From the Philosophy of AI to the Philosophy of Information

Luciano Floridi

Summertime, and a bottle of juice lies half-empty on the grass. Attracted by the smell, wasps get inside it but cannot get out of it and eventually drown. Their behaviour is stupid in many ways: they try to fly through the very surface on which they walk; they keep hitting the glass, until they are exhausted; they see other corpses inside the bottle and yet fail to draw any conclusion; they cannot tell each other about the danger, despite their communication abilities; even if they escape the danger, they do not register it and will come back inside the bottle; they cannot use any means to help the other wasps. If you did not know better, you would think the *vespula vulgaris* to be some kind of robot. Descartes would certainly agree with you.

As a family of insects, wasps got lucky. Had nature produced juice-bottle-flowers, they would have long been extinct. Wasps and their environment have been tuned to each other by natural selection. To us, they are a reminder that fatally stupid behaviour comes in a bewildering variety of forms. Unfortunately, so does intelligence.

Common sense, experience, learning and rational abilities, communication skills, memory, planning capacities: these are only some of the essential ingredients that can make a behaviour (hence the agent so behaving) intelligent. If you think of it, they are all ways of handling information (mind, not symbols or uninterpreted data, but information in the semantic sense of the word, more on this presently). So, could it be that stupid or intelligent behaviour is a function of some hidden information processes? The question is “too meaningless to deserve discussion”, to quote Turing, but it does point in the right direction: information is the key.

Suppose the necessary information processing is already in place: although intelligent behaviour cannot be defined in terms of necessary and sufficient conditions, it could still be tested contextually and comparatively, as Turing rightly understood. After all, we do have a sample of intelligent agents, and that's us, modestly.

Suppose, instead, that the necessary information processing is not yet in place: could it be engineered? If it could, it may be Turing-tested. Yet, whether it can, it is still anybody's guess, or rather faith, despite half a century of research in Artificial Intelligence (AI). One thing, however, seems to be clear: talking of information processing helps to explain why our current AI – or, better, AIB (Artificial Intelligent Behaviour) – systems are overall more stupid than the wasps in the bottle. Our present technology is actually incapable of processing any kind of information, being impervious to semantics. IT is as misnamed as "smart weapons". If you find this puzzling, consider the following example.

Wasps can navigate very successfully. They can find their way in the garden, avoid obstacles, collect food, fight or flee other animals, and so forth. This is already far more than any current AIB system can achieve. A recent confirmation came last March, when 13 vehicles took part to the *Grand Challenge* (http://www.darpa.mil/grandchallenge/), a race sponsored by DARPA (the American Defence Advanced Research Projects Agency). The rough and unpredictable course consisted in 142 miles through of the Mojave desert between California and Nevada. Each vehicle had to navigate unmanned, unaided, un-pre-programmed and without any remote-control, by relying only on its AIB. The prize was $1m to the team whose vehicle was first to cross the finishing line within ten hours. Eventually, the best performance was offered by Sandstorm, a machine built by a team from Carnegie

Mellon University (http://www.redteamracing.org/). Sandstorm managed to navigate 7.4 miles before its tires caught fire.

This is how bad the situation is with state-of-the-art AIB systems. Despite its simplicity, navigation seems to require some sort of intelligence. Nobody really knows how to achieve the same result by means of advanced sensors and computational capabilities.

Sometimes one may forget that the most successful AIB systems are those lucky enough to have their environments shaped around their limits, like a robomower (http://www.friendlyrobotics.co.uk/), not vice versa. Put artificial agents in their digital soup, the Internet, and you will find them happily buzzing. The real difficulty is to cope with the unpredictable world out there, which may also be full of other collaborative or competing agents. This is known as *the frame problem*: how a situated agent can represent a changing environment and interact with it successfully. Nobody has much of a clue, so human intervention is constantly required, as with the robots on Mars (http://marsrovers.jpl.nasa.gov/mission/spacecraft_rover_brains.html). Our most successful artificial agents operating in the wild are those to which we are related as homunculi to their bodies.

Consider now the explanation of AI failure, namely the lack of information processing capacities. Our current computers, of any architecture, generation and physical making, analogue or digital, Newtonian or quantum, sequential, distributed or parallel, with any number of processors, any amount of RAM, any size of memory, whether embodied, situated, simulated or just theoretical, never deal with information only with data. No philosophical hair-splitting here. Data are mere (patterns of physical) differences and identities. They are uninterpreted and tend to stay so, not matter how much they are crunched or kneaded. Nowadays, we think of data in Boolean terms –

ones vs. zeros, ups vs. downs in the spin of an electron, high vs. low voltage – but of course artificial devices can detect and record analogue data equally well. The point is not the binary nature of the vocabulary, but the fact that strings of data can be more or less well-formed according to some rules, and that a computer can then handle the latter rather successfully. So, whenever the behaviour in question is reducible to a matter of transducing, encoding, decoding or modifying uninterpreted data according to some syntax, computers are likely to be successful. This is why they are often and rightly described as purely syntactic machines.

Of course, "purely syntactic" is a comparative abstraction, like "virtually fat free". It means that traces of information are negligible, not that they are completely absent. Computers are indeed capable of (responding to) elementary discrimination. They can detect identities as equalities and differences as simple lacks of identities between the relata (but not in terms of appreciation of the peculiar and rich features of the entities involved). Admittedly, this is already a proto-semantic act. So, to call a computer a syntactic machine is to stress that discrimination is a process far too poor to generate anything resembling semanticisation. It only suffices to guarantee an efficient manipulation of syntactically-friendly data. Given that it is also the only vaguely proto-semantic act that (present) computers are able to perform as "cognitive systems", the *Grand Challenge* resembles more a *Mission Impossible*.

Problems become immediately insurmountable when their solutions require the successful manipulation of information, that is, of well-formed data that are also meaningful. Semantics is the snag. How do data acquire their meaning? Solving what is known as the *symbol grounding problem* in a way that could be effectively engineered would be a crucial step towards solving the frame problem. Unfortunately, once again we still lack a clear understanding of how precisely the symbol grounding problem is solved in animals, including primates like us, let alone having a blue print of a physically implementable approach.

What we do know is that processing information is exactly what intelligent agents like us are good at. So much so that fully and normally developed human beings seem cocooned in their own informational space. Strictly speaking, we do not consciously cognise pure meaningless data. The genuine perception of completely uninterpreted data might be possible, perhaps under very special circumstances, but it is not the norm, and cannot be part of a continuously sustainable, conscious experience, at least because we never perceive pure data in isolation but always in a semantic context, which inevitably forces some meaning onto them. We are so used to dealing with rich semantic contents that we mistake dramatically impoverished or variously interpretable information for something completely devoid of any semantic content. Yet what goes under the name of "raw data" are data that might lack a specific and relevant interpretation, not any interpretation.

To sum up, data, as (interpretable but still) uninterpreted (patterns of physical) differences and identities, represent the semantic upper-limit of current and foreseeable AIB systems. They also are the semantic lower-limit of natural intelligent behaviour (NIB) systems, which normally deal with (semantic) information. Ingenious layers of interfaces exploit this threshold and make possible human-computer interaction.

The suggestion concerning human informational-cocooning and machines' data-entrapment becomes less controversial once is carefully distinguished from five theses that it does not deny. One may argue that:

1) young NIB systems, for example Wittgenstein's young Augustine, seem to go through a formative process in which, at some stage, they experience only data, not

information. There is a stage in the history of a human being at which we are information virgins;

2) adult NIBs, for example Turing's clerk, the adult John Searle or a medieval copyist, could behave or be used as if they were perceiving only data, not information. One could behave like a child – or an Intel processor, or a Turing Machine – if one is placed in a Chinese Room or, more realistically, while copying a Greek manuscript without knowing even the alphabet of the language but just the physical shape of the letters;

3) cognitively, psychologically or mentally impaired NIBs, including the old Nietzsche, might also act like children, and fail to experience information (like "this is a horse") when exposed to data;

4) there is certainly a neurochemical level at which NIBs process data, not yet information;

5) NIBs' semantic constraints might be comparable to, or even causally connected with, AIBs' syntactic constraints, at some adequate level of abstraction.

These five theses are perfectly fine and consistent with the point made above, which is that (current) AIBs' achievements are constrained by syntactical resources, whereas NIBs' achievements are constrained by semantic ones.

There is a semantic threshold between us and our machines and we do not know how to make the latter overcome it. Indeed, we know very little about how we ourselves build the cohesive and successful informational narratives that we inhabit. If this is true, then artificial and human agents belong to different worlds and one may expect them not only to have different skills but also to make different sort of mistakes. Some evidence in this respect is provided by the Wason Selection Task.

Imagine a pack of cards where each card has a letter written on one side and a number written on the other side. You are shown the following four cards: [E], [T], [4],

[7]. Suppose, in addition, that you are told that if a card has a vowel on one side, then it has an even number on the other side. Which card or cards – as few as possible – would you turn over, in order to check whether the rule holds?

While you think about it, it may be consoling to know that only about 5% of the educated population gives the correct answer, which is [E] and [7]. However, most people have no problems with a semantic version of the same exercise, in which the rule is “if you borrow my car, then you have to fill up the tank” and the cards say: [borrowed the car], [did not borrow the car], [tank full], [tank empty].

In both cases, a computer obtains the correct answer by treating each problem syntactically. The test reminds us that intelligent behaviour relies on semantic understanding more than on syntactical manipulation and that, while both can easily achieve the same goals efficiently and successfully, semantically- and syntactically-based agents are prone to different sort of potential mistakes.

All this should be fairly trivial, yet it is still common to find people comparing human and artificial chess players. In 1965, the Russian mathematician Alexander Kronrod remarked that chess was the fruit fly of artificial intelligence. This may still be an acceptable point of view had AI tried to win chess tournaments by building computers that (learn how to) play chess the human way. But it hasn’t, and as a result the chess-fly has caused some conceptual confusion.

Playing chess well requires quite a lot of intelligence if the player is human, but no intelligence whatsoever if played computationally. When IBM computer *Deep Blue* won against the world chess champion Garry Kasparov in 1997, it was a sort of Pyrric victory for classic AI (http://www.sciencemag.org/cgi/content/full/276/5318/1518). *Deep Blue* is only a marvelous syntactical engine, with a great memory but virtually zero AIB.

John McCarthy, one of the fathers of AI, immediately recognised that *Deep Blue* said more about the nature of chess than about intelligent behaviour. He rightly complained about the betrayal of the original idea (http://www.sciencemag.org/cgi/content/full/276/5318/1518), but he drew the wrong lesson. Contrary to his suggestion, AI should not try to *simulate* human intelligent behaviour. This is the glass we should stop hitting. AI should try to *emulate* its results.

Emulation is not to be confused with some form of functionalism, whereby the same function – lawn-mowing, dish-washing, chess-playing – is implemented by different physical systems. Emulation is rather connected to "outcomism": agents emulating each other can achieve the same result by radically different strategies and processes. The end underdetermines the means.

Outcomism is techologically fascinating and rather successful, witness the spreading of IT in our society. Unfortunately, it is eyes-crossingly dull when it comes to its philosophical implications, which can be summarised in two words: "big deal". So should this be the end of our interest in the philosophy of AI? Not at all.

The failure of mimetic AI (AIB must simulate NIB) has been conceptually very fertile. By showing that what matters is information, the philosophy of AI has hushered in a new paradigm, the *philosophy of information*.

Elsewhere I have defined PI as the philosophical field concerned with (a) the critical investigation of the conceptual nature and basic principles of information, including its dynamics, utilisation and sciences, and (b) the elaboration and application of information-theoretic and computational methodologies to philosophical problems. It would be impossible to analyse here critically and in detail the nature of PI, but a reference to the *Blackwell Guide to the Philosophy of Computing and Information* (http://www.blackwellpublishing.com/pci/) and three schematic points may suffice to give a general idea.

First, by trying to circumvent the semantic threshold and squeeze some information procesing out of mechanics and syntax, AI has opened up a large and very rich variety of research areas, which are conceptually challenging *per se* and very interesting for their potential implications and applications in philosophy. Part of this innovation goes under the name of *New AI*. Consider, for example, situated robotics, neural networks, multi-agent systems, Bayesian systems, machine-learning, cellular automata, artificial life systems, epistemic logic and non-monotonic reasoning. Philosophical issues no longer look the same once you have been exposed to any of these fields.

Second, ironically, artificial simulations have failed to reproduce NIB, but have made available environments where philosophical theories can be simulated and tested "in silico". This is true not only for logic-based problems, as one may expect, but also for ethical, linguistic and epistemological issues, for example, which can be modelled through digital simulations.

Third, by realizing that it is not so much the process (computing) that has revolutionized our society and our conceptual schemes, as the broader phenomenon of information, the philosophy of AI has helped to call attention to new conceptual problems that require our philosophical attention. Computer ethics is a good example.

So the Philosophy of AI has not lived in vain. Obviously there is plenty of exciting work that lies ahead. All we need to do is to replace an I for an I.