



**John Haugeland, Costas Pagondiotis**

## Intelligence and the ability to take responsibility

*An interview with John Haugeland*

In an interview with *Cogito* (Greece), philosopher John Haugeland proposes that the defining feature of human intelligence is responsibility. On the ethical level, this means being able to decide between what one is told to do and what one ought to do; on the cognitive level, it involves abandoning a certain theory if it fails to comply with observation.

**Costas Pagondiotis:** In your book *Artificial Intelligence -- the Very Idea*, (1985)<sup>1</sup> you coined the term "Good Old-Fashioned Artificial Intelligence (GOFAI)" in order to name the then dominant approach to AI. What are the main theoretical assumptions of this approach and is it still dominant?

**John Haugeland:** "GOFAI" is a term for the first serious attempt to build machines that would have the general, flexible kind of intelligence that is characteristic of ordinary people. The fundamental idea was that intelligence has two essential components: first, a lot of factual knowledge about the world; and, second, a flexible ability to think about the world (on the basis of that knowledge) so as to draw conclusions and make decisions.

In order for this traditional idea to support a scientific research programme, there had to be a way to formulate and test empirical hypotheses; and brain science at the time was too "low level" to be of much help.

The essential breakthrough (early in the 1950s) came with the availability of (relatively) large electronic computers and, even more important, a profound conceptual abstraction in understanding them. Through the work especially of Alan Turing, John McCarthy, and Allen Newell, it became clear that numerical calculation is just one special case of what computers can do. As long as you could precisely specify the symbol structures, the operations that could be performed on them, and the conditions under which those operations were to be performed, the machine itself was completely neutral.

That's why computers also came to be called "*general-purpose symbol-manipulating systems*". But that insight is very exciting, because all kinds of things can be conceived as "symbol manipulation". In particular, formal logic is a kind of symbol manipulation, governed by explicit rules. Logic, however, going all the way back to Aristotle, has always been associated with thinking and reasoning. Of course, real human thinking is much more than just logical inference. But it suddenly became plausible to suppose that the right way to generalize logical thinking beyond the textbooks might be at the symbol-manipulation level. And the beauty of computers is not only that they can implement any such symbol structures, but that they can also

actually perform the prescribed manipulations.

It was this inspiration that spawned and supported the vibrant research programme that I called "good old-fashioned" artificial intelligence. I think that it can legitimately be called a "paradigm" in Thomas Kuhn's sense. It dominated not only AI but also cognitive psychology and even linguistics for something like a quarter of a century. By the mid-1980s, however, when I introduced the expression, its dominance had begun to wane. That's the point of the phrase "good *old-fashioned*". New ideas were on the scene — most notably "PDP" or "connectionist" systems — and many of the younger scientists were heading in that direction.

**CP:** Both you and Hubert Dreyfus stress that one of the biggest problems that GOFAI faces is how to represent and access common sense knowledge. What exactly is this problem, and do you think that there is any in-principle argument that shows that GOFAI could not possibly solve it?

**JH:** There are really two problems here. The first is also called the problem of relevance; it can be illustrated with a simple example proposed by Yehoshua Bar-Hillel in 1960. Suppose you want to design a system to translate English texts into other languages; and consider the problem you face with even so simple a text as: "The box was in the pen." The trouble is that the word "pen" is ambiguous between (just to keep it simple) "play pen" and "fountain pen". But since most other languages won't have that same ambiguity, your system will in general have to resolve it — figure out which sense was intended. The question is: how is it going to do that?

A human translator would know that a typical box wouldn't fit inside a fountain pen; so, in this context, the word must mean "play pen". That's the kind of knowledge that's sometimes called "common sense". A symbol manipulating system could, of course, store lots of information about boxes, fountain pens, and play pens — including their sizes. But how is it going to determine *which* of these facts is relevant? How will it determine that size is what matters in *this* case?

You might think that it could just work through *all* its recorded facts about play pens, fountain pens, and boxes, until it found a combination that settled the question. The difficulty is that, when you look at *combinations*, the number of possibilities "explodes". And it only gets worse when you consider context. What if we look at two sentences: "They never found the microfilm. The box was in the pen." In that case, it might mean a very tiny box. Then imagine what could happen if we looked at a whole conversation, or a short story?

That's the problem of relevance. AI struggled with it for at least a generation, starting around 1970. It was the defining problem of GOFAI; and the results were never better than dismal.

The other problem seems simpler, but it's ultimately deeper. Because of its roots in symbol manipulation, GOFAI always focused on *linguistic* abilities. Of course, language is crucial to our distinctively human intelligence. But it doesn't follow that language is sufficient for that intelligence, or even that it is possible without other distinctively human capabilities of perception, action, and organization. People can, for instance, make and "read" maps, drawings, and works of art; they can design, build, and operate machines; they can institute intricate social practices and statuses, and maintain them normatively; and so on.

Language and these other "skilful" capabilities are thoroughly intertwined and interdependent; neither side could be what it is without the other. But GOFAI was almost totally oblivious to non-linguistic skills — regarding them, in effect, as mere input/output "interfaces" for the symbol-manipulating system (and so, not part of the problem). But the real reason for this negligence was not that skilful perception and action are unimportant, but that nobody had any serious idea how to implement them.

I don't know whether there is any argument that the knowledge-access problem is unsolvable in principle. Instead, what I think we have learned from the way GOFAI got itself mired down in a tangle, and then collapsed, is that the symbol-manipulation paradigm, which looked so promising at the outset, was in fact misguided. In my own view, the inability to address perception and sophisticated skills is ultimately more telling than the combinatorial explosion.

**CP:** The most discussed argument against the idea that the mind is a computer is Searle's Chinese room argument.<sup>2</sup> With this argument, Searle attempts to show that the mere syntactic manipulation of symbols (what computers are supposed to do) is not sufficient for having semantics and intentionality. What do you think about this argument?

**JH:** Searle is addressing the GOFAI "natural-language processing" systems that I discussed above; but his point is quite different. Whereas I noted the matter-of-fact empirical failure of those systems, and tried to give some insight into why it occurred, Searle argues that, even if they had worked beautifully, they still would not have *understood* anything. In other words, no matter how smart they *seemed*, they would not have had any *real* intelligence at all.

Here's the way he sets it up. A GOFAI system that is supposed to understand some language — Chinese, let's say — consists of four things: (i) a way of taking in and sending out Chinese sentences; (ii) a common-sense knowledge-base (also in Chinese); (iii) a set of rules specifying which Chinese sentences to send out, as a function of those that are received and/or already in the knowledge base; and (iv) a processor that actually follows those rules.

The argument that no such system would really understand Chinese is this. It doesn't matter which code or language the *rules* are expressed in, as long as the processor can reliably follow them. In GOFAI systems, they're expressed in some computer programming language. But they could just as well be expressed in English; and then Searle himself could follow them — which is to say, he could himself play the part of the processor. But that wouldn't mean that he himself understood any Chinese, no matter how fluent in Chinese the system as a whole seemed to be.

And, of course, that's right; but it's also quite beside the point. Nobody ever suggested that the *processor* would understand Chinese. The hope was that a certain kind of complex *system* — of which the processor, the rules, and so on, would be proper parts — could understand natural languages like Chinese. As it happened, that didn't work out; but the failure had nothing to do with the fact that human beings, who speak some other language, could take over various internal functions of the system.

**CP:** Perhaps the attempt to construct a computational machine that can understand natural language might be too ambitious. But, according to a common view, a computational machine applies algorithms by following rules.

In what sense, however, can a computational machine follow a rule?

**JH:** It depends, of course, on what one means by "following". Thus, one could say that the moon faithfully follows a rule like this: always go around the earth in orbit O at speed S. All physical objects follow rules like that; therefore, it's not an interesting point about computers (or people).

Amore interesting kind of case presupposes a way of explicitly formulating any of various different rules — such as a language or some kind of code — and a system that would reliably behave in accord with whatever rule you gave it in that form (at least up to some practical limit). Computers are, of course, prime examples of this kind of system; and, to a certain extent, so are people.

But there is a fundamental difference. People can choose whether or not to follow the rules they are given; but, on the face of it, computers cannot. To be sure, modern computers can execute many different programs at a time; and a higher-priority program (such as a secure operating system) can prevent the execution of a lower-priority one. But the machine itself cannot refuse to execute whatever program has highest priority, so long as the code is well-formed.

Now, it might be argued that the cases are not as different as I'm suggesting. Perhaps, when a person refuses to follow some rule — on the grounds of love, faith, or conscience, say — he or she is just following a higher-priority rule (like an operating system or a security protocol). But I think that's absurd, and even vile.

The ability to tell the difference between what one is told to do and what one *ought* to do, the ability to take *responsibility* for what one values and what one does, is not a mere matter of following rules ("because they're rules"), but rather of standing up and *deciding* who one is. This capacity is, I believe, the deepest essence of humanity. And I am quite confident that computational machines are incapable of it.

**CP:** Is this ability to take responsibility related to *intentionality*? In a recent Paper<sup>3</sup> you sketch a positive account of what is required for a system to have, what you call, "original intentionality" and your main position is that original intentionality presupposes an ability to accept responsibility. What is original intentionality and in what sense does it presuppose the ability to accept responsibility?

**JH:** Intentionality is the character of being "of" or "about" something — in the way, for instance, that many mental states, speech acts, diagrams, pictures, and so on, can be of or about things (even things that don't exist). Thus, intentionality is closely related to meaning. Now, intentionality (or meaning) can be conferred on something from something else that already has it. Such conferred intentionality is what I call "derivative"; and intentionality that is not derivative is original. Until a few decades ago, most philosophers tacitly took it for granted that only the intentionality of mental phenomena could be original. I introduced the original/derivative distinction so that we could articulate that assumption, ask whether it is really defensible, and — above all — investigate alternatives.

To ask how there can be original intentionality is to ask how there can be any intentionality at all: how in the world can anything be about anything? Now, I maintain that this is ultimately equivalent to another question: how can there

be a difference between being right or wrong about things — a difference that depends on those things themselves? For, if there were no way for intentional phenomena to fail, then it would be empty to say that any of them succeeded, and the concept would be useless.

I will try to illustrate this with a quick sketch of a special case: the intentionality of scientific claims and theories.

What separates science from superstition and folklore is that scientists put their claims and theories at risk. The structure of this risk-taking is what we need to understand. As I see it, there are two main factors. On the one hand, there is a subtle and refined experimental practice, including not only skilful practitioners but also sophisticated equipment and procedures. On the other hand, there is a body of theoretical laws and principles, such that only some of the conceivable empirical results would in fact be compatible with those laws and principles.

This two-fold structure implies that there is a serious possibility of internal conflict among the various discoveries and assumptions of the discipline itself; and that ever-present possibility is essential to science as such. For it is only because it consistently "sticks its neck out" in this way that its surviving results are so compelling as accounts of reality. (The reason that astrology, folklore, and the like, are not similarly credible is that they don't stick their necks out.)

But this has implications for what scientists themselves must be like. Someone who didn't give a damn about such internal conflicts couldn't possibly be a scientist. What ultimately drives scientific research forward is the committed effort to resolve those conflicts in an empirically responsible way. And this is why I say that science itself, by its very nature, presupposes the possibility and actuality of genuine human commitment and responsibility.

There is, however, a deeper point; for there are two fundamentally different kinds of commitment and responsibility — not only in science, but in life generally. The first kind, which we might call "ordinary" responsibility, is what I referred to above as "giving a damn" about whether theory and observation are in accord. This is what drives most research and discovery — what Kuhn called "problem-solving" — and mostly it succeeds.

Sometimes, however, even after great effort, it does not succeed — leading eventually to what Kuhn called a "crisis" and Heidegger called "anxiety". Such a crisis is what requires the second and more profound sort of responsibility — ownedness or authenticity, in Heidegger's terms — namely, facing the possibility of having to give the whole thing up (in a radical "paradigm shift", for instance). History shows that this is not an empty prospect.

My point is that this sort of responsibility, too, at least sometimes and in some people, is prerequisite to the kind of original intentionality that is (so far as we know) uniquely distinctive of humanity.

**CP:** In an older paper<sup>4</sup> you hold that the mind should be understood as embodied and embedded in the world. In this way, you think, we can overcome the residual Cartesianism that is almost invisible in many current anti-Cartesian approaches to the mind, such as Davidson's and Rorty's. Could you please tell us a little more about this?

**JH:** Descartes's legacy has several parts. One of these is mind–body dualism, which entails a kind of independence of individual minds from the material world. Most philosophers — including Davidson, Rorty, and I — now reject any such idea. Another part of the same legacy, however, is Descartes's cognitivism: the thesis that people are essentially thinking things. Only a smaller group (many, like me, inspired by Bert Dreyfus) reject this aspect of Cartesianism as well.

According to this minority, much of what constitutes our grasp of the world, and our lives within it, doesn't have intellectual or cognitive content at all. In other words, it's not a matter of beliefs, desires, or other attitudes toward propositions. The kind of counterexample that's easiest to see is skilful know–how — like knowing how to knit, ride a bicycle, or eat with a knife and fork. Clearly, such know–how must be en rapport with the world "here and now", or it wouldn't yield satisfactory results. Likewise, the idea of improving skills wouldn't make sense if the world didn't render a verdict in the form of success or failure.

But this point can be pressed much further. It's not just skills and performances on which the world (eventually) renders a verdict, but also the design of equipment, the structure of institutions, even our sense of what is worth striving for. Indeed, the very vocabulary on which cognitive abilities themselves depend can itself be mastered only in a kind of skilful know–how. If we say that such phenomena do not fall within the categories of knowledge or theory, then so much the worse for those categories. The ultimate challenge is not how best to fit things into our current categories, but how best to adapt our categories to what there is.

---

<sup>1</sup> John Haugeland, *Artificial Intelligence, the Very Idea*. MIT Press 1985

<sup>2</sup> John Searle, "Minds, Brains, and Programs", *Behavioural and Brain Sciences*, 3 (1980), 417–457.

<sup>3</sup> John Haugeland, "Authentic Intentionality" in: Scheutz, M. (ed.), *Computationalism: New Directions*, MIT Press 2002, 159–174.

<sup>4</sup> John Haugeland, "Mind Embodied and Embedded" in Haaparanta, L. and Heinamaa, S. (eds): *Mind and Cognition. Acta Philosophica Fennica*, 58 (1995), 233–267.

---

Published 2006–09–05

Original in English

Contribution by Cogito (Greece)

First published in *Cogito* (Greece) 4 (2006)

© John Haugeland, Costas Pagondiotis/Cogito (Greece)

© Eurozine