Prolegomena to a Theory of Communication and Affect

Aaron Sloman

School of Cognitive and Computing Sciences
University of Sussex
(Birmingham University from August 1991)

Abstract [1]

As a step towards comprehensive computer models of communication, and effective human machine dialogue, some of the relationships between communication and affect are explored. An outline theory is presented of the architecture that makes various kinds of affective states possible, or even inevitable, in intelligent agents, along with some of the implications of this theory for various communicative processes. The model implies that human beings typically have many different, hierarchically organized, dispositions capable of interacting with new information to produce affective states, distract attention, interrupt ongoing actions, and so on. High "insistence" of motives is defined in relation to a tendency to penetrate an attention filter mechanism, which seems to account for the partial loss of control involved in emotions. One conclusion is that emulating human communicative abilities will not be achieved easily. Another is that it will be even more difficult to design and build computing systems that reliably achieve interesting communicative goals.

1 Introduction

It isn't only for poets that communication and affect are often inextricably linked. We all know of cases where ill-chosen phrases provoke bad feeling then hurtful responses, escalating in a positive feedback loop that ends in tragically damaged relationships. Similarly, felicitous phrases can soothe painful wounds, or plant the seeds of an intense and fruitful relationship. What are the mechanisms that make all this possible? Affective states (as loosely defined below) are among the most important to us, and also among the least understood. Will cognitive science ever be able to explain the diversity of communicative processes found in, for example, flirting, teasing, taunting, threatening, consoling, enthusing, enthralling, entertaining, demoralising, deflating, encouraging, inspiring, and inducing uncontrollable fits of laughter or rage? Can these processes be simulated on computers? Will intelligent machines ever be able to advise, help or teach people, without falling foul of the complexities of human affective reactions?

Full answers would need to delve into many difficult questions concerning the relationships between affect and information transfer. It is not only content that can have affective significance: there are diverse communicative modes, media and formalisms, many with emotional, motivational, or aesthetic significance. Besides words, we react to tone of voice, intonation contours, amplitude, facial expression and gestures. In all cultures, the mutually enhancing effect of words and music is found in songs expressing joy, sorrow, resolve, and many other emotions and attitudes.

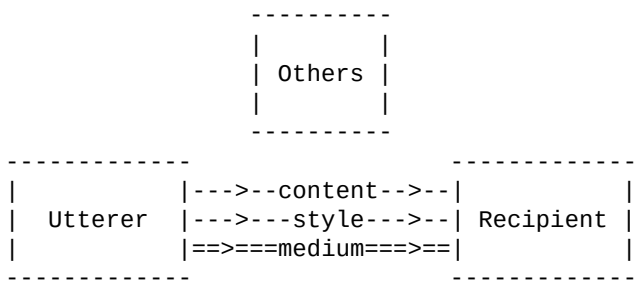
Not only deliberate communication needs to be taken into account, especially as regards affective states. Involuntary changes in facial expression and color, changes in posture, weeping, wincing, gasping and the like all

[1] Paper presented, Nov 1990, to NATO Advanced Research Workshop on "Computational theories of communication and their applications: Problems and Prospects". Published in Ortony, A., Slack, J., and Stock, O. (Eds.) *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues.* Heidelberg, Germany: Springer, 1992, pp 229-260. Also available as Cognitive Science Research Paper, CSRP-91-05, The University of Birmingham. Accessible online at http://www.cs.bham.ac.uk/research/cogaff/

help to reveal to others states that we might prefer to conceal, and sometimes even states (e.g. anger) that we are not aware of ourselves. Such behavioral expression of affect is surely not just accidental: from Darwin on, theorists have speculated about the evolutionary pressures in social organisms that could account for them. These involuntary communications could have functions for the group or the species, even if they do not serve the explicit purposes of the individuals concerned. A society of intelligent agents all able fully to conceal their internal states at will might be too inefficient: the constant second-guessing would lead to too many errors of judgement in which individuals ended up harming each other unnecessarily because they suspected non-existent malevolence, etc. This point is especially important in view of the fact, discussed at length in this volume by Cristiano Castelfranchi, that communication is often not co-operative.

A further complication is that "third" persons may also be involved, in a variety of different ways. Their real or imagined affective states may be referred to, expressed by, a cause of, or caused by the communication. Once again, different aspects of the communication can be involved in these relations, for instance the content, the style, or the medium. An example might be the use by fearful underlings of a whisper, or a special tone of voice, to avoid offending an awesome overseer.

In order to appreciate the full complexity of our topic the reader may imagine adding arrows to the diagram below to represent all the possible semantic, expressive or causal relations between aspects of an utterance and the



people indicated, including unintended as well as intended effects. It may help to use different kinds of arrows for semantic, for expressive, and for causal relations. It will quickly become apparent that there is a very tangled network of relationships. The tangle depicts the topic of this paper. No wonder so many people are so poor at communication. Can machines do half as well?

I have suggested that an affective state can be involved in communication, as one of its causes, as a reason or justification, as its intended effect, as unintended effect, as a side effect of the medium or mode, or as the content of what is communicated, implicitly or explicitly. Although I cannot hope to address all of these issues fully, I shall try at least to give a flavour of some of the implications for the design of human-machine interfaces and natural language systems. I hope to show that designers of interactive systems need caution and humility: it is too easy to overlook something, when dealing with such a complex and ill-understood set of mechanisms. I shall present an outline theory of some of the mechanisms involved in human affective states and use it to account for some of the ways in which communication can interact with affective states in the hearer. This will explain why it is often difficult for communicators to achieve their goals and avoid unintended effects. This has implications both for computer models of human communication and for intelligent machines designed to communicate with humans, where the communication has a practical purpose, such as teaching, persuading or counselling.

2 What Is Affect?

The words "affect" and "affective" do not have a clear, well-defined, meaning. Different professional groups and even the same individual at different times use the words in different ways. The survey by Ortony et al. (1987) implicitly defined affect by contrast with other categories of mental states. But that assumes we can already intuitively make

the distinction. Moreover their interest was in describing the structure of the affective lexicon, not in our topic, namely explaining how affective states come about and interact with other states and processes. Some people and some dictionaries treat affect as primarily concerned with emotions. The Concise Oxford Dictionary hints at a more general definition:

"affect n. (Psych) feeling, emotion, desire, esp. as leading to action."

Two questions arise: (a) What does "leading to action" mean here? (b) What are feeling, emotion and desire supposed to have in common that distinguishes them from other mental phenomena, such as thinking, attending and remembering, since all can lead to action? I shall first give a partial answer to (a), in terms of dispositional control hierarchies, then discuss (b), and then return to issues of control and their implications for communication.

Many kinds of mental states are connected with action, but not as necessary or sufficient conditions. Rather they are *dispositionally* related to action. The importance of dispositional analyses of mental states was argued at length by Ryle (1949, e.g. chapters IV and V). He was widely misunderstood as proposing a behaviorist analysis of mental states as dispositions to produce external behavior. In fact he was far more subtle than that, allowing some dispositions to produce mental states and processes, which themselves involve dispositions to produce other dispositions. In short, dispositions can be related to one another hierarchically, some having very indirect links with external behaviour. There are several states that are dispositional in that they are causally related to possible forms of behavior that may not actually occur: for example propensities, inclinations, tendencies, abilities, aptitudes, flairs, traits, potentialities and the like. All these are capable of issuing in, and explaining, behavior without constituting necessary or sufficient conditions for that behavior. They produce their effects via indirect, defeasible causal links.

The number of steps in the dispositional hierarchy can vary. For example, a personality that covets popularity might include a high-level disposition to self-modification so as to acquire attitudes, knowledge and skills that involve more specific dispositions to do things that will produce approval and liking in others. Insincere individuals simply have dispositions to produce the behavior to make them popular, whereas more sociable and generous personalities have a disposition to react to the needs of others by desiring to do what will please them, and these desires involve dispositions to produce the required behavior. All these dispositions have causal powers that are defeasible: they can be overridden by other attitudes or desires such as concern for one's family or fear of losing one's job. Noticing conflicts between attitudes or desires usually depends on reasoning about the consequences of various actions. Such inferences are not always made automatically, so inconsistencies may go unnoticed if the agent lacks the reasoning ability or is distracted by something else.

So causal connections between mental states and action are sometimes fairly direct (as in startles), but more often very indirect, going through several layers of dispositions, where, at each stage, other dispositions may cause the the effects to be suppressed. (This is one reason why it is almost impossible to find universal *laws* governing human behavior.)

Later I shall try to show how the hierarchy of dispositions could explain some of the unintended effects of communication: dispositions at different levels in the hearer may interact with incoming information in unexpected ways.

Since dispositions can form hierarchies, the trait/state distinction, often assumed by psychologists, is not really an absolute distinction. There are not two kinds of properties, but two kinds of relationships between properties. What is a state relative to one trait may be trait relative to another state. Your trait of generosity may produce in you a state of wanting to help me, and this new state, which is also a trait, may itself produce in you a state of wanting to know what my troubles are, and so on. So something can be both a state (produced by a higher level trait) and a trait (producing lower level states). In this paper, therefore, whether something is described as a "trait" or as a "state" will depend on the context.

A partial answer to question (a) then is that affective states are often dispositions, in various levels of a complex hierarchy. Question (b) remains unanswered: what distinguishes affective states? Is there something common to affective states and processes such as moods, emotions and desires that sets them apart from other mental states and processes such as perceiving, attending and noticing? Since some non-affective states are also dispositionally linked to action (e.g. believing), something else is needed to characterise affect. An easy answer often attracts people: identify affect with emotion. That might be acceptable (though possibly too narrow) if there were an accepted, empirically accurate, theoretically based, notion of emotion. There isn't. Moreover the suggestion ignores affective states that

are not emotional, such as wanting, liking, and enjoying some physical sensation. It will be useful to investigate mechanisms underlying affective states before returning to the question.

3 The Need For Design-Based Theories

I believe a proper analysis of the concept of an "affective" state or process must be based on a more general theory of the coarse-grained architecture of mind. Such a theory, should describe the main sub-mechanisms, showing how they are related and how their causal roles within the total system differ. Various functions for mechanisms and states can be distinguished, but only relative to the whole architecture. For instance, you can't describe something as a brake except in the context of an architecture that has at least one component whose speed is controlled by another sub-mechanism; and without the presence of a brake you can't describe a system as being in a state of "braking", though you can describe it as slowing down. Similarly, a collection of electronic switches cannot be described as a memory unless they are used by other parts of the system to store information. Like "braking", and "memory", many mental states, including affective states, presuppose a certain architectural richness. I shall argue that some aspects of the architecture have implications for communication.

A design-based theory about mental states can be contrasted with two more common kinds, semantics-based theories and phenomena-based theories. Semantics-based theories attempt to make sense of the structure of some portion of the lexicon of ordinary language (e.g. Ortony et al., 1987; Clore & Ortony, 1988; and Johnson-Laird & Oatley, 1989, unlike Oatley & Johnson-Laird, 1987, which is a design-based theory). Phenomena-based theories abound in the writings of psychologists: they assume that some particular kind of phenomenon can be intuitively recognized (e.g. emotional states) and then investigate other phenomena that are correlated with it in some way, e.g. physiological causes, physiological effects, behavioral responses, cognitive processes. The three kinds of theories are not incompatible: some authors combine two or more.

A design-based theory locates human mechanisms within a space of possible designs, covering both actual and possible organisms and also possible non-biological intelligent systems. Considering alternative possible designs leads to deeper theories, partly because the contrast between different design options helps us understand the trade-offs addressed by any one design, and partly because an adequate design-based theory of human affective states would describe mechanisms capable of generating a wide range of phenomena, thereby satisfying one of the criteria for a good scientific theory: generality. Such a theory can also demonstrate the possibility of new kinds of phenomena, which might be produced by special training, new social conditions, brain damage, mental disturbance, etc. An ambitious outline design-based theory that overlaps with those discussed below, is sketched in Johnson-Laird (1988).

There is *no* requirement that a good design-based theory of the mechanisms underlying affective states should fit colloquial concepts well, since our language reflects only our common presuppositions about how minds work, and these presuppositions may well be erroneous. Although ordinary language is often a suggestive starting point, it is also, as Simon (1967) warns, a source of muddle and confusion: " 'Emotion', like 'learning' and other traditional categories, refers to a mixed bag of phenomena, which may involve diverse mechanisms" (page 35n.) Dennett's paper 'Why a computer can't feel pain' (in Dennett, 1978) illustrates some of the incoherence in the concept of pain – the real reason why a computer can't feel pain. Related muddles pervade many words and phrases of ordinary language denoting mental phenomena, especially nouns, such as "consciousness", "sensation" and "emotion". When we have an architectural theory on the basis of which we can systematically derive good ways of classifying possible states and processes, we should expect to revise ordinary language with a taxonomy of affective states, just as the development of physics and chemistry gave us a richer and more systematic set of concepts for talking about different kinds of matter.

4 Towards a Design-Based Theory of Affect

What should a design-based theory of affect look like? Since there is no well-defined generally used concept of "affect" this is not a well-defined question. In the light of the preliminary analysis given above, I suggest that affective states are (a) dispositional states (long term and short term), (b) at various levels in a control hierarchy, (c) that include positive or negative evaluations of something and (d) have at least a tendency to produce motivational states, (e) which in turn have a tendency to produce behavior. (Some of the terms used will not be defined here: their ordinary intuitive

meanings, partially clarified in the rest of this section, will have to suffice for the purposes of this paper.) We'll see that there is no sharp boundary between affective and cognitive states, and affective states need not be states the agent is aware of since one can, for example, be angry or have an attitude, all unawares.

In what follows I shall describe features of an architecture in fairly non-technical terms. This should be treated as a preliminary specification, that still needs to be fleshed out with further implementation details. For example, I have already mentioned dispositional control hierarchies, without describing any mechanism that could produce them. There are obviously various possible implementations. For example, in a neural net, higher level networks could modify some of the weights in lower level networks; and in a rule-based mechanism there could be a hierarchy of production-systems where the actions of one system include creating or modifying rules in another lower level system. The neural network model has the advantage of appearing to be closer to the structure of the brain. The rule-system model has the advantage of great flexibility if it allows new collections of rules, corresponding to new attitudes, to be created easily, as opposed to simply modifying weights associated with a fixed set of rules. Other implementations would have to be considered in a full design-based study, along with detailed analyses of the implications of the differences.

States in which something is enjoyed, relished, found pleasant, admired etc., involve explicit positive evaluations and therefore tend to produce motivations to preserve or extend or reproduce the current state, or something that causes it, while states in which something is suffered, disliked, found painful, found unpleasant, despised, etc., involve negative evaluations and therefore tend to produce motives to terminate, shorten, or prevent the state, or something that causes it. The preceding uses of the word "therefore" presuppose a mechanism linking positive and negative evaluations with specific classes of dispositions to produce motivational states. (Having such dispositions may, in certain simple organisms, be all that such an evaluation amounts to.)

"Higher level" affective states, for instance, liking jazz or disliking people of a certain race, may be dormant dispositions until some cognitive process interacts with the general attitude to create a specific state directed positively or negatively at some instance. The new more specific affective state can either involve a motive to do something, or further relatively high level dispositions (possibly also dormant for a while) to acquire motives as a result of later events.

The motivational states themselves (wanting, desiring, etc.) may also be called "affective", though some people might wish to exclude motivations that do not involve a positive or negative evaluation of what is being done or achieved, as in compulsive desires like nail biting.

People often assume that affective states necessarily involve some kind of self-awareness. However, dispositions can lie dormant until their triggering conditions occur. Such states exist without being involved in anything one is currently conscious of, attending to or doing. For example, a person who dearly loves his or her children need not be constantly aware of that fact, or even thinking about the children: That would be obsession, not love. Love in its most common form is best classified as an attitude not an emotion, though most people, if asked, tend to classify it as an emotion.

The connection between affective states and the motivation that they can generate is only dispositional, since other things (e.g. guilt, masochistic enjoyment, fear of consequences, ethical standards, or distraction of attention) can override or suppress any particular tendency inherent in an affective state. The same goes for the link between motives and behavior: you may want to do something, but decide not to, or believe you don't have the opportunity, etc.

The fact that affective states are related to action only indirectly, sometimes via several levels in the hierarchy, makes possible great flexibility of response to the presence, promise or threat of harm or benefit. The flexibility is achieved by allowing cognitive states to modulate the way affective states control behavior. In particular, when conflicting motives are generated, cognitive processes can play a role weighing up the pros and cons of the options, or devising a compromise plan, or selecting a means that achieves one goal while doing as little damage as possible in relation to the other.

So cognitive states like believing, imagining, attending, perceiving, pondering, planning, etc. also influence motivation and behavior, but not normally on their own: they require affective states.

5 Cognitive Reflexes

The rich array of ways in which communication can trigger dispositions associated with affective states is part of the general flexibility of control achieved by allowing interacting affective and cognitive states. This is one of the ways in which intelligent agents differ from organisms that are dependent primarily on innate reflex responses.

Sometimes, however, there is no time for evaluation, reasoning, planning, etc., so even intelligent agents often need highly trained and very rapid responses to new information – "reflexes" in the non-technical sense. These are cases where a *cognitive* state directly produces (internal or external) action, bypassing processes of decision-making, production of motives, formation of intentions or plans, etc. Many skilled performances depend on this kind of shortcut mechanism linking cognition and behavior. It is not worth arguing whether the pre-scientific concept of "affect" applies to these processes or not. The important thing is that they have a role that can be described precisely in terms of the architecture within which they function: they short-circuit, or over-ride the normal "intelligent" control hierarchy, trading risk of error for speed.

Reflex links between cognition and external action can be directly triggered by verbal communication, but that is rare: they seem to be mainly designed to cope with fast moving entities in the physical environment. However, internal trained reflexes play an important role both in production and in understanding of speech.

All this is still very vague. In order to be able to design a machine that can model the affective states involved in communication, or one that takes account of potential affective states in deciding what to say to other agents, we need to know more about the nature of these various dispositional states, the control hierarchies they can participate in and how they interact with new information to generate and modify both internal processes and external behavior.

6 Control Hierarchies and Ease of Change

One purpose of communication is to produce or change affective states. For some states this is much harder than for others. If affective states include attitudes, moods, emotions, desires, impulses, and also flushes of embarrassment or sudden pain that produces involuntary wincing, then different time scales are involved as well as more or less indirect links with behavior. Moreover, not all of these states are subject to communicative influence (except perhaps through hypnosis). This variety of time scales and links with behavior is to be expected in a dispositional control hierarchy. If state S1 involves a disposition to produce and maintain state S2 or S3 or S4 depending on circumstances, then S1 must be a longer lasting state than S2, S3 or S4. Compare this with non-hierarchical relations: if S1 merely *initiates* the other states, which then don't require S1 to keep them going, then they may last longer than S1. X's intense but short term infatuation for Y can produce long term hatred if Y betrays or callously rejects X.

Closely related to different time scales are different degrees of changeability and ease of external control. Many painful or pleasurable sensations can be turned on or off by the addition or removal of the appropriate external stimulus. Some desires are externally triggered and turned off (e.g. finding half a worm in the apple you are enjoying), but many are not directly under external control and often persist till satisfied or forgotten. Some emotional states can come and go fairly quickly, but many are deeper and more persistent. Certain moods, like depression, happiness, calmness, and optimism, are persistent states with the dispositional power to color and modify a host of other states and processes. Such moods can sometimes be caused by cognitive events with semantic content, though they need not be: their causes can, for instance, be chemical. Similarly, their control function does not require specific semantic content, though they can influence cognitive processes that do involve semantic content. Sometimes depression reduces the likelihood of believing any positively evaluated hypothesis and increases the likelihood of adopting negatively evaluated ones. Thus if X is depressed he may be less likely to believe Y's assertion "You are sure to do well in your examinations", even though X's depressed state has no cognitive content linking it with examinations. So a machine that tries to cheer you up may fail miserably if all it can take account of are your beliefs and attitudes, and not your current mood.

There is a use of "mood" as in "I'm in the mood for X (or to do Y)" that is close to a desire or inclination and is not a global semantics-free control state. Although not all such moods can be influenced by communication, some can, for instance bad tidings that remove a mood to go to a party. I suspect that there is no general category of moods. Rather what sort of state the word "mood" refers to depends on what other words it is combined with: depressed mood, energetic mood, contented mood, mood for dancing, mood for a cup of coffee, are all different kinds of temporarily dominant state, some of which are no more than desires.

Attitudes are longer lasting than moods, yet more changeable than personality traits, and more semantically directed than either. They are persistent high level clusters of beliefs and preferences concerned with the object of the attitude (a political party, a style of art, a person, etc). Many different (and potentially inconsistent) attitudes can coexist in one person and lie dormant until some specific item of information turns up that interacts with one or more of them and generates more specific, shorter-lived affective states.

Thus a prejudiced attitude to people of a certain color may make one more likely to believe stories about their bad behavior even when there is little evidence, and can make one both want harsher punishment for them, and want it more intensely. So, among the internal states that an attitude can influence are other attitudes, beliefs, the contents of desires, and the intensity of desires. Sometimes such attitudes can be changed by new information (e.g. you stop liking someone when you learn that he has been deceiving you for a long time), but often general attitudes are very hard to change, a point that skillful communicators, human or machine, need to understand. (Social psychologists have explored some of the conditions under which such changes occur, but I am not aware of any detailed models of the control hierarchies and cognitive mechanisms involved.)

Unlike global moods, many attitudes, including potentially conflicting ones, can co-exist, as dormant dispositions, in the same individual. This means that a communication that is intended to do something about changing a manifest attitude risks triggering some unwanted reaction from another hitherto unrevealed attitude. Since dormant attitudes usually have no visible manifestation, the only ways to check which attitudes must be allowed for are either to do thorough investigations of the people one talks to, or to use prior knowledge about their attitudes. The former option is impractical in most communicative contexts, and the latter is generally available only if one knows the person very well. We are therefore inevitably driven to rely on stereotypes. Intelligent machines will have to do the same.

One aspect of being part of a culture is learning about the attitudes likely or unlikely to be found in different sorts of people in that culture. A person (or machine) that lacks this culture-specific know-how can slip up badly, as sometimes happens in advertisements in foreign markets, or television propaganda prepared in one country for broadcasting to another.

Personality and character traits are generally even more durable than attitudes, and their relationship to action is very indirect, usually via attitudes that flow from them, and which in turn produce beliefs, motives, emotional reactions, intentions, etc. and thereby (sometimes) plans and, ultimately, actions. Communications intended to alter or mould personality are to be found in psychotherapy, which currently appears to be more an art than a science. It is sometimes supposed that computers might one day be psychotherapists (perhaps with more patience and lower fees than human psychotherapists). However, we need far better theories about the nature of these high level dispositions and the mechanisms through which they are formed and modified if this is to be done well. Apart from psychotherapy there are many forms of personal counselling to which similar comments apply.

Probably the most common communications intended to shape or change personality or character are to be found between adults and children at home and in schools. Judging by the state of the world, it would appear that computers could hardly do worse than most human adults.

Communication is often more directly aimed at initiating or changing behavior than at changing personality. Persuading and giving advice both aim to influence the behavior of the hearer, and this is normally much easier than changing personality or attitudes. If you know that a person has a particular attitude this can be used to generate behavior, simply by giving information that will interact with that attitude ("There's a spider creeping up behind you"). However, there are pitfalls: if the communicator does not know about all the relevant attitudes in the hearer, or does not understand the differences between relatively low level, easily changed desires, preferences, plans or intentions, and relatively high level enduring attitudes or traits then the communication may, at best, fail, and at worst be crass and insensitive, e.g. "If you think the manager doesn't value your work enough, why don't you try flattery, or offering sexual favours: it has worked for others."

All this suggests the following very crude design-based taxonomy of types of affective states (addressing different issues from the taxonomies proposed by Ortony et al., and Johnson-Laird & Oatley, though similar in outline).

1. Dispositional
   1.1. Long term, and hard to change by communication
        E.g. Personality, character, temperament

1.2. Medium term, and easier to change, though still hard
Co-existing, enduring, mostly dormant, clusters of sub-states: e.g. ideals, principles, attitudes, interests, desires,
Global states: moods (only one of which can exist at a time), e.g. depression, undirected happiness.
1.3. Short term, and still easier to change
E.g. Emotions, pains, pleasures, irritations, inclinations.

2. Episodic (with dispositional elements)
2.1. Very short term, some externally controllable, some not
E.g. Pangs, twinges, urges, startles, orgasms, "reflexes" (innate, trained, etc.)
Other involuntary responses tied to feelings, e.g. wincing, gasping, smiling, along with the *urge* to do these things (which is sometimes suppressed).

This is not intended to imply that there are only three or four control layers in the hierarchy of affective dispositions. There is probably no fixed number: the number of intermediate dispositional states between a disposition and the behavior that it influences may be different in different cases. Some architectures would allow new layers to be added on the basis of learning.

7 Cognition, Affect and Architecture

What has been said so far implies that a theory of communication has to allow for several kinds of relationship between affective states and cognitive states with which they interact. For example:

- A cognitive state may interact with a dispositional state to produce or modify or terminate a manifestation of the disposition in lower level dispositions, or behavior.

  Learning that a spider is approaching can produce a desire to move away. Learning that most spiders are not at all dangerous may change the manifestations of one's fear of spiders, even though the fear persists.

- A cognitive state or process may modify an existing affective state, e.g. by increasing or reducing its intensity, or redirecting it.

  Learning some new facts may increase or reduce anger, or redirect it towards another target.

I claimed earlier that a theory of affect requires a design-based theory of the architecture of intelligent agents. Different architectures permit different sorts of affective states and processes, just as the variety and relations of sub-mechanisms in an engine determine the kinds of dynamic states it can be in. Intelligent agents (e.g. mice, monkeys, human infants, human adults, and perhaps one day intelligent machines) can have different architectures, including different cognitive sub-mechanisms, and this limits the kinds of affective states they can be in. The sub-mechanisms within an architecture will include some that determine representational capabilities, which depend for example on the kinds of structural variation that can be supported. This will make a difference to the kinds of semantic content available in affective states: could a goldfish wish its mother were still alive, or have any other affective state whose cognitive content used the concepts of *mother, being alive,* etc.? Not if the available representational apparatus cannot support the representation of absent individuals, non-existent states of affairs and abstract relationships, for example.

Even when mechanisms are available that in principle could have this representational power, they may not be embedded in an architecture that gives them this functional role in affective states. Functions of available sub-mechanisms depend on relationships with other sub-mechanisms: a pedal in a vehicle would not be an accelerator if it were not connected to an engine in such a way as to alter its speed. So certain descriptions and questions that make sense for one organism or machine may not make sense for another: could a goldfish weigh up present pain against future pleasure? Could a mouse desperately want its children to do well? Not if they don't have architectures supporting a sufficiently rich collection of motivational sub-processes.

If humans have different cognitive architectures at different stages of development then this needs to be taken into account when communicating with them. This is grasped intuitively by some parents: they would not dream of trying to dissuade a very young child from eating sweets by talking of the sufferings likely to follow in old-age. This is a

pointless exercise if the child lacks the representational apparatus required to conceive of himself as an old man. Even if suitable representational apparatus is present, it may be pointless for other reasons: the child might not yet have developed an architecture with the functional differentiation required for sub-processes weighing up short term and long term costs and benefits and deciding accordingly. (Some people never seem to develop this.)

I suspect that the vast majority of parents do not have any insight into these matters, though the patient and sensitive few manage without any explicit theory, by using genetically determined and socially learnt communicative strategies enhanced and modified by feedback from the child that indicates whether their attempts at communication are working or not. Intelligent machines lacking adequate knowledge about the mechanisms in people would also have to use feedback, but this will often require powerful perceptual mechanisms.

8 H. A. Simon's Theory

A system that can support a collection of simultaneously active dispositional states that interact with other states needs a coarse-grained parallel architecture that makes possible coexisting, intercommunicating processes. By a coarse-grained architecture I mean a division into major co-existing, functionally distinct, components that can influence one another. This functional division does not imply physical distinctness, as shown by the way in which a collection of interacting *virtual* processes can be implemented on a time-shared computer. This coarse-grained parallelism is different from the fine-grained parallelism of neural-nets studied by connectionists, though neural nets can support coarse-grained parallelism, if different sub-nets perform different tasks.

I know of no theory in psychology, AI or cognitive science that begins to address the full range of architectural requirements, including the relationships between the dispositional hierarchy and mechanisms of perception, reasoning, learning, attention, motor-control, self-monitoring, and so on. There have been some useful initial moves, however. One influential theory is by H.A. Simon (1967). He attempted to address criticisms of information processing theories made by Neisser (1963), by presenting a computational theory according to which all human thinking involves "an intimate association with emotions and feelings", and "almost all human activity, including thinking, serves a multiplicity of actions at the same time."

Simon sketched a fairly sophisticated computationally inspired architecture, aspects of which are often re-invented, and used it to give an analysis of emotions as states in which ongoing activities are interrupted or disrupted (an idea that others before and after him have espoused). He did not define the architecture in great detail, and, as far as I know, no attempt has been made to produce a detailed implementation, which would require a solution to many hard problems in AI, including modelling perception, learning, reasoning, planning and motor control. I shall first summarize Simon's theory and then describe some recent developments, including a version that allows emotional states to have the *potential* for interruption without requiring actual interruption. An important distinction will be made between interrupting behavior and interrupting attention. The mechanisms proposed will then be related to the task of understanding how communication can interact with affective states, and some of the implications for computer models and interfaces discussed.

Simon's work is based on several fairly obvious key ideas, some already discussed. Human beings (and other animals) have multiple independent sources of motivation (including social motives). In particular, new motives can arise in response to changing external situations or changing internal states. Humans are also resource-limited (both relatively slow and largely serial), which is a problem in view of the fact that the environment (including other agents) is complex, partly unpredictable and often fast moving, so that constant, asynchronous, monitoring is required in order to detect unforeseen dangers, obstacles or opportunities. He outlines some of the control issues, and suggests suitable mechanisms, inspired in large part by developments in computer science and AI, including software techniques for generating new sub-goals at run time, techniques for queueing and scheduling processes, techniques for forming plans in order to achieve goals, techniques for assigning priorities and resolving internal conflicts, and techniques for generating and handling interrupts. He claims that these processes, especially the interruptions resulting from new information, account for the states we typically call emotional.

Although he stressed the importance of multiple, changing, motives he did not say much about where they come from, apart from postulating a collection of drives whose intensity is a function of the length of time of deprivation. However, it is clear this does not cover all cases. Additional sources of motivation are attitudes and other high level dispositions discussed previously, interacting with incoming information. More generally, goals that produce new

actions or interrupt or disturb existing actions can arise from new information coming from: (a) the environment, (b) new internal physiological needs, and (c) cognitive processes in which associations are triggered, for instance suddenly remembering something that implies that what you are doing has a serious danger. Case (a) would include communications from others. Case (c) could be thought of as communication with oneself.

Simon notes that monitoring of other agents is an important feature of human processing, and a major source of emotional states. Human beings need to be particularly good at detecting and interpreting the behavior of others. Moreover, because people are so complex we often misinterpret what is happening. Hence, two important kinds of learning are required: one that increases sensitivity to others and thereby the likelihood of emotional disturbance, another that improves one's ability to anticipate, forestall, or cope with interrupting stimuli and therefore reduces the likelihood of emotional disturbances in social interaction. Which of these predominates will vary from individual to individual and over time within one individual. He suggests that adolescence is often a peak period for the dominance of the first over the second.

Many human beings find such learning very difficult, and consequently never learn to get on well with others. Some manage to cope only with others in their own, rather restricted, sub-culture. These are pointers to the great difficulty of giving machines such knowledge. Perhaps, like people, intelligent machines in the foreseeable future will be able to interact well only with members of a restricted sub-culture. Unless they have powerful learning capabilities of the kinds mentioned by Simon, along with more general inductive and abductive capabilities, they could not be expected to deal with a wide variety of individuals. However, our analysis will show that such learning is very difficult indeed.

9 The Global Signal Theory of Emotions

There are at least two different more detailed lines of development of Simon's theory. The first can be found in the "global interrupt signal" theory of Oatley and Johnson-Laird (1987; Johnson-Laird, 1988) and the second, in the "attention filter penetration" theory of Sloman and Croucher (1981; Sloman, 1985b; Sloman, 1987).

Oatley and Johnson-Laird (like Sloman, 1978 and 1981b) postulate a hierarchy of parallel processors all asynchronously dealing with different tasks, but ultimately managed by some kind of "top level", or "central", control system. This sort of architecture makes it possible for processors that detect problems to send out signals that propagate through the system. Oatley and Johnson-Laird suggest that the global effect of such signals sets the whole system into a new state when a problem occurs with an ongoing plan or activity. The new state may interrupt ongoing plans. The spread of such signals, they claim, is essentially what an emotion is, though many emotions involve additional phenomena produced as a result of this disturbance.

They claim that the global signals have no "propositional content". I suspect that this does not express precisely what they mean. Propositional content would require the use of bivalent predicates applied to arguments, logical connectives, quantifiers and the like. It seems unlikely that most organisms use internal states with a propositional syntax. But there are many other ways of expressing semantic content, as I have tried to show, for example, in Sloman (1971; 1985a). It is probable that many or most animal and human mental states have *semantic* content but no *propositional* content. If so, saying that emotion signals have no propositional content does not distinguish emotions from other states, including some cognitive states using non-propositional representations. I suspect that what they really mean is that the signals have no *semantic* content — they simply have a *control* function, without including elements that refer to, depict, or describe objects, events, processes, relationships etc. Related points are made in Sloman (1989) concerning the way perceptual processes can sometimes directly produce control signals that change behavior, without always going via descriptive databases.

One way in which communication could produce an emotional state, according to Oatley and Johnson-Laird is by providing information that reveals a new obstacle in an ongoing plan. This would trigger the global signals that indicate a need to interrupt or perhaps modify the plan. It is not clear what the full range of possible interactions between communication and affect would be on this theory, partly because the theory does not attempt to account for higher level dispositional affective states.

The global signal theory mainly stresses episodic phenomena as constituting emotions, i.e. states in which signals are actually generated and disturb or modify behavior. Our alternative approach, the filter penetration theory, described below, stresses dispositions, tendencies, and the like: strong jealousy need not actually divert attention and disturb

or reorganize other processes, if, for instance, some unrelated activity is temporarily engrossing; but, given half a chance, the green gremlin will emerge to color thoughts, decisions and plans. So the theory allows temporarily dormant emotions (if strong jealousy is an emotion). It also permits more complex and subtle interactions between information communicated and resulting affective states, by stressing the impact of emotional states on the potential to disturb current cognitive processing (attention) rather than on more general disturbances. I shall now explain this in more detail.

10 Insistent Motives and Filter Penetration

The filter penetration theory is similar to the global signal theory, but requires extra architectural complexity (implicit in Simon's paper). Cognitive processes can sometimes involve switches of attention without interrupting ongoing actions. For example, while driving on an important errand one can think about other things or enjoy the scenery before reaching one's destination. This requires an architecture that allows several high level concurrent processes, including monitoring the environment whilst controlling actions. Some activities require a strong focus of attention: perceptual attention especially, but also sometimes concentrated thought processes are needed for dealing with a stream of difficulties. If one's thoughts wander whilst listening to complex instructions the consequences could later be disastrous. This suggests a need for variable-threshold interrupt filters to control the ability of new motivators, thoughts, or percepts to disturb or divert attention. An example would be the soldier or football player who does not notice an injury that occurs during a battle or an important match, even though the pain would normally divert attention. (Attention filters need not be separate mechanisms: all that is required is that the overal architecture ensures that the potential for new information to interrupt or disturb ongoing perceptual or thinking processes is highly context sensitive.)

I am using "attention" in its ordinary non-technical sense: what one is attending to in this sense is what one is currently thinking about, looking at, taking care over. (There is a complex family of concepts related to attention, whose full analysis is not possible here.) The filter concept makes an architectural assumption that some activities use cognitive and physical resources that are limited, and that in some situations diverting them is either dangerous or likely to sabotage some important goal. Attention filters provide protection against this. Variable-threshold filtering allows the level of protection to depend on the importance and vulnerability of the current task.

This mechanism is important only when interruption or diversion of attention would undermine important activities, which is not necessarily the case for for all important tasks, for instance those that are automatic or non-urgent. Keeping the car on the road while driving at speed on a motorway is very important, but a skilled driver can do it while thinking about what a passenger is saying, whereas sudden arm movements could cause a crash. However, in situations where speed and direction of travel must be closely related to what other cars are doing, even diverting a driver's attention could be dangerous. So our theory's focus on diverting or interrupting cognitive processing is different from the focus in Simon and the global signal theory on disturbing or interrupting current *actions.* We'll see that some emotional states involve a disposition to divert attention without necessarily disturbing any action.

All this has implications concerning dimensions of variability of affective states. Simon mentioned drives, whose intensity increases with deprivation until a threshold is exceeded whereupon they interrupt ongoing activities. This does not account for all the phenomena. I have argued (e.g., in Sloman, 1987) that it is necessary to distinguish several dimensions along which motivational states can vary, including insistence, importance and urgency.

"Insistence", of a motive is the basis of its ability to generate emotional states. This depends on its disposition to divert attention from other activities. So insistence is defined as the propensity to get through attention filtering processes and thereby divert and hold attention. A highly insistent motive will divert attention unless current activities set the filter threshold level very high. Even then, the potential for diversion may persist. This is different from other dimensions of variation of motives, such as "subjective urgency" (perceived or supposed time remaining before it will be too late) and relative "importance" to the agent (which may depend on more or less sophisticated comparisons of means, goals and higher level ideals and aspirations). These depend on the agent's assessments, which may, of course be mistaken: subjective urgency may or may not correspond to actual urgency, and the agent's view of long term importance of a goal may prove quite mistaken.

Although insistence is different from these other dimensions, it should, in a well engineered system, bear some relationship to them. Thus the mechanisms that assign an insistence level, or which do the filtering, should tend

to ensure that only relatively important and urgent topics will be allowed to divert attention. But this must happen without complex cognitive processing of the kind that would divert resources! So insistence is assessed on the basis of relatively quick and simple (possibly learnt) heuristics, which could be erroneous in some cases.

High insistence of a new motive can cause attention to be diverted without actually causing any current action to be interrupted or disturbed. For example feeling very hungry can make a driver consider whether to stop for a meal, without interfering with the driving. Interruption might occur if the new goal is judged more important than, and inconsistent with, the purpose of the current activity, or if the new one is judged to be very urgent (although not necessarily very important) whereas the (more important) current activity is not time-critical and can be temporarily suspended: for instance stopping for a meal because one has plenty of time before the important meeting. Alternatively a highly insistent motive that gets through the filters can be considered and then rejected as relatively unimportant, without interrupting any important current action. So insistence, the propensity to divert attention, is not the same as a propensity to interrupt current actions, except those that require full attention.

None of these kinds of variability of affective states (insistence, urgency, importance) need be measured on ordinal or other uni-dimensional scales. Relative importance, for instance, has a number of facets relating to different needs, ideals, preferences, plans, etc. and there may be only a partial ordering. Moreover in some cases only descriptive assessments make sense: "the importance of this goal is that it contributes to such and such objectives and supports such and such ideals". In these cases the process of selection between competing goals could involve very complex reasoning.

There is a lot more to be said about the hardware and software mechanisms that might implement attention filters, assign insistence levels, and control thresholds. Different designs will have different implications, and only some will be fully consistent with how human beings work. In some designs the filter mechanisms may be only partially effective, so that although a certain state with high insistence does not get through the filter it can nevertheless reduce the efficiency or accuracy of ongoing cognitive processes. This could, for example, be a consequence of a design based on activity propagation through networks, where filtering is a matter of degree. We'll ignore such details and turn to some of the implications of the architecture so far described.

11 Insistent Motives and Loss of Control

In certain situations, insistent motives (and other states) have a strong tendency or disposition repeatedly to get through the filter, divert attention and possibly interfere with other ongoing processes, even if they have already been considered and rejected or even adopted for future action, for instance when you can't put the worry, shame, or fiendishly clever scheme for revenge out of your mind; or rather you can do so only when engaged in some other powerful attention-grabbing activity (such as sex, gossiping, or watching a film). This sort of state seems to be common to what we normally think of as strong emotions: states in which we are "moved" and at least partly out of control, or would be partly out of control if there weren't a current activity that makes the filter thresholds unusually high. They are dispositional states, where the disposition may or may not be realized in actual diversion of attention.

Insistence, on this analysis, is a *dispositional* state: the highly insistent motive or thought need not actually get through the filter and interrupt anything. Even if it does get through it need not actually disturb any current activity. I suggest it is this strong *potential* for such disturbance and diversion of attention that characterizes many of the states we describe as emotions. Such states can exist whether or not attention is actually diverted, and whether or not actions are thereby interrupted or disturbed. Thus, like jealousy, anger (in the form of a very insistent desire to harm someone because of something he is believed to have done that is strongly negatively evaluated) can persist even though something else occupies attention for a while. During that time there is no diversion of attention or disturbance of any action. Dormant dispositions include such emotional states.

As explained above, (and in Sloman & Croucher, 1981; Sloman, 1987) the assignment of interrupt priorities in a resource-limited agent cannot depend on complex cognitive processing, for that would defeat the purpose of the interrupt filter. Such filtering therefore depends on "quick and dirty", potentially fallible, heuristics. So one of the tragedies of resource-limited agents in complex, fast changing, only partly knowable environments, will be that insistence is imperfectly correlated with subjective and objective importance. What tends to grab and hold attention is therefore not necessarily always in our long term interests: one source of inspiration for much great literature.

A detailed computational model of language understanding able to cope with stories will have to include some understanding of these mechanisms. Otherwise it might fall into the 'rationalist' error of assuming that human agents always act and think in accordance with their judgement of relative importance of various options. Some philosophers have even suggested that such rationality is a defining characteristic of having such states as beliefs and desires. However, our analysis of design options for resource-limited agents shows that this is wrong. The philosophers have not analysed design requirements for intelligent agents with resource limits (compare the discussion of the 'intentional stance' in Dennett 1978, pages 7–12). Nevertheless, people have an implicit comprehension of these mechanisms, as shown by their ability to understand stories of the sort hinted at above. Unfortunately, their more explicit theories usually oversimplify by assuming insufficiently sophisticated architectures. (An accurate computer simulation would make the same mistake!)

Since insistence, as I have defined it, is a matter of degree, the theory implies that there are only differences of degree between emotional and non-emotional motivational states. It also implies that there is much in common between emotional states and those cognitive states where a particular thought or something like a remembered experience or tune has high insistence, but does not involve any particular motivation or positive or negative evaluation.

Some people would restrict the word "emotion" to the case where the disturbance actually occurs. This seems to be the view of Simon, though Oatley and Johnson-Laird suggest that an emotional tone can persist without actually interrupting any action. I think they require this tone at least to be noticed by the agent, whereas on our theory the state can persist without any effect on consciousness. It is not clear whether they would require attention to be diverted. A theory of emotion that requires *actual* interruption or diversion would imply that if X's anger is temporarily put completely out of mind by the need to deal with some unrelated emergency, then X would temporarily not be in a state of emotion. Someone holding such a theory would either have to say that X was temporarily no longer angry or else that being angry is not necessarily an emotional state.

The important point for us is not how to use ordinary language, but to note that certain designs make it possible for a certain potentially disturbing dispositional state to continue to exist in a 'dormant' form, during an interval in which some other high priority state dominates attention and uses a high filter threshold to prevent the disturbance from occurring. Quibbling about whether the anger or the emotion 'really' exists during this interval is fruitless. The important point is that there is a state that persists, and can manifest itself as soon as the diversion is over, and for proper descriptions of the human mind we need a vocabulary for describing such states, whether or not it accords with ordinary language.

An implication for a working machine communication system is that an individual who appears not to be angry, jealous or afraid, may actually be in a strong, but temporarily dormant, emotional state, masked by some temporarily engrossing distraction. In such a state, badly chosen words or phrases might alter insistence levels or filter thresholds, triggering the emotion to manifest itself and thereby defeating the purpose of the communication.

The theory also implies that several different emotional states can co-exist, since many different things may simultaneously have high insistence with the potential to divert attention. In fact one emotional state can cause another, even while the first persists.

12 Consequences of the Theory

The extra architectural complexity postulated by the filter penetration theory beyond what is required for the global signal theory has several consequences including allowing different kinds of learning. For instance one kind of learning made possible by the presence of filters is discovering how to map goals and activities of various degrees of importance onto appropriate filter thresholds so that when they are active they will not be interrupted by other things of lesser importance. Another kind of learning, more like training of a skill, is modification of the mechanisms or rules for assigning *easily computed* insistence levels to new motives and other potential sources of disturbance. Too low an insistence level can mean that a matter of life and death fails to interrupt an activity of only medium importance. Too high a level can cause attention to be diverted by trivia during important tasks. Resource limits imply that all such strategies will be dependent on fallible heuristics, although if there is enough regularity in the environment, the heuristics can be improved with experience.

Whereas Simon postulated a queue of pending plans or goals, the current proposal allows several different kinds of queues or information stores containing motives, for instance one containing motives awaiting consideration as

to whether they should be accepted or rejected and one containing accepted motives that have not yet been assigned temporal priorities or conditions for action. Information stores are also required for current actions, some being pursued in parallel, and some temporarily suspended for one reason or another. This architecture allows for idle wishes, and other affective states that generate no plans or actions.

Different kinds of affective states would be related to whether anything is in these various stores, what their content is, how they relate to one another (e.g. inconsistent desires) and what their causal powers are (e.g. whether they tend to get through the interrupt filter again and again, and whether they influence other processes, as optimistic and pessimistic moods do). There is still much work to be done to clarify this theory, including analysing the similarities and differences between motives and other attention distractors.

The theory allows that a new piece of information interacting with a dormant attitude, or with the fact that there is something one has always wished for but thought could not be achieved, can raise the insistence level of a motive, thereby generating great excitement, even if there is no need to interrupt or disturb any current action. For instance, one might learn unexpectedly that a long-standing wish, on which one had given up hope, could be achieved by taking action on the following day, for which one had not yet formed any plans. This example shows another way in which emotional states need not involve actual interruption or disturbance of any action or current plan, though the potential is there, insofar as they would have caused a disturbance had the current actions been incompatible with the newly awakened goal.

13 Some Computer Modelling Goals

Before discussion the implications further it will be useful to distinguish different purposes for which a model of affect might be used. There are different sorts of goals in designing systems intended to communicate with human beings, including (in order of increasing difficulty):

G1 Getting machines to understand human utterances in the ways that people do, for various practical purposes, such as answering questions or obeying commands.

G2 Getting machines to achieve various goals by communicating with human beings, such as teaching them, advising them, helping them solve their emotional problems, etc.

The distinction between **G1** and **G2** is orthogonal to the distinction between the scientific attempt to model and explain human capabilities and the engineering goal of building something potentially useful: both engineers and scientists may adopt either **G1** or **G2**, though they will pursue them in different ways. The scientific modelling task could be characterized as:

G3 Getting machines to simulate human affective states and processes in detail, including those involved in communication.

In what follows I shall concentrate on the practical goals **G1** and **G2**, while showing how they differ in their requirements for a deep scientific understanding of the human mind. In particular, if ordinary language is based, in part, on erroneous theories, that implies that **G1** and **G2** have different requirements. A machine designed for purpose **G1**, or an organism that can do **G1**, needs only a system of concepts based on the same set of possibly erroneous presuppositions as ours. On the other hand, a machine (or organism) designed for **G2**, i.e. one for which communication is not just an end in itself, but is a means to such ends as teaching, giving advice effectively, and more generally communicating with people about various topics without upsetting, confusing or misleading them, will need to have, or be based on, a good (i.e. non-erroneous) theory about how minds work. It may need a whole family of theories, if different people's minds work partly in different ways (e.g. people in different cultures, children and adults, men and women). In short, a machine to achieve **G2** reliably would either have to be far superior to most people in its understanding of the human mind, or else have a collection of heuristic rules based on a superior theory. For reasons indicated below, it is unlikely that a really good set of rules could be learned empirically. Moreover without a good theory underpinning such rules, the designers couldn't understand how such a machine worked.

Having a (correct) explicit theory about how other minds work might not be very helpful in practical contexts, if deducing from it how best to communicate took a very long time. Shallow heuristics might work much faster. However,

a comprehensive set of heuristics would have to take account of a huge variety of combinations of mental states that might be dealt with more economically by a good model. The deeper theory would enable one to cope with a wider range of contexts and communicative goals, so ideally an intelligent communicator should have both, even though the more explicit theory would be useful only when there's time for complex reasoning. (Compare knowing multiplication tables and understanding the principles from which they are derived.) Few people have both kinds of knowledge about affective mechanisms, and most have neither, except for a smattering of more or less shallow heuristics. Because we lack good theories of affective processes it will be a long time before we can program either kind into a robot nursemaid, or even provide it with the wherewithal to learn enough itself.

14 Implications of the Theory for Communication

Several implications for processes of communication and the design of models of communication have already been mentioned, especially the potential for unanticipated responses when a communication interacts with one or more dormant dispositions. This is a source of difficulty for goal **G2**. In addition, the attention filter hypothesis implies that when a person is in an emotional state, communication with that person may be made more difficult because of the competition for his or her attention. In order to communicate effectively in such cases special devices may be necessary to gain and hold the hearer's attention. This could be important in designing human-machine interfaces intended to deal with safety-critical situations. If human beings are likely to become emotionally disturbed in such situations, then strong action may be needed to ensure that they attend properly to instructions, etc. Conversely, when a serious problem arises and the human concerned is in a relaxed and unemotional state, it may be necessary for the machine to take action to generate an appropriate state of anxiety, or concern, etc. in order to ensure that the problem receives the person's full attention. This is obviously the sort of factor that has (wittingly or unwittingly) influenced the design of alarm systems. However in contexts where just making a disturbing noise to get people moving is not enough, for example because detailed instructions have to be given, it will be necessary for the machine to adopt more sophisticated means of generating an appropriate level of concern, possibly even using some irrelevant but attention-getting or anxiety-arousing stratagem to start with.

In teaching situations more subtle techniques are sometimes required. These include such things as the use of entertaining examples or relating what is taught to something the pupil already knows and is interested in. Detecting situations when these techniques are needed, and inventing strategies to suit particular individuals, will be beyond the likely capabilities of intelligent tutoring systems for some time to come.

Another implication of the theory is that new information provided in communicative acts can interact not only with current plans and actions, but also with higher level attitudes, motives, plans, and actions that are queued, pending, or suspended. A new piece of information can interact with an attitude to generate a new motive, or can reveal an unexpected opportunity to achieve something you were not planning to do but would very much like to do. These processes can interfere seriously with what the communication was intended to achieve.

Finally, the filter penetration theory offers the germ of an account of pain and pleasure: both include attention-holding and attention-grabbing capabilities: the one combined with disposition to want the current state to continue and the other with a desire for it to end. Some states of pleasure and pain, bound up with perceptual processes, also include semantic content, such as reference to a bodily location and what is happening there.

A comprehensive theory of human emotions would probably have to incorporate both the global signal theory and the filter penetration theory, along with, no doubt, other mechanisms.

15 Are Computer Models Possible?

A major difference between the theory proposed here and most other theories of emotions, is that the latter are intended primarily as accounts of human and animal emotions, whereas the filter penetration theory emerges from a general theory about design requirements and possibilities for intelligent systems, whether biological or artificial. So this theory regards as peripheral certain aspects of human emotion (e.g., facial expression and physiological reactions) which other theories treat as central. They are peripheral because essentially similar emotional states, with similar

social implications, could occur in alien organisms or machines lacking anything like our expression mechanisms. Could such states occur in computers?

Although Simon (1967) (and some of my own early papers) suggests that the kinds of processes we are concerned with here are all computational, I now think that the notion of mechanism is more general than the notion of a (Turing-equivalent) computer (see Sloman, forthcoming). For example, moods might be influenced by chemical, or hormonal, processes in ways that would not naturally be described as "computational" except in such a loose sense as to cover almost any process, thereby making the description trivial and uninformative. What remains in question is whether the kinds of affective states that can be produced by such non-cognitive, non-computational, non-semantic mechanisms are very different in kind from those that are influenced by cognitive processes. Is a mood of depression or euphoria that is produced by chemical processes a totally different state from the depression produced by repeatedly failing to pass your examinations or the euphoria produced by passing with distinction? Or does the latter kind of process actually use the former? Even if certain mechanisms that don't look like computations are used in human beings, functionally similar states might be produced in robots implemented using standard computing techniques.

Certain human affective states are intimately bound up with our physiology and the presence of massive and constant feedback from physiological processes to central monitors. Organisms or machines without this kind of architecture, e.g., organisms whose bodies had relatively few built in proprioceptive sites, would not be capable of having those of our affective states in which a bodily gestalt plays a key role e.g. physical disgust or sorrowful weeping. Nevertheless, such alien beings (or intelligent machines), if they had the other mechanisms I have described, could have closely related states: for instance they could find certain information disturbing, extremely objectionable and hard to put out of mind, and the strong desire to take remedial action might also have high insistence. Thus even without our physical reactions they could share significant aspects of emotional states such as disgust or anger. Moreover, like human mathematicians they might feel great disappointment at the news that some mathematical conjecture had been refuted: an affective state that need have no accompanying bodily reaction (outside the brain).

Doubts about the possibility of exact replication of emotions might be based on the common assumption that emotional states are necessarily bound up with various kinds of external expression tied to our physical structure: e.g. smiling, weeping, grimacing, wincing etc. This causal linkage certainly occurs in human beings and in some cases may be part of a primordial communicative mechanism, like mating dances in birds, chemical signals in moths, etc. But there is nothing intrinsic to the general concept of emotional states like anger and fear, or attitudes like love, hate or pride, that implies that they should have any direct bodily effects. Certainly these bodily processes are not essential for emotions to be important in our social relationships. What makes intense jealousy matter to us is not the feeling in the belly or the tension, sweating, shaking or weeping, but the way it interacts with motives, decisions, beliefs and tendencies to act, and its strong disposition repeatedly to gain control of one's thought, making it impossible to get on properly with other tasks, or deal properly with other personal relationships, etc. Even if a pill could remove the physical effects, that would not get rid of the jealousy.

The discussion so far suggests that many of the physiological processes involved in human affect should be regarded as *contingent* features, relics of our evolutionary history (though perhaps some of them still play an important role in involuntary mechanisms for revealing internal state in social agents). The more socially significant, 'intrinsic' features of states like grief, fear, joy, love or hate, could occur in machines or organisms that lack automatic mechanisms of external expression. For agents without such uncontrollable external reactions, the communication of an affective state would *always* have to be a deliberate act, just as it *sometimes* is with us.

One reason why it is important that human physiology is not required for many of the familiar features of human affective states is that it leaves open the possibility of certain kinds of empathy between machines and people. It is a commonplace that one way of trying to understand another person's reactions is to "put yourself in his place". This is generally thought of as doing a kind of simulation of the other person by using one's own imagined reactions. If computer-based machines that lacked our physiology, and therefore were incapable of simulating our physiological reactions, could nevertheless simulate the other significant features of our emotional states, that might help them cope with the some of the problems of communicating with us. This may not be all that different from empathising in human beings: one can imaginatively put oneself the place of a happy or suffering friend without experiencing any of the physiological reactions.

16 The Boundary between Cognitive and Affective States

I have so far implied that there is a relatively clear distinction between affective and cognitive states, so that for example believing that there is a tiger in the next room would be a cognitive state and being afraid of tigers in general and that one in particular would be affective states. However the division is not so clear cut, as can be seen by considering beliefs such as that the tiger is dangerous.

It is clear that many words and phrases that function grammatically like factual descriptions of properties of things have a wider significance than this, relating those things to the needs, preferences, hopes, fears etc. of speakers and listeners. Words like "dangerous", "safe", "lovely", "poisonous", "nourishing", "wicked", "admirable", do not only describe things, they also have affective connotations. Somehow, they are linked with high level affective states and dispositions. Philosophers have called these "emotive" or "evaluative" words. In the classification of Ortony et al. (1987) these would probably go into something like their category "Subjective evaluations of external objects".

By using these terms to sum up inferences from factual information to practical or affective consequences, and storing the conclusions using such affectively laden words, we can remove the necessity to make the inferences repeatedly, and we reduce the necessity for separate types of information store for different kinds of information, e.g. with "X is a tiger" in one store and some specific association between tigers and avoidance behavior in another. By combining representations or symbols with different kinds of meaning, in this sort of way, we seem to be able to economize on mechanisms for both 'internal' and 'external' communication. The same mechanisms that are used for processing factual information and drawing inferences can therefore be directly integrated with mechanisms for generating or controlling motives and actions.

Some affective influences of factual descriptions are not part of the meaning in this extended sense but derive indirectly from what the words denote, for example the name of someone you hate or fear, a place name that reminds you of a terrifying experience. Even privately reading or thinking of words for human body-parts can produce embarrassment or erotic feelings. Although there are many theoretically possible computational models that would explain such triggering, it will not be clear what the design trade-offs between them are until we have a more complete model of affective states and processes.

17 Further Developments of the Theory

A more complete and detailed model for all this will have to take account of a wide range of cases, including the effects of known falsehoods. For example, a story that you know is pure fiction can make you weep or feel apprehension, relief, joy and the like. Stories, plays and films work on some deep-seated general link between the mechanisms used for envisaging or contemplating possible states of affairs and the mechanisms that trigger affective responses. An odd example of this mechanism at work is the way verbal mention of fingernails scratching on a blackboard can cause the same kind of horrible cold shudder as hearing the event itself. Sorry about that! Much of the power of literature and effectively persuasive communication depends on these mechanisms that apparently by-pass or override rational cognitive processes.

A peculiarity of the affective states induced by fictional or imagined situations is that although they can trigger behavioral expressions of affect (weeping, shuddering, etc.), and can also generate states that are close to having real motives (you hope the murderer will be caught, you want the hero to forgive the heroine, you are disappointed that a clue is not noticed, etc.) nevertheless these states are decoupled from the mechanisms that generate decisions and plans for action in the real world: you are not for a moment tempted to report the murder to the police, or pass on information about the culprit, unlike children who want to warn the endangered heroine at a pantomime. There are clearly different kinds of routes between dispositional affective states and behavior.

Additional phenomena that need to be explained by a general design-based theory of the mechanisms involved in communication and affect include the following:

- Emotional contagion in crowds, as exploited for example by rabble-rousers and pop stars.
- How requesting, pleading, begging, exhorting can influence short term or long term motivational affective states, without depending on power or authority in the speaker and without relying on arguments to the self-interest of the hearer.

- How 'powerful' stories, novels, allegories can work on high level dispositions that are not easily changed.
- How humour works
- Hypnotism
- The nature of aesthetic responses.
- The influence of communicative media (as opposed to content) on emotions, including the combined impact of words and music.

18 Implications for Human-Machine Interaction

I have tried to present a design-based theory (admittedly very incomplete) about some aspects of affective states, along with some of the implications both for modelling human communication (goals **G1** and **G3**) and for designing machines that communicate with human beings in order to achieve other goals, such as teaching, advising, helping, or dealing with emergencies (goal **G2**).

The latter designs will be very risky unless based on a good theory of human affective mechanisms. Even people designing unintelligent interfaces need to remember that users can have all sorts of affective responses even to relatively simple machine behavior. [2] Intelligent interface designers need to remember that affective states, whether produced deliberately or not, can have all sorts of influences on other processes, including the ability to attend, to take in new information, to perform in tests, etc.

Often the problems are much harder than the obvious HCI issues of choosing suggestive labels for boxes, selecting the word to denote errors, designing an 'attractive' screen layout, etc. A tutoring system that always interprets student failings in terms of lack of ability or lack of knowledge and tries always to give help on that limited basis will be a failure for many (though not all) students (du Boulay & Sloman, 1988). A good teacher often has to know what a student is feeling and why, and be able to help the pupil overcome bad feelings that are getting in the way of both learning and performing. This requires insight into particular kinds of emotional states in order (a) to be able to detect their presence and (b) know how to deal appropriately with them. A student who is frightened of the task or the teacher has to be treated differently from one who is worried about a suspected disease, or who is in pain after dental treatment.

However, detecting some affective states, especially dormant attitudes, that have no behavioral symptoms, can be very difficult. Predicting how a known attitude will interact with the next utterance can also be difficult, because it will depend on which other, possibly dormant, states are present. The problems cannot be dealt with simply by doing some initial tests to determine the student's affective state (e.g. using a standard questionnaire) and then assuming the result for the rest of the interactive session. Depending on the type of interaction it may be very important to be continually on the lookout for evidence of motivational or emotional change, and then to adjust the communicative strategy accordingly.

19 Doing Without a Good Model of the User

I have tried to show that for certain purposes, e.g. producing a machine that understands human linguistic utterances as well as people do (i.e. for goals of type **G1**) it may not be necessary to have a *correct* theory about how people work. It will generally suffice that the machine shares the presuppositions, right or wrong, that are built into our ordinary language for describing mental states. Examples of pioneering attempts to give machines this kind of ability are reported in Colby (1982), Lehnert (1987), and Dyer (1987). Doing this well requires the designers, if not the machine, to have a correct theory about the (possibly wrong!) theories presupposed by ordinary language. Alternatively it may be possible to design machines that learn such things by interacting with people.

However, designing machines in such a way that they not only share human understanding, but can actually use communication effectively for the other purposes listed above (i.e. goals of type **G2**) requires a correct theory of how human minds work. Practical success, however, does not require that the whole theory be explicitly formulated: some aspects may be implicit in successful heuristic rules, including rules that allow the interaction to be modified using

[2] We found, when we first started teaching programming using the language Pop-11, that printing the word "Error" caused distress in some students. We therefore changed it to "Mishap".

corrective strategies based on feedback from the client or student. This would require great advances in computer vision and speech understanding in order to detect relevant changes in facial expression, or tone of voice, etc.

For scientific purposes the use of heuristic rules that give adequate results, without any explanatory theory, would not be so satisfactory. Moreover the communicative task is harder when there is no scope for correction because the instruction or information has to be expressed the right way the first time. This is more likely to be achieved if based on good models of listeners, even if they are only models of general types of user rather than very detailed models developed separately for each individual. Often there is not enough time, so using some stereotype is inevitable: this may be satisfactory as long as (a) there are enough different stereotypes to match the variety of potential users, (b) the tests for identifying the appropriate stereotype are good enough, and (c) the stereotypes are based on rich and deep enough models or rules to cope with the complexities described above.

Failure to understand the extreme context sensitivity of affective control hierarchies can mislead mathematically minded psychologists into postulating purely probabilistic mechanisms linking situations with behavior. Equally, it may tempt designers of intelligent interfaces to use probabilistic rules for predicting or interpreting human behavior, or for generating communicative actions. It must be admitted that statistical inferences do give remarkably precise predictions for some kinds of mass behavior (otherwise psephologists would be out of business), but predicting what proportion of the masses will vote for X is no basis for knowing how to talk to an individual. Successful communication often needs to take account of that individual's precise set of relevant affective states, including those that are currently dormant but might easily be triggered. There is probably no foolproof way of achieving that, for humans or for machines. Dormant dispositions that interfere with communicative intentions are often impossible to detect, and may surprise even the individual concerned. So machine advice, tutoring, etc. will be at least as risky as communication between humans.

20 Conclusion

I have presented the outlines of a theory of affect that shows that some of the more ambitious attempts to design machines that can communicate effectively with human beings will need to take account of at least the following:

- A hierarchy of long term affective dispositional states that can be difficult to discern, difficult to change, and likely to interact with new information in a manner that depends on other affective and cognitive states.
- Relatively easily changeable global states, like depressed or optimistic moods that can influence responses to new information without having any relation of cognitive content.
- Relatively short term emotional states that may interfere with the attention required for coping with new information, or may interact with new information in unexpected ways.
- The need to generate new emotional states in order to hold the receiver's attention properly.
- The existence of dormant plans, goals, wishes, including actions that were suspended, previously formed intentions and conditional plans waiting for opportunities.
- The existence of learnt 'reflexes' linking the advent of new information with fast, uncontrollable mental or physical reactions.

All of these phenomena are difficult to cope with in others because discovering the relevant information about an individual's state of mind can be very difficult, though sometimes involuntary expressive behavior gives clues. These phenomena would also need to be understood in order to produce realistic models of human communication.

- The need to base communication on culture-specific and other stereotypes, because of the impossibility of checking all relevant facts about the hearer.

Sometimes an alternative to discovering the individual's state of mind is to use a collection of general models of the kinds of attitudes and high level dispositions to be expected in certain sub-classes of people, for instance members of a sub-culture, or children of a certain age. The task of assigning an individual to such a class is often easier than discovering his beliefs and attitudes. This strategy is apparently (unconsciously) employed by people, but has obvious dangers and limitations. A powerful learning system would be needed in order to build up a set of such models, but because of the difficulty of checking the classification of individuals even the learning task is inherently error prone.

Even when a lot is known about an individual's current affective state, it may be very difficult to give effective advice or persuasion because changing some of the higher level states can be difficult. In extreme cases nothing short of writing powerful novels will suffice.

A more complete design-based investigation would consider the brain mechanisms underlying these functional states, and would survey different possible mechanisms in which all this might be implemented, in order to understand fully the advantages and disadvantages of any one design. This would give insight into selective pressures that could help us understand how human mechanisms might have evolved.

I think it is clear that giving machines all, or even most, of these abilities will be well beyond the state of the art for many years to come. But it is important to keep trying, both as one of many ways to increase our self understanding and because there may be worthwhile practical results. At the very least, studying the problems may give us clues as to how to remedy some of the many deficiencies in communication between people.

Acknowledgements

I am grateful to several colleagues who have either commented on an earlier draft or discussed these issues with me, including Aluizio Araujo, Margaret Boden, Monica Croucher, Glyn Humphreys, Keith Oatley (the respondent at the workshop), Helen Petrie, David Young, Nicola Yuill and Andrew Ortony, whose editorial criticisms and suggestions have been particularly helpful. Some of the ideas reported here were developed while the author held a GEC Research Fellowship.

Bibliography

du Boulay, J.B.H., Sloman, A.: 'Bread today, jam tomorrow: the impact of AI on education' in Collins, J.H., Estes, N. & Walker, D. (eds) *Proceedings Fifth International Conference On Technology And Education,* Edinburgh, (1988) 82–85

Clore, G.L., Ortony, A.: 'The Semantics of the affective lexicon' in V. Hamilton, et al. (eds) *Cognitive Perspectives on Emotion and Motivation,* Kluwer Academic Publishers (1988) 367–397

Colby, K.M.: 'Modelling a paranoid mind' *Behavioral and Brain Sciences* **4**, (1982) 515–560

Dennett, D.C.: *Brainstorms* Bradford Books and Harvester Press, (1978)

Dyer, M.G.: 'Emotions and their computations: Three computer models', *Cognition and Emotion,* **1**,3, (1987) 323–347

Johnson-Laird, P. N.: *The computer and the mind: An introduction to cognitive science,* Fontana, (1988)

Johnson-Laird, P. N., Oatley K.: 'The language of emotions: an analysis of a semantic field' *Cognition and Emotion* **3**,2 (1989) 81–123

Lehnert, W.G., Vine, E.W.: 'The role of affect in narrative structure' *Cognition and Emotion,* **1**,3, (1987) 299-322

Neisser, U.: 'The imitation of man by machine' in *Science* **139**, (1963) 81–97

Oatley, K., Johnson-Laird, P.N.: 'Towards a cognitive theory of emotions' *Cognition and Emotion,* **1**, (1987) 29–50

Ortony, A., Clore, G.L., Foss, M.A.: 'The referential structure of the affective lexicon' *Cognitive Science,* **11**,3 (1987) 341–364

Ortony, A., Clore, G.L., Collins, A.: *The Cognitive Structure of the Emotions,* New York, Cambridge University Press (1988)

Ryle, G.: *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) 29–38

Sloman, A.: 'Interactions between Philosophy and artificial intelligence: the role of intuition and non-logical reasoning in intelligence' *Proceedings 2nd International Joint Conference on Artificial Intelligence,* London 1971, reprinted in *Artificial Intelligence,* **2**, (1971) 209–225, and in J.M. Nicholas (ed), *Images, Perception, and Knowledge* Dordrecht-Holland: Reidel 1977. (Also Cognitive Science Research Paper 191, University of Sussex)

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

Sloman, A, Croucher, M.: 'Why robots will have emotions', in *Proceedings 7th International Joint Conference on Artificial Intelligence*, Vancouver, (1981) 197–202, also available as Cognitive Science Research Paper 176, Sussex University

Sloman, A.: 'Skills, learning and parallelism', in *Proceedings Cognitive Science Conference*, Berkeley, (1981b) 284–285, also available as Cognitive Science Research Paper 013, Sussex University

Sloman, A.: 'Why we need many knowledge representation formalisms', in *Research and Development in Expert Systems*, ed. M Bramer, Cambridge University Press (1985a) 163–183

Sloman, A.: 'Real time multiple motive-expert systems', in *Expert Systems 85*, ed. M. Merry, Cambridge University Press, (1985b) 213–224

Sloman, A.: 'Motives Mechanisms and Emotions' in *Emotion and Cognition* **1**,3, (1987) 217–234 reprinted in M.A. Boden (ed) *The Philosophy of Artificial Intelligence* "Oxford Readings in Philosophy" Series Oxford University Press, (1990) 231–247

Sloman, A.: 'On designing a visual system (Towards a Gibsonian computational model of vision)' *Journal of Experimental and Theoretical AI* **1**,4, (1989) 289–337

Sloman, A.: 'Beyond Turing Equivalence' in P. Millican and A. Clark (eds) *Proceedings Turing90 Colloquium* (forthcoming) also available as Cognitive Science Research Paper 164, Sussex University

NOTE:
Since this paper was written many of the theoretical ideas have been further developed. For more information see
http://www.cs.bham.ac.uk/ axs/cogaff.html
http://www.cs.bham.ac.uk/research/cogaff/

This article was processed using the LaTeX macro package with ICM style