Towards a grammar of emotions

Aaron Sloman, Cognitive Studies Programme, the University of Sussex.

My favourite leading question when teaching Philosophy of Mind is 'Could a goldfish long for its mother?' This introduces the philosophical technique of 'conceptual analysis', essential for the study of mind (Sloman 1978, ch. 4). By analysing what we mean by 'A longs for B', and similar descriptions of emotional states, we see that they involve rich cognitive structures and processes, i.e. computations. Anything which could long for its mother would have to have some sort of representation of its mother, would have to believe that she is not in the vicinity, would have to be able to represent the *possibility* of being close to her, would have to desire that possibility, and would have to be to some extent pre-occupied or obsessed with that desire. That is, it should intrude into and interfere with other activities, like admiring the scenery, catching smaller fish, etc. If the desire were there, but could be calmly put aside, whilst other interests were pursued, then it would not be truly a state of longing. It might be a state of preferring. Thus longing involves computational interrupts. The same seems to be true of all emotions.

All this talk of desires, beliefs, and processes involving them seems to presuppose that there is some kind of machine within any sort of intelligent ghost. That is there must be mechanisms for constructing and manipulating a variety of symbolic structures representing things, properties, relationships, actions, etc. Of course, it may be that the operation of mind is pure mystery, and ultimately inexplicable: some kind of magic. That view is essentially undiscussable. To help undermine it, if it is false, we can only try to show in detail how to build a machine which has all the allegedly inexplicable abilities, including the ability to have emotions. However, a defender of the mystery thesis can always claim that some unanalysable unobservable 'something' will forever be missing from such a machine. People who like such

mysteries are welcome to them. The rest of us will continue to try to understand how the human mind works. That is a long and arduous task. This paper merely illustrates some of the work to be done.

Example: what does 'X is angry with Y' mean?
X must believe Y to be responsible for something which violated one of X's motives, e.g. X believes Y did something X disliked, or that Y failed to do something X wanted done. ('Motive' is here used to refer to a type of representation which is capable of generating certain sorts of processes of decision and action. Further details will be published elsewhere.)

This is not sufficient, for X might merely *regret* the occurrence, or *forgive* Y, and not be *angry*. For X to be angry, his belief must interact with a motive-generator (Sloman and Croucher 1981) to produce a new motive: he must want to do something to violate one of Y's motives. X's new motive need not be selected for action: it need not become an *intention*, for instance because X is afraid of the consequences of acting on it, or disapproves of vengeance. Alternatively, execution of the intention may be postponed.

This is not yet enough. X 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 a fairly *intense* desire to do something to Y: the desire should frequently 'request attention', so it will frequently come back into X's thoughts, making it hard for him to concentrate on other activities. This aspect of anger, and other emotions, makes essential use of 'interrupt' mechanisms which can be shown (Sloman 1978, ch. 6, Sloman and Croucher 1981) to be desirable within any intelligent animal or robot with multiple motives in a complex and partly unpredictable world.

For *rational* anger, the desire to harm Y must not be one which X would have had in any case: it must be *based on* the belief that Y has violated one of X's motives: the desire to harm Y should be redirected to Z on learning that it was Z, not Y, who was responsible. Further, X must to some extent regard Y as a responsible agent who intended to do what he did. This sort of anger is possible only in animals able to represent others as having beliefs, motives, etc., and capable of suffering. More primitive species

might merely have learnt to respond with violence to violators of their motives. (People are often like this.)

Unconscious or irrational anger is possible too, and would require some of the conditions to be modified. In irrational anger X's desire to harm Y might not be indexed in such a way that it disappears when it is discovered that Y was not the cause of the violation of X's motive. Some of the states and processes might be conscious, others unconscious. In particular, even if the information about internal processes is accessible to self-monitoring processes, the monitoring routines need not have the descriptive resources to characterise what is happening accurately (see Sloman 1978, ch. 10, for more on the unconscious).

Besides producing mental disturbances, i.e. constantly intruding into X's decision making, anger may also produce physical disturbances, such as sweating, shaking, feelings of tension, tendencies to perform violent actions, such as stamping, thumping objects. These are sometimes related to mechanisms required for survival and achievement of complex goals. For instance, it is sometimes necessary suddenly to reorganise the movements of hands, arms, legs, etc. However, it is not *necessary* that anger involve any such physical effects. If X satisfied all the other conditions he could rightly describe himself as being angry, even very angry, despite not having the physical symptoms. The anger might then be described as 'cold' or 'cold-blooded'. Yet the feeling could be strong, insofar as it constantly intruded into thoughts and decisions, and insofar as X strongly desired to make Y suffer, and suffer a great deal.

Although anger can interrupt and disturb other activities, it need not actually do so. For instance, X may be engaged in pursuing his new motive of hurting Y. Current activity is produced by the anger: the anger is not therefore interrupting some other activity. The characteristic of anger and other emotions is a *disposition* or *tendency* to interrupt and disturb other activities. The disturbance may be entirely mental, of course: we are not talking about behavioural dispositions, but about computational processes.

The strength of anger, like other emotions, can vary along different dimensions. It can vary according to how much X minds what he thinks Y has done, which, in turn, will depend on how important the violated motive was. It can vary according to how

much harm X wants to do to Y. The strength can vary also with the priority assigned to the wish to harm Y: the desire may be very hard to override, or it may be relatively easy. It can vary according to how much the new motive, and thoughts about the violated motive, demand attention and disturb other processes. The new desire may be hard to ignore (intense) yet easily overridden (low priority). Finally, the strength of the anger can vary according to how much mental or physical disturbance it is capable of producing in X.

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

What are emotions? A sketch of the possibilities.

Many of the states we ordinarily refer to as emotions (anger, fear, delight, pity, awe, embarrassment, shame, pride, etc.) need have little or nothing to do with physiological changes or the awareness thereof. In fact, the emotional states are often accompanied by physiological processes, but they are not essential components. There are some emotions, such as being 'nauseated' by something, feeling cold shivers of fear, which do inherently involve awareness of bodily processes. Yet many of our ordinary emotion concepts, with all their subtlety and richness, could be applied to intelligent beings with totally different physiological processes from us, even beings in whom emotions were not correlated with any bodily changes (except the brain-processes required to support the computational processes). We can now summarise some of the main features of emotions.

There is at least one fairly strong motive: a desire for something to be the case or not be the case: past present or future. It may be currently active, or dormant (Sloman and Croucher 1981).

There must be some belief-state concerning the motive: a belief that it has or has not been satisfied or violated, or will or will not be, or explicit uncertainty about the matter. 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. A further distinction can be made as to whether the belief concerns the past or the future. Believing something *undesirable* has happened can generate despair, believing that it will happen can generate desperate clutching at straws. Believing that something *desired* has happened can produce an emotional state involving considerable pleasure or satisfaction, while the belief that it will happen can produce a desire to make doubly certain that it will be taking additional precautionary steps.

All this accounts for similarities and differences between anger, sorrow, despair, shame, anxiety, hope, anguished expectation, anxious or elated anticipation, and so on.

The belief produces a disposition to interrupt other mental processes (possibly via new motives): it is not possible calmly to put the matter out of mind and get on with other things. (One is *moved*.)

The belief (or uncertainty) may interact with a motive-generator to produce a new motive, e.g. a desire to right what has gone wrong, or to take new precautions. Instead of a new motive there may be constant dwelling on what has happened. Regrets or happy recollections may constantly intrude. (Interrupts again).

New motives need not be selected for action: they may remain inoperative, though disturbing. Some emotional states such as fright may involve the direct production of actions, by-passing processes of deliberation and decision. This includes 'impulsive' action. Mechanisms making this possible would also make it possible to take very rapid remedial action in times of great danger, or when sudden opportunities are recognized. (In less intelligent organisms, this might be the only way motives generate action.)

Some emotions arise directly out of actions by the individual. Some actions require great precision, and the risks of error may be very serious. During the action there may be considerable fear about possible errors, and secondary motives generated to take extra care, suppress interruptions, etc. These secondary motives may, in extreme cases, generate so much disturbance that they defeat the original motive.

Emotional states may be deliberately sought, at the theatre, or fun-fair. There need not be actual interruption or disturbance of other activities, though there will often be physiological distur-

bances. Yet the potential to interrupt mental processes remains. Attempting to turn ones attention completely on to proving theorems, or planning a holiday would be difficult. Otherwise the emotion is simulated, and the screams or tears fraudulent.

The possibility of interrupting processes is essential in coping with a complex and unpredictable environment. There may be rapid disturbance of a large number of ongoing processes, controlling different parts of the body, to cope with some new danger or opportunity. If those parts of the body contain sensors, then the local changes produced by the interruptions will be monitored, and self-awareness will be changed.

Feeling the emotion

The individual may or may not be aware of the disturbance or potential disturbance. Awareness may be partial. Awareness, if present, constitutes *feeling* anger, embarrassment, etc. It need not involve awareness of body states. If there is some internal representation of the state, this can activate dormant motives or motive-generators, e.g. producing a desire to control the emotion, or not to show it; or a second-order emotion, e.g. if being in the emotional state is itself undesirable. In social animals, these second-order effects may depend on the behaviour of others.

Attitudes and moods

Other mental states and processes have features in common with emotions. A mood is partly like an emotion in that it involves global disturbance of mental processes. But it need not be the intrusion of specific thoughts, desires and inclinations to act, rather a very general tendency to have thoughts and desires of certain kinds. Thus having a certain mood involves being disposed to have certain emotions rather than others. But it need not actually involve having any emotion: the dispositions or tendencies may remain mere potentialities.

Attitudes may be 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, *when the opportunity arises*, which need not be often. The person you hate may be out of your mind nearly all the time. Such attitudes include motives and motive-generators which are longer-lasting than the emotions they produce.

Extreme attitudes or moods may be emotions, if they are not easily ignored, and continually disturb other mental processes. If the hate or love is obsessive, and cannot be put out of mind, then it is an emotion. There is no sharp dividing line. The space of possible mental states and processes to too rich and complex for simple divisions to be useful.

Emotions and intelligence

Analysis of the architecture of a mind (fitting the bits and pieces studied in A.I. projects on perception, inference, planning, etc. into a larger framework) needs to be related to constraints imposed by the physical and social environment, the demands of a complex world where things change rapidly and predictions often fail. Humans and other animals will not have a simple set of hierarchically organised goals (e.g. all derived from some single goal such as survival), but a large collection of different motives and motive-generators derived from different needs and influences. Inconsistencies and violations will therefore occur. Mechanisms for coping intelligently will inevitably have the potential to produce emotional states (Sloman 1978, 1981). This suggests that the notion of a super-intelligent robot may be entirely mythical.

Some implications

Analysis of such global constraints and mechanisms for coping with them can lead to a survey of *possible* types of mind, a sort of 'grammar' for minds, including possible robots. Different mechanisms will suffice for coping with different sets of constraints. Different architectures seem to occur in different animals. For instance, in some animals selection of a motive for action is probably inseparable from the process of initiating the action, so there are no postponed intentions. Emotions related to dormant intentions would then be impossible. Some animals may be capable of postponing intentions, but incapable of interleaving their execution. Some may be incapable of storing reasons for actions and therefore abandoning them when goals are accidentally achieved. Such variations point to a generative grammar for possible types of mental mechanisms.

Analysis of computational architectures making the processes possible, has implications for philosophical discussions of the nature of mind and the relation between mind and body. For

instance, whereas behaviourists try to define mental events and processes in terms of their functional relationships to external stimuli and behaviour, we see how they can more fruitfully be analysed in terms of their relationships to internal processes. For instance a motive may be present, and request attention from time to time, yet always be overridden by others, and thus generate no behaviour. Further, a motive may be represented in the relevant database, yet be associated with such a low interrupt priority level that it never even gets enough attention to be rejected. It has the potential to be considered, and even to be accepted: but only if there are drastic changes in the rest of the store of beliefs and motives. (This illustrates the thesis of chapter 2 of Sloman 1978: science is more fundamentally a description and explanation of what *can* occur than of *laws*: a point understood by few psychologists.)

The study of mental illness and learning failures requires an explanation of *normal* functioning. A model of a mind as involving stores of motives and motive-generators, which can interact with one another and with percepts, beliefs and current activities, may be important for psychotherapy and education.

For instance, a system with the complexities sketched above (with motive generators, and motive generator generators, etc.) will be potentially very sensitive to experiences, and similar individuals may develop very differently owing to slight differences in the environment, the divergences feeding back into the system to cause still more divergence. The end state may be totally unlike anything envisaged by the programmer (if there is one) and in that case it will be entirely misleading to say that the motives which drive the 'mature' system are not its own, but the programmer's. There is enormous scope for 'bugs', for instance, recursive escalation of emotions, or interrupt mechanisms which are invoked far more often than is good for the system. 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.

As for education, the theory sketched here implies that processes of learning and cognitive development occur within the framework of a complex and frequently changing collection of motives and motive-generators, capable of interrupting other processes. These, and the emotional states they generate, must

have a profound influence on what is learnt when, what is perceived, which problems are solved, etc., and it is to be expected that there will be enormous variation between individuals. If education is ever to become a science, these processes will have to be understood.

All this in turn generates constraints on the computational resources individuals require for coping with a social environment: if other individuals are enormously complex and varied, simple and uniform social strategies will not suffice for interacting with them.

This article sketches work in progress. There are still many gaps, not least the lack of an adequate computational analysis of pleasure and pain, or of what it is to find something funny (which may have something to do with being a social animal).

Acknowledgment: Margaret Boden, Monica Croucher and various colleagues have helped me.

References

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

Aaron Sloman and Monica Croucher, 'Why robots will have emotions', in *Proceedings International Joint Conference on Artificial Intelligence* (University of British Columbia, Vancouver, 1981).