, an account of pleasure and pain is missing, and I am not yet able to give an acceptable analysis of what it is to find something funny! There are states like being thrilled by rapid motion, spellbound by a sunset, moved to tears by reading a book or watching a play, that require more detailed analysis. I have not discussed the many aspects of human emotional life that arise contingently from our evolutionary history and would not necessarily be found in well-designed robots. So there is much yet to be done. Nevertheless, the theory provides a framework for thinking about a range of possible types of intelligent systems, natural and artificial -- part of our general study of the space of possible minds. Attempting to test the ideas in working computer simulations will surely reveal gaps and weaknesses.

Acknowledgements

This work is supported by a fellowship from the GEC Research Laboratories, and a grant from the Rennaisance Trust. Some of the ideas derive from Simon’s seminal paper. My ideas have benefited from discussions over several years with Monica Croucher. Her thesis developed the ideas of this paper in greater depth. The view of ordinary language as a source of information about the human mind, sketched here and expanded in chapter 4 of Sloman 1978, owes much to Austin (1961). Keith Oatley’s editorial comments have been most helpful.

Biblography

Abelson, R.A. ’The structure of belief systems’ in, R.C.Schank and K.M.Colby (eds) *Computer Models of Thought and Language,* W.H.Freeman 1973

Austin, J.L. (1961), ’A Plea for Excuses’ in his *Philosophical Papers,* Oxford University Press, 1961, reprinted in A.R.White (ed), *Philosophy of Action,* Oxford University Press, 1968.

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

Boden, Margaret *Artificial Intelligence and Natural Man,* Harvester Press, 1978.

Croucher, Monica. (1985). *A Computational Approach to Emotions*, Unpublished Thesis, University of Sussex, 1985

Dennett, D.C., *Brainstorms,* Harvester Press, 1979.

Dyer, Michael G., ’The role of TAUs in narratives’, *Proceedings Cognitive Science Conference,* Berkeley, 1981.

Edelson, Thomas, ’Can a system be intelligent if it never gives a damn’, in *Proceedings AAAI-86*

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

Lehnert W.G, J.B.Black and B.J.Reiser, ’Summarising Narratives’, in *Proceedings 7th International Joint Conference on A.I.* Vancouver 1981.

Lyons, William, *Emotion,* Cambridge University Press, 1980

Oatley, Keith and P.N.Johnson-Laird, ’Sketch for a cognitive theory of the emotions’, Cognitive Science Research Paper No CSRP.045, Cognitive Studies, University of Sussex, 1985.

Pylyshyn, Zenon W. *Computation and Cognition: Toward a Foundation for Cognitive Science,* The MIT Press, 1986.

Roseman, Ira, ’Cognitive aspects of emotion and emotional behaviour’, presented to 87th Annual Convention of the American Psychological Association, 1979.

Ryle, Gilbert, *The Concept of Mind,* Hutchinson 1949.

Simon, H.A., ’Motivational and Emotional Controls of Cognition’ 1967, reprinted in *Models of Thought,* Yale University Press, 1979.

Sloman, Aaron ’How to derive "Better" from "Is"’, *American Philosophical Quarterly,* 1965

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

Sloman, Aaron ’Skills Learning and Parallelism’, Proceedings Cognitive Science Conference, Berkeley 1981.

Sloman, Aaron and Monica Croucher ’Why robots will have emotions’, in *Proceedings 7th International Joint Conference on A.I.* Vancouver 1981.

Sloman, A. ’Why we need many knowledge representation formalisms’ in M. Bramer (ed) *Research and Development in Expert Systems,* Cambridge University Press, 1984.

Sloman, A. 'Real-time multiple-motive expert systems' in Martin Merry (ed), *Expert Systems 85* Cambridge University Press, 1985