# Notes for a 2004 workshop on self-aware machines

## DARPA WORKSHOP ON SELF-AWARE COMPUTER SYSTEMS

**Organised by John McCarthy and Pat Hayes, Virginia 2004**

**Participant statements are available here:**
**https://www.ihmc.us/users/phayes/DWSAS-statements.html**

**Aaron Sloman**
**http://www.cs.bham.ac.uk/~axs/**

Note: this is a totally unrelated document:
http://www.ihmc.us/users/phayes/dwsas-statements.html

**Last modified:**
26 Jan 2004; 26 Jan 2016 (Added links); 5 Aug 2021 (fixed refs)

**Note added 5 Aug 2021:**
A brief summary/overview of conclusions from the workshop is here:
https://apps.dtic.mil/sti/pdfs/AD1002393.pdf
Report on DARPA Workshop on Self-Aware Computer Systems
Eyal Amir, Michael L. Anderson, and Vinay K. Chaudhr

---

## CONTENTS:

## GENERALISING THE QUESTION

> *The purpose of the workshop is to discuss ways in which computer systems can be made to be aware of themselves, and what forms of self-awareness will be useful for systems with various functions.*

The stated purpose of the workshop is to discuss a special subset (computer systems) of a class of systems that can be made aware of themselves.

I would like to try to relate this to a broader class (information-processing systems), by combining biological and engineering viewpoints, partly because I think a more general theory provides deeper understanding and partly because biological examples may extend our ideas about what is possible in engineered systems.

I have been trying since the 1970s to do a collection of related things including:

1. Provide a unifying conceptual framework based on the design stance for talking about 'the space of possible minds', including both natural and artificial minds.

2. Use that conceptual framework for developing both:

    2.1. scientific explanations of various kinds of natural phenomena, in many kinds of animals, including humans at different stages of development, humans in different cultures, and humans with various kinds of brain damage, or other abnormalities

    2.2. solutions to engineering problems where the solutions are theoretically well-founded, and it is possible to explain why they are solutions, and not just the result of searching for designs that pass various tests.

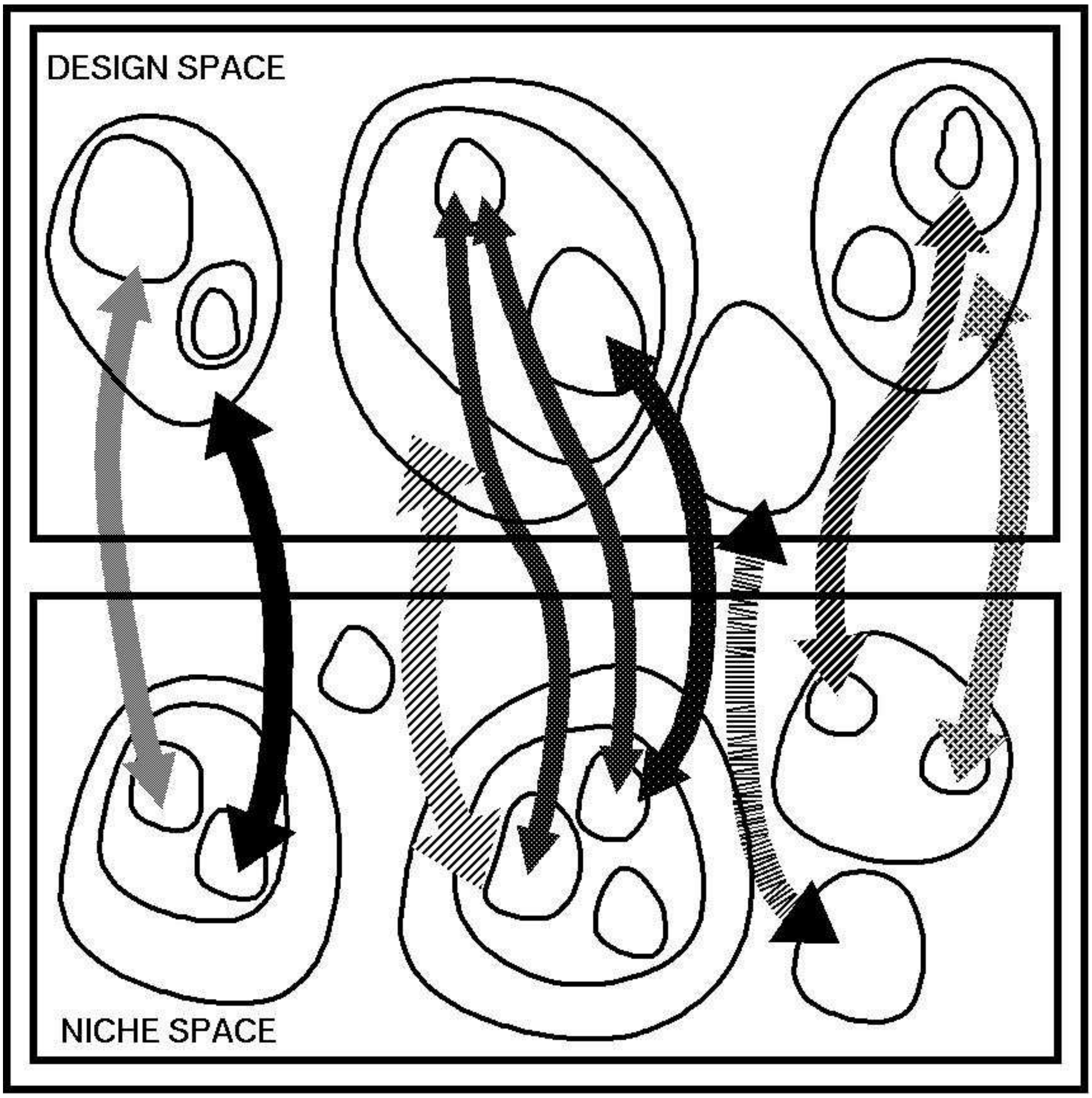
Using the design stance includes:

A. Developing both an ontology for *requirements* as well as for *designs*, and also a way of describing *relationships* between requirements and designs that is deeper and richer than the now commonplace use of fitness functions. (For natural systems the analogue of a set of requirements is a biological niche).

B. Developing ontologies for characterising designs at different levels of abstraction including the physical level, physiological levels, and the kinds of functional levels familiar in AI and software engineering.

This includes the description of virtual-machine architectures in terms of their components, the kinds of functions they have, the ways the components interact (types of causal interactions in virtual machines), the kinds of information they process and the forms of representation they use.

Here is a picture of design space and niche space, and relationships between sub-regions:



## RELATIONSHIPS BETWEEN DESIGN SPACE AND NICHE SPACE

In this framework, the workshop task:

> *to discuss ways in which computer systems can be made to be aware of themselves, and what forms of self-awareness will be useful for systems with various functions.*

can be seen as just a special case of the investigation of the relationship between designs and requirements. However, there are many different sets of requirements and often the solutions that satisfy the requirements are not unique, unless further constraints are added to the requirements.

**Note added 19 Aug 2019:**
Interestingly, as far as I can recall nobody at the workshop fell into the trap of arguing that in order to be self-aware a machine or organism needs to have a special component of type self, unlike many philosophers (and others) who have wasted words on arguing about what sort of thing a self is (and whether there is such a thing as a "true self", as in https://digest.bps.org.uk/2017/08/22/there-is-no-such-thing-as-the-true-self-but-its-still-a-useful-psychological-concept/ among many other discussions). I suppose some people may think that because a powered lawnmower can damage itself, it must have a self to damage, but most will see through that trap.

For a more detailed discussion and rejection of "self-mythology" see http://www.cs.bham.ac.uk/research/projects/cogaff/misc/the-self.html.

## CONSEQUENCES OF ADDING CONSTRAINTS TO REQUIREMENTS

E.g. if a requirement is specified in purely behavioural terms, then meeting the requirement over any finite time period can always be done (in principle) by a design with a sufficiently large collection of shallow condition-action rules (where the conditions include histories).

However, if further constraints are added, such as that the memory should not exceed a certain size, then it may be necessary to replace the set of shallow rules with something more abstract with generative power, in order to produce the same (potentially huge) set of behaviours in the same environment. (I.e. making the same set of counter-factual conditionals true of the agent.)

If an additional constraint is that the environment can change in ways that cannot be predicted by the designer, or, in the case of a natural system, if the environment changes in ways that produce requirements not covered by the specific evolutionary history of a species then that requires a higher level of abstraction in the system, namely a learning capability. (Something can be generative without learning, like a parser for a fixed grammar.)

If different niches require different sorts of learning capabilities then the architecture may be provided with different learning mechanisms and in addition the ability to detect when to switch between them.

However, if the sorts of learning capabilities needed are not themselves known in advance by a designer, nor acquired in the evolutionary history of a species, then a higher level ability to learn new learning abilities will be needed.

## LEARNING TO LEARN

Conjecture: Humans, and possibly some other animals, somehow evolved an ability to learn to learn, though it requires each new-born individual to go through a period in which older individuals systematically increase the difficulties of the challenges set, either through explicitly designed tasks or by steadily reducing the amount of protection and help provided to the learner, or some combination.

There are many other sub-divisions among sets of requirements (niches) that impose constraints on designs that can meet the requirements.

## DEVELOPING AN ONTOLOGY FOR INTERNAL STATES

For example, if the environment contains other individuals that process information then it is helpful to be able to predict or explain their actions. There are many ways this could be done, e.g. either by using very large numbers of shallow correlations between observed contexts and observed behaviours.

However, it may be possible to achieve far greater generality more economically if the description of other agents uses an ontology that refers not only to their behaviours, but also to information-processing mechanisms within them, e.g. mechanisms for forming beliefs, generating desires, dealing with conflicts of desires, forming and executing plans, acquiring information by perception and reasoning, etc. where the notions of beliefs, desires, etc. are not simply abstractions from behaviour but defined in terms of assumed internal architectures and mechanisms.

I.e. evolution may have given some organisms the ability to use the ’design stance’ in dealing with others: this is much deeper and more general than using the intentional stance (or Newell’s ’Knowledge Level’), since the latter presupposes rationality, and therefore cannot be applied to animals that are not rational, since they do not reason, but nevertheless process information, e.g. insects, mice, frogs, chickens, human infants, etc.

## OTHER/SELF SYMMETRY

Many of the benefits of being able to take the design stance directed at information-processors in dealing with *others* can also come from taking it in relation to *oneself*, e.g. learning to anticipate desires and beliefs one is going to have, being able to reason about percepts one would have as a result of certain actions without actually performing the actions, being able to notice and either compensate for or in some cases remedy flaws in one’s own reasoning, planning, preferences, conflict-resolution strategies, emotional relations, or learning procedures.
[John McCarthy has a useful list in his paper on self-aware machines.]

This is sometimes referred to as having a ’reflective’ layer in an architecture, though my students, colleagues and I have been using the label ’meta-management’ since (a) in common parlance reflection is a rather passive process and (b) some members of the research community use the label ’reflective’ to cover a narrower range of processes than we do. The label is not important as long as we understand the variety of possible types of functions and mechanisms that can support them.

There are some people who argue that inward-directed reflective or meta-management capabilities require no other mechanisms than the mechanisms that suffice for adopting the outward-directed design stance.

For just as one can observe the behaviour of others, form generalisations, make predictions, construct explanations, in terms of the internal information-processing of other individuals, so also can an agent use *exactly* the same resources in relation to itself, though there will be differences. E.g. there are differences arising out of the difference of viewpoint, which will sometimes make it easier to infer the state of another than to infer one’s own state (e.g. because facial expressions of others can more easily be seen), and sometimes make it easier to infer one’s own state because one’s own behaviour is more continuously observable and sometimes in more detail.

The above line of reasoning is sometimes used to claim that self-awareness requires no special architectural support.

This claim is especially to be found in the writings of certain positivist or behaviourist philosophers, and those who like to repeat the quotation:

> ’How can I know what I think unless I hear what I say?’
> (Attributed variously to E.M.Forster, Graham Wallas, Tallulah Bankhead, and possibly others....)

## WHEN DO EXTRA ARCHITECTURAL FEATURES HELP?

However, from a design stance it is clear that

(a) not necessarily all internal states and processes of information-processing systems will be externally manifested (e.g. because available effectors may not have sufficient bandwidth, or for other reasons)

(b) it is in principle possible for internal states and processes to be internally monitored if the architecture supports the right sort of concurrency and connectivity

(c) in some cases the additional abilities produced by the architectural extensions may have benefits for the individual or the species, e.g. supporting high level self-debugging and learning capabilities, or supporting the ability to short-circuit ways of asking for and requesting help (e.g. when an oculist asks the patient to describe visual experiences instead of simply using behavioural tests of accuracy in catching, throwing, discriminating, etc.).

In my case, for instance, I believe that the reason I have not suffered from RSI despite spending a huge amount of time at a terminal each week, is that I earlier learnt to play musical instruments and discovered there (with the help of teachers) the importance of sensing the onset of stress and deliberately relaxing in order to produce a better tone, or smoother phrasing, or even simply in order to be able to play faster pieces at the required speed.

Because I developed that internal monitoring capability when playing a flute or violin I also use it when typing and at the first sign of stress am able to adjust various aspects of how I type. This requires architectural support for internal self-monitoring, not just the ability to observe my own behaviour as I would observe others.
[I can give many other examples.]

## ONTOLOGIES AND FORMS OF REPRESENTATION FOR INFORMATION-PROCESSING

What I’ve written so far is very schematic and abstract. There is a lot more to be said about the different sorts of ontologies and different forms of representation and reasoning required for different kinds of self-awareness, and the different architectural options to be considered.

One of the questions to be addressed is how good the ’folk psychology’ ontology is for the purposes of a machine that thinks about machines that think.

There is not just one FP ontology but several, found at different stages of individual human development, and in different cultures.

FP ontologies (plural) like ’naive’ conceptions of physical reality, are products of biological and social evolution, and, like all such products, they need to be understood in relation to the niches they serve.

I claim that there is a core of those FP ontologies which is a useful precursor to a more refined, scientific, ontology for mental phenomena which is based on explicit recognition of the nature and functions of virtual machines.

This useful core includes notions like: perception, desires, beliefs, preferences, learning, hopes, fears, attitudes, moods, puzzlement, understanding, unintentional action, surprise, pleasure, pain, introspection, etc. etc. But we can re-interpret all of those on the basis of an architectural theory. (Much work of that kind is going on in the Cognition and Affect project at Birmingham.)

I.e. by adopting the design stance, we, and intelligent machines of the future, can improve on the core FP ontology in the same way as adopting something like the design stance to the physical world enabled us to extend and refine the ’folk physics’ (for kinds of physical stuff, properties of physical stuff, processes involving physical stuff, etc.) That required a new theory of the architecture of matter.

Likewise we’ll need new theories of the architectures of minds: not ’architecture of mind’, as some philosophers think, because there are many possible types of mind, with different architectures, whereas one architecture exists for matter (though there are different architectures in the same physical system, at different levels of abstraction, e.g. as studied by chemistry, materials science, geology, astrophysics, etc.)

In some cases it may be possible to program an ’improved’ ontology for mental phenomena and the appropriate forms of representation for expressing it, directly into some self-aware machines.

In other cases it may be necessary for the ontology to be bootstrapped through processes in which the machine interacts with information processors (including itself), for instance using some kind of self-organising introspective mechanism, just as self-organising perceptual mechanisms can develop ontologies that effectively classify certain classes of sensory inputs (e.g. Kohonen nets).

A self-organising self-perceiver can, in principle, develop an ontology for self-description that is a product of its own unique internal history and the particular initial parameters of its internal sensors, etc.

Such agents may be incapable of expressing in any kind of language used to communicate with others the *precise* semantics of its self-descriptors. (At least those developed in this manner: others may be defined more by their structural and functional relationships, and those can be communicated.)

The ’non-communicable’ (private) concepts developed in self-organising self-perceivers are ’causally indexical’, in the sense of J.Campbell (1994) *Past Space and Self*, as explained in my 2003 JCS article with Ron Chrisley.

This causal indexicality suffices to explain some of the features of the notion of ’qualia’ that have caused so much confusion among philosophers and and would-be philosophers in other disciplines.

(Some intelligent artifacts may end up equally confused.)

PERCEPTUAL SHORT-CUTS

In general, getting from observed behaviour of something to a description of its internal information-processing may require convoluted and lengthy reasoning. (’It must have known X, and wanted Y, and disliked Z, because it did A and B, in circumstances D and E’ etc.) (Compare debugging complex software systems.)

However if getting the description right *quickly* is important, organisms can evolve or learn extra perceptual techniques for rapidly and automatically interpreting certain combinations of cues in terms of internal states, e.g. seeing someone as happy or intending to open the door, provided that there are certain regularities in the designs and functions of the systems being observed, which are revealed in patterns of behaviour. Those patterns can then be learnt as powerful cues (e.g. the pattern of behaviour indicating that someone is trying hard to see what's happening in a particular place.)

It may also be useful for organisms to evolve *expressive* behaviours that make it easier for others to infer their mental state. Once that starts, co-evolution of expressive behaviours and specialised perceptual mechanisms can lead to highly expert perceptual systems.

(I argued in a paper published in 1992 that if there were no *involuntary* expressions of behaviour, the costs to intelligent species would be too high, e.g. too much hedging of bets would be necessary: http://www.cs.bham.ac.uk/research/cogaff/81-95.html#10 )

This suggests that cooperating 'high level' expressive and perceptual capabilities co-evolved, as manifested in our immediate reaction to these pictures:

http://www.cs.bham.ac.uk/~axs/fig/postures.gif
http://www.cs.bham.ac.uk/~axs/fig/faces.gif

Conjecture: both high level action mechanisms and high level perceptual mechanisms linked to the ontology and forms of representation of a meta-management (reflective) architectural layer evolved in nature, and will be needed in intelligent machines interacting with other intelligent machines, e.g. humans.

(Note: this is not the same thing as training low level neural nets etc. to label faces as 'happy', 'sad', etc. as many researchers now do, for such systems have no idea what happiness and sadness are.)

## SUMMARY

The ontology and forms of representation that are useful in thinking, reasoning and learning about the information-processing going on in other agents can also be useful in relation to oneself.

To some extent the application of such an ontology to oneself can use the same perceptual mechanisms as suffice for its application to others. (Self/Other symmetry works up to a point.)

However, the development of special-purpose architectural features can support additional self-referential capabilities that may produce designs specially suited to certain sets of requirements (niches). We need to understand the trade-offs.

It is possible to produce very long lists of examples of ways in which self-awareness of various kinds can be useful.

However, it will be useful if we can put such examples in the context of a general conceptual and theoretical framework in which trade-offs between different design options can be discussed in relation to different sets of requirements and design constraints.

The analysis of such trade-offs may be far more useful in the long run than the kinds of arguments often found in the literature as to whether one design is better than another, or whether everything can be achieved with some particular class of mechanisms, where people are often merely attempting to support their own design preferences. I.e. we need to understand trade-offs when there are no right answers.

In particular, such an analysis will not only clarify engineering design requirements, but may also give us a better understanding of how various kinds of self-awareness evolved in nature.

We can thereby become more self-aware!

See also http://www.cs.bham.ac.uk/~axs/misc/talks/#inf
TALK 26: WHAT ARE INFORMATION-PROCESSING MACHINES? WHAT ARE INFORMATION-PROCESSING VIRTUAL MACHINES?
Notes from a workshop on models of consciousness, Sept 2003, and a presentation in York, Feb 2004.

## BACKGROUND

Self-knowledge is a topic on which I have been thinking (and writing) for a long time, as part of a larger project to understand architectural requirements for human-like systems.

E.g. my IJCAI 1971 paper on logical and non-logical varieties of representation, chapters 6, 8 and 10 of *The Computer Revolution in Philosophy* (1978) -- now online at
http://www.cs.bham.ac.uk/research/cogaff/crp/

Also relevant:
the IJCAI 1981 paper on Why robots will have emotions, my IJCAI 1985 and ECAI 1986 papers on semantics and causality, and more recent work on architectures, for instance:

Invited talk at 7th Conference of Association for the Scientific Study of Consciousness.
Expanded slides are online here:
http://www.cs.bham.ac.uk/research/cogaff/talks/#talk25
What kind of virtual machine is capable of human consciousness?

Virtual machines and consciousness, A. Sloman and R.L. Chrisley, (2003,) *Journal of Consciousness Studies*, 10, 4-5, pp. 113--172,
Online here:
http://www.cs.bham.ac.uk/research/cogaff/sloman-chrisley-jcs03.pdf

The architectural basis of affective states and processes, A. Sloman, R.L. Chrisley and M. Scheutz, in *Who Needs Emotions?: The Brain Meets the Machine*, Eds. M. Arbib and J-M. Fellous, Oxford University Press,
Online at
http://www.cs.bham.ac.uk/research/cogaff/sloman-chrisley-scheutz-emotions.pdf