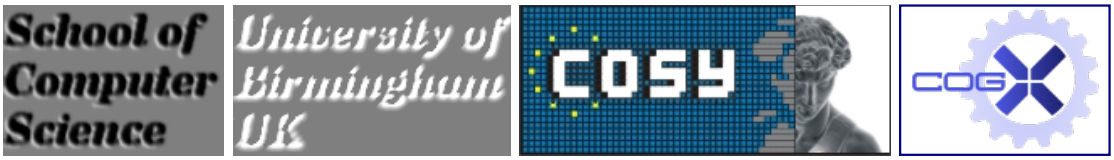


Discussion Note

Architecture-based motivation vs Reward-based motivation

Aaron Sloman
School of Computer Science
The University of Birmingham
Last updated: 13 Jul 2009
Installed: 25 May 2009 (Liable to change.)

CONTENTS (Provisional)

From time to time I shall create a PDF version of this file here. (Created by Firefox's 'print to pdf' option.) It will become out of date if I forget to update it after editing the html version.

Introduction

> "Reason is, and ought only to be the slave of the passions, and
> can never pretend to any other office than to serve and obey them."
> David Hume *A Treatise of Human Nature* (2.3.3.4), 1739-1740
> http://www.class.uidaho.edu/mickelsen/ToC/hume%20treatise%20ToC.htm

Whatever Hume may have meant by this, and whatever various commentators may have taken him to mean, I claim that there is at least one interpretation in which this statement is obviously true, namely: no matter what factual information an animal or machine A contains, and no matter what competences A has regarding abilities to reason, to plan, to predict or to explain, A will not actually **do** anything unless it has, in addition, some sort of control mechanism that selects among the many alternative processes that A's information and competences can support.

In short: *control* mechanisms are required in addition to factual information and reasoning mechanisms A is to *do* anything. This paper is about what forms of control are required. I assume that in at least some cases there are motives, and the control arises out of selection of a motive for action. That raises the question where motives come from. My answer is that they can be generated and selected in different ways, but one way is not itself motivated: it merely involves the operation of mechanisms in the architecture of

A that generate motives and select some of them for action. The view I wish to oppose is that all motives must somehow serve the interests of A, or be rewarding for A. This view is widely held and is based on a lack of imagination about possible designs for working system. I summarise it as the assumption that all motivation must be reward-based. In contrast I claim that at least some motivation may be architecture-based, in the sense explained below.

Instead of talking about "passions" I shall use the less emotive terms, "motivation" and "motive". A motive in this context, is a specification of something to be done or achieved (which could include preventing or avoiding some state of affairs, or maintaining a state or process). The words "motivation" and "motivational" can be used to describe the states, processes, and mechanisms concerned with production of motives, their control and management and the effects of motives in initiating and controlling internal and external behaviours. So Hume's claim, as interpreted here is that no collection of beliefs and reasoning capabilities can generate behaviour on its own: motivation is also required.

This view of Hume's claim is expressed well in the Stanford Encyclopedia of Philosophy entry on motivation, though without explicit reference to Hume:

> "The belief that an antibiotic will cure a specific infection may move an individual to take the antibiotic, if she also believes that she has the infection, and if she either desires to be cured or judges that she ought to treat the infection for her own good. All on its own, however, an empirical belief like this one appears to carry with it no particular motivational impact; a person can judge that an antibiotic will most effectively cure a specific infection without being moved one way or another."
> http://plato.stanford.edu/entries/moral-motivation

That raises the question: where do motives come from and why are some possible motives (e.g. going for lunch) selected and others (e.g. going for a walk, or starting a campaign for election to parliament) not selected?

If Hume had known about reflexes, he might have treated them as an alternative mode of initiation of behaviour to motivation (or passions). There may be some who regard a knee-jerk reflex as involving a kind of motivation produced by tapping a sensitive part of the knee. That would not be a common usage. I think it is more helpful to regard such physical reflexes as different from motives, and therefore as exceptions to Hume's claim. I shall try to show that something like "internal reflexes" in an information-processing system can be part of the explanation of creation and adoption of motives. In particular, adopting "the design-based approach to the study of mind" yields a wider variety of possible explanations of how minds work than are typically considered in philosophy or psychology, and paradoxically even in AI/Robotics, where such an approach ought to be more influential.

This proposal opposes a view that all motives are selected on the basis of the costs and benefits of achieving them, which we can loosely characterise as the claim that all motivation is "reward-based".

In the history of philosophy and psychology there have been many theories of motivation, and distinctions between different sorts of motivation, for example motivations related to biological needs, motivations somehow acquired through cultural influences, motivations related to achieving or maximising some reward (e.g. food, admiration in others, going to heaven), or avoiding or minimising some punishment (often labelled positive and negative reward or reinforcement),

motivations that are means to some other end, and motivations that are desired for their own sake, motivations related to intellectual or other achievements, and so on. Many theorists assume that motivation must be linked to rewards or utility. One version of this (a form of hedonism) is the assumption that all actions are done for ultimately selfish reasons.

I shall try to explain why there is an alternative kind of motivation, architecture-based motivation, which is not included even in this rather broad characterisation of types of motivation on Wikipedia:

> "Motivation is the set of reasons that determines one to engage in a particular behavior. The term is generally used for human motivation but, theoretically, it can be used to describe the causes for animal behavior as well. This article refers to human motivation. According to various theories, motivation may be rooted in the basic need to minimize physical pain and maximize pleasure, or it may include specific needs such as eating and resting, or a desired object, hobby, goal, state of being, ideal, or it may be attributed to less-apparent reasons such as altruism, morality, or avoiding mortality."
> http://en.wikipedia.org/wiki/Motivation

Philosophers who write about motivation tend to have rather different concerns such as whether there is a necessary connection between deciding what one morally ought to do and being motivated to do it. For more on this see the afore-mentioned entry in the Stanford Encyclopedia of philosophy.

Motivation is also a topic of great concern in management theory and management practice, where motivation of workers comes from outside them e.g. in the form of reward mechanisms (providing money, status, recognition, etc.) sometimes in other forms, e.g. inspiration, exhortation, social pressures, ... I shall not discuss any of those ideas.

In psychology and even in AI, all these concerns can arise, though I am here only discussing questions about the mechanisms that underlie processes within an organism or machine that select things to aim for and which initiate and control the behaviours that result. This includes mechanisms that produce goals and desires, mechanisms that identify and resolve conflicts between different goals or desires, mechanisms that select means to achieving goals or desires.

Achieving a desired goal G could be done in different ways, e.g.

- select and use an available plan for doing things of type G
- use a planning mechanism to create a plan to achieve G and follow it.
- detect and follow a gradient that appears to lead to achieving G
  (e.g. if G is being on high ground to avoid a rising tide, walk uphill while you can)

There is much more to be said about the forms different motives can have, and the various ways in which their status can change, e.g. when a motive has been generated but not yet selected, when it has been selected, but not yet scheduled, or when there is not yet any clear plan or strategy as to how to achieve it, or whether action has or has not been initiated, whether any conflict with other motives, or unexpected obstacle has been detected, etc.

For a characterisation of some of the largely unnoticed complexity of motives see
http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#16
L.P. Beaudoin, A. Sloman, A study of motive processing and attention,
Prospects for Artificial Intelligence, IOS Press, 1993
(further developed in Luc Beaudoin's PhD thesis).

Where do motives come from?

It is often assumed that motivation, i.e. an organism's or machine's, selection, maintenance, or pursuance of some state of affairs, the motive's content, must be related to the organism or machine having information (e.g. a belief, or expectation) that achievement of the motive will bring some rewards or benefit, sometimes referred to as "utility". This could be reduction of some disadvantage or disutility, e.g. a decrease in danger or pain.

Extreme versions of this assumption are found in philosophical theories that all agents are ultimately selfish, since they can only be motivated to do things that reward themselves, even if that is a case of feeling good about helping someone else.

More generally, the assumption is that selection of a motive among possible motives must be based on some kind of prediction about the consequences of achieving or preventing whatever state of affairs is specified in that motive. This document challenges that claim by demonstrating that it is possible for an organism or machine to have, and to act on motives for which there is no such prediction.

My claim

My claim is that an organism (human or non-human) or machine may have something as a motive whose existence is merely a product of the operation of a motive-generating mechanism -- which itself may be a product of evolution, or something produced by a designer, or something that resulted from a learning or developmental process, or in some cases may be produced by some pathology. Where the mechanism comes from and what its benefits are are irrelevant to its being a motivational mechanism: all that matters is that it should generate motives, and thereby be capable of influencing selection of behaviours.

In other words, it is possible for there to be *reflex* mechanisms whose effect is to produce new motives, and in simple cases to initiate behaviours controlled by such motives. I shall present a very simple architecture illustrating this possibility below, though for any actual organism, or intelligent robot, a more complex architecture will be required, for reasons given later.

Where the reflex mechanisms come from is a separate question: they may be produced by a robot designer or by biological evolution, or by a learning process, or even by some pathology (e.g. mechanisms producing addictions) but what the origin of such a mechanism is, is a separate question from what it does, how it does it, and what the consequences are.

I am not denying that some motives are concerned with producing benefits for the agent. It may even be the case (which I doubt) that most motives generated in humans and other animals are selected because of their benefit for the individual. For now, I am merely claiming that something different can occur and does occur, as follows:

> Not all the mechanisms for generating motives in a particular organism O, and not all the motives produced in O have to be related to any reward or positive or negative reinforcement for O.
>
> What makes them motives is how they work: what effects they have, or, in more complex cases, what effects they *tend* to have even though they are suppressed

(e.g. since competing, incompatible, motives can exist in 0).

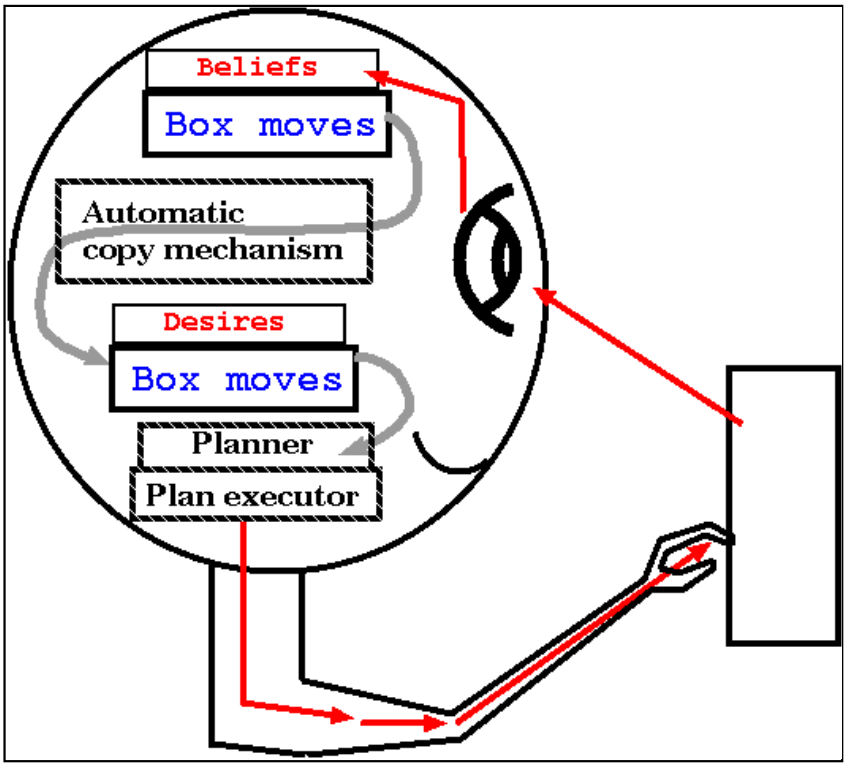
Learning and motivation

Many researchers in AI and other disciplines (though not all) assume that learning must be related to reward in some way, e.g. through positive or negative reinforcement.

I think that is false: some forms of learning occur simply because the opportunity to learn arises and the information-processing architecture produced by biological evolution simply reacts to many opportunities to learn, or to do things that could produce learning because the mechanisms that achieve that have proved their worth in previous generations, without the animals concerned knowing that they are using those mechanisms nor why they are using them.

Architecture-based motivation

Consider a very simple design for an organism or machine. It has a perceptual system that forms descriptions of a process occurring in the environment. Those descriptions are copied/stored in a data-base of "current beliefs" about what is happening in the world or has recently happened.



At regular intervals another mechanism selects one of the beliefs about processes occurring recently and copies its content (perhaps with some minor modification or removal of some detail, such as direction of motion) to form the content of a new motive in a database of "desires". The desires may be removed after a time.

At regular intervals an intention-forming mechanism selects one of the desires to act as a goal for a planning mechanism that works out which actions could make the desire come true, selects a plan, then initiates plan execution.

This system will automatically generate motives to produce actions that repeat or continue changes that it has recently perceived, possibly with slight modifications, and it will adjust its behaviours so as to execute a plan for fulfilling the latest selected motive.

Why is a planning mechanism required instead of a much simpler reflex action mechanism that does not require motives to be formulated and planning to occur?

A reflex mechanism would be fine if evolution had detected all the situations that can arise and if it had produced a mechanism that is able to trigger the fine details of the actions in all such situations. In general that is impossible, so instead of a process automatically triggering behaviour it can trigger the formation of some goal to be achieved, and then a secondary process can work out how to achieve it in the light of the then current situation.

For such a system to work there is NO need for the motives selected or the actions performed to produce any reward. We have goals generated and acted on without any reward being required for the system to work. Moreover, a side effect of such processes might be that the system observes what happens when these actions are performed in varying circumstances, and thereby learns things about how the environment works. That can be a side effect without being an explicit goal.

A designer could put such a mechanism into robot as a way of producing such learning without that being the robot's goal. Likewise biological evolution could have selected changes that lead to such mechanisms existing in some organisms because they produce useful learning, without any of the individual animals knowing that it has such mechanisms nor how they were selected or how they operate.

More complex variations

There is no need for the motive generating mechanism to be so simple. Some motives triggered by perceiving a physical process could involve systematic variations on the theme of the process e.g. undoing its effects, reversing the process, preventing the process from terminating, joining in and contributing to an ongoing process, or repeating the process, but with some object or action or instrument replaced. A mechanism that could generate such variations would accelerate learning about how things work in the environment, if the effects of various actions are recorded or generalised or compared with previous records, generalisations and predictions.

The motives generated will certainly need to change with the age and sophistication of the learner.

Some of the motive-generating mechanisms could be less directly triggered by particular perceived episodes and more influenced by the previous history of the individual, taking account not only of physical events but also social phenomena, e.g. discovering what peers seem to approve of, or choose to do. The motives generated by inferring motives of others could vary according to stage of development. E.g. early motives might mainly be copies of inferred motives of others, then as the child develops the ability to distinguish safe from unsafe experiments, the motives triggered by discovering motives of others could include various generalisations or modifications, e.g. generalising some motive to a wider class of situations, or restricting it to a narrower class, or even generating motives to oppose the perceived motives of others (e.g. parents!).

Moreover some of the processes triggered instead of producing external actions could produce internal changes to the architecture or its mechanisms. Those changes could include production of new motive generators, or motive comparators, or motive generator generators, etc.

For more on this idea see chapter 6 and chapter 10 of The Computer Revolution in Philosophy (1978).

Mechanisms required

In humans it seems that architecture-based motivation plays a role at various levels of cognitive development, and is manifested in early play and exploration, and in intellectual curiosity later on, e.g. in connection with things like mathematics or chess, and various forms of competitiveness.

Such learning would depend on other mechanisms monitoring the results of behaviour generated by architecture-based motivational mechanisms and looking for both new generalisations, new conjectured explanations of those generalisations and new evidence that old theories or old conceptual systems are flawed -- and require debugging.

Such learning processes would require additional complex mechanisms, including mechanisms concerned with construction and use of powerful forms of representation and mechanisms for producing substantive (i.e. non-definitional) ontology extension.

For more on additional mechanisms required see

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang
Evolution of minds and languages. What evolved first and develops first in children: Languages for communicating, or languages for thinking (Generalised Languages: GLs)

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#prague09
Ontologies for baby animals and robots From "baby stuff" to the world of adult science: Developmental AI from a Kantian viewpoint.

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#toddlers
A New Approach to Philosophy of Mathematics: Design a young explorer, able to discover "toddler theorems" (Or: "The Naive Mathematics Manifesto").

The mechanisms constructing architecture-based motivational sub-systems could sometimes go wrong, accounting for some pathologies, e.g. obsessions, addictions, etc. But at present that is merely conjecture.

Conclusion

If all this is correct, then humans, like many other organisms, may have many motives that exist not because having them benefits the individual but because ancestors with the mechanisms that produce those motives in those situations happened to produce more descendants than conspecifics without those mechanisms did. Some social insect species in which workers act as 'slaves' serving the needs of larvae and the queen appear to be examples. In those cases it may be the case that

> **Some motivational mechanisms reward the genomes that specify them, not the individuals that have them.**

Similarly, some forms of learning may occur because animals that have certain

learning mechanisms had ancestors who produced more offspring than rivals that lacked those learning mechanisms. This could be the case without the learning mechanism specifically benefiting the individual. In fact the learning mechanism may lead to parents adopting suicidal behaviours in order to divert predators from their children.

If follows that any AI and cognitive science research based on the assumption that learning is produced ONLY by mechanisms that maximise expected utility for the individual organism or robot, is likely to miss out important forms of learning. Perhaps the most important forms.

One reason for this is that typically individuals that have opportunities to learn do not know enough to be able to even begin to asses the long term utility of what they are doing. So they have to rely on what evolution has learnt (or a designer in the case of robots) and, at a later stage, on what the culture has learnt. What evolution or a culture has learnt may, of course, not be appropriate in new circumstances!

This discussion note does not *prove* that evolution produced organisms that make use of architecture-based motivation in which at least some motives are produced and acted on without any reward mechanism being required. But it illustrates the possibility, thereby challenging the assumption that ALL motivation must arise out of expected rewards.

Similar arguments about how suitably designed reflex mechanisms may react to perceived processes and states of affairs by modifying internal information stores could show that at least some forms of learning use mechanisms that are not concerned with rewards, with positive or negative reinforcement, or with utility maximisation (or maximisation of expected utility). My conjecture is that the most important forms of learning in advanced intelligent systems (e.g. some aspects of language learning in human children) are architecture-based, not reward based. But that requires further investigation.

The ideas presented here are very relevant to projects like CogX, which aim to investigate designs for robots that 'self-understand' and 'self-extend', since it demonstrates at least the *possibility* that some forms of self-extension may not be reward-driven, but architecture-driven.

Various forms of architecture-based motivation seem to be required for the development of precursors of mathematical competences described here:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#toddlers

Some of what is called 'curiosity-driven' behaviour probably needs to be re-described as 'architecture-based' or 'architecture-driven'.

[This document is still under construction. Suggestions for improvement welcome. It is likely to change frequently in the first few years of its life!]

This is one of a series of notes explaining how learning about underlying mechanisms can alter our views about the 'logical topography' of a range of phenomena, suggesting that our current conceptual schemes (Gilbert Ryle's 'logical geography') can be revised and improved, at least for the purposes of science, technology, education, and maybe even for everyday conversation, as

explained in http://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html

NOTE
Marvin Minsky wrote quite a lot about goals and how they are formed in *The Emotion Machine*. It seems to me that the above is consistent with what he wrote, though I may have misinterpreted him.

Something like the ideas presented here were taken for granted when I wrote The Computer Revolution in Philosophy in 1978. However, at that time I underestimated the importance of spelling out assumptions and conjectures in much greater detail.

Acknowledgements

I wish to thank Veronica Arriola Rios and Damien Duff for helpful comments on an earlier, less clear, draft.

Maintained by Aaron Sloman
School of Computer Science
The University of Birmingham