Artificial Intelligence: The harder you try to define it the more you understand it.

Maxwell J. Roberts
July 2024

Downloaded from:
www.tubemapcentral.com/writing/webarticles/Roberts_AI-Definitions_2024.pdf

Also at LinkedIn:

https://www.linkedin.com/pulse/artificial-intelligence-harder-you-try-define-more-maxwell-roberts-fj9cc/

More articles at:

www.tubemapcentral.com/writing/writing.html

Artificial Intelligence is difficult to define but the various attempts over the decades can be placed in one of three categories: task-based, behaviour-based and process-based. Being able to identify a definition that fits the third category is important because people can be fooled that software is performing intelligently even when it is clearly using dumb procedures. A perfect definition of Artificial Intelligence might be difficult or impossible to attain, but the very act of attempting to formulate one increases our understanding.

Cartoon by BAT ©2024

It's becoming a cliché to say that Artificial Intelligence (AI) has been in the news lately, but has AI *really* been in the news? That depends on how we define It. This is difficult but it is important to try because, even if we can't come up with a perfect definition, the very act of attempting this gives many insights into whether we are either on the verge of incredible breakthroughs or else being served up with yet another dose of snake oil with a more sophisticated flavour than before.

If we start with the media definition of AI: *Artificial Intelligence is when computers do really cool stuff that looks clever*, it is not difficult to see that a definitional equivalent of 'Trip Advisor ratings' will get us nowhere. Current AI packages can do all sorts of useful things but it is very easy to be fooled by impressive results. This is called the Eliza effect, named after an extremely stupid 1960s computer program that, despite its limitations, still managed to fool people into thinking that they were conversing with a human — provided that they were off their guard and not interacting for any length of time. We need to attempt a better definition, but that is where the problems start.

Human intelligence is hard enough to define — although psychologists have made enormous strides in understanding it over the last few decades — but at least we know it exists. Machines are not by their nature intelligent: they have to be configured to be that way by intelligent humans. Hence, defining whether or not a machine is intelligent must take account of the fact that, no matter how spectacular it may seem on the surface, inside the case it may not be intelligent at all.

When looking through the various definitions of AI that have been attempted, they can be categorised into three basic types: behaviour-based, task-based and process-based. Each, in turn, can be sub-divided into lowbrow and highbrow versions. All have strengths and weaknesses. [See, also, Note 1, at the end of this article.]

Behaviour-based definitions. *A piece of software displays intelligence if it exhibits certain characteristic behavioural qualities*. In the literature we see criteria ranging from demonstrates *adaptive behaviour in its environment* to passing some sort of corruption of the *Turing Test*. For this type of definition, the exact domain of operation is not specified, so that the software would be expected to behave in the same intelligent way no matter what task or environment it is given. The trouble with behaviour-based definitions is that:

(1) if the scope of the definition is modest, then modest behaviour becomes intelligent by definition. For example, a simulated rat that learns to avoid simulated flavours associated with poisons could be argued to fulfill an adaptivity criterion, and therefore has attained intelligence;

(2) ultimately, these definitions rely on human evaluations of computer performance, requiring that the human has sufficient knowledge and intelligence to be a skilled arbiter of the output.

Task-based definitions. *A piece of software displays intelligence if it is able to perform tasks that would require intelligence if they were undertaken by a human.* These types of definition must make explicit what sorts of tasks will act as markers of intelligence and, inevitably, run into difficulties as a result. One problem is that tasks which are difficult for machines – such as seeing and/or navigating a complex environment – tend to be so easy for humans that it doesn't even occur to them that they are being intelligent while they perform them. Conversely, many tasks that are taken to point towards the pinnacle of human achievement, such as playing chess, are not too difficult for machines to attain or surpass good human-level performance. This is what I call the *baseline of Intelligence problem*: should a definition either encapsulate the basic cognitive foundations underlying what people can do, and call that intelligence, or else encapsulate the pinnacle of human achievements? The other difficulty with task-based definitions concerns the generality the specification. Identifying a particular, well-defined task can result in computers that are really good at this and nothing else: human intelligence is *general* by definition. Identifying a more vaguely specified task, such as effortlessly making conversation in any situation, is more in the realm of the behaviour-based definitions above.

Process-based definitions. *A piece of software can be said to be intelligent if it implements intelligent decision making processes.* At first sight, this type of definition seems merely to shift the issue. Instead of the difficulties of defining whether the system as a whole is intelligent, now we have to decide whether the processes that it implements qualify. This is a valid objection but, even so, the mere knowledge of the existence of this definitional category, and why it is important, transforms our ability to evaluate AI. The need to think about whether processes are intelligent alerts us to the possibility that the computer might be attaining impressive results by cheating somehow. For example, a close look at the dark side of AI programming reveals numerous *ad hoc* patches (also known as *kludges*) — these are cheats that intercept and prevent inappropriate or erroneous outputs — added out of necessity by an intelligent human programmer. Even without these, we will later see that some decision making processes are, arguably, nothing to do with intelligence irrespective of what the system achieves. If it is possible to identify non-qualifying tricks and procedures then we can, at the very least, formulate useful negative process-based definitions that can enhance our evaluative powers. For example, *if a computer output depends on X then AI has NOT been implemented*. [See, also, Note 2, at the end of this article.]

Let's look at current generative AI, such as ChatGPT — the most famous of the various AI attempts at the time of writing. The way this software works – optimistically named *deep learning* – is based on principles and architectures that have been investigated since the 1980s. Information is stored in a computer-simulated network of interconnected nodes which, at a very superficial level, might be said to bear some sort of resemblance to neurones in the human brain. The network is configured (or trained) by presenting it with information (e.g., textual for ChatGPT, or visual) that has been encoded by humans, and the network is prompted to generate outputs (e.g., to queries such as: *What is a valid next word in this sentence? What is the object in the picture?*). If the computer gives an inappropriate output, as pre-determined by the humans, then it is reconfigured by mathematically calculating the distance between an acceptable output and the actual output, and feeding this back into the network to adjust the interconnectedness of the nodes, so a more appropriate output is slightly more likely next time. Repeat and repeat and repeat. And that's it.

ChatGPT is not clever in a human/intelligent processes sense, it has merely been configured to perform *statistical language extrapolation*: As a result of directed trial and error configuration an enormous multiple regression equation has effectively been created. High level concepts such as truth, understanding, and morality are only displayed in the output to the extent that they were statistically detectable in the input, and also were required for responses during the configuration phase. If the computer demonstrates erudition this is only because (1) intelligent human writers exhibited erudition previously in time and (2) another intelligent human included this writing in the input materials and (3)implemented the configuration in such a way that the computer was able to (4) identify and encode a statistical relationship that (5) enabled it to output a combination of words that (6) yet another intelligent human was able to judge as being erudite (or not). Effectively, the current protagonists of AI are attempting to compile an ersatz version of *Wikepedia* by conducting a statistical evaluation of everything that they can get their hands on.

Artificial Intelligence is usually at its strongest when the intention is entertainment rather than accuracy. However, sometimes the outputs are entertaining for the wrong reasons. Here are some recent results from my prompts for an Art Nouveau railway station built in Lego. All of the results were bizarrely weird in their failings but my favourite is the one at the top. Here, a minifigure with a bizarrely disassembled face stares blankly outwards from the mayhem in an almost Magritte-esque parody.

Now we know the full process, we can see that ChatGPT might be a useful tool in some circumstances, but to claim that statistical language extrapolation (along with any added patches to override the output as necessary) embodies the concept of intelligence stretches the notion somewhat. This methodology is simply not psychologically plausible as a model of intelligent behaviour. If we want a name for the current state of

the art, *Artificial Plagiarism* would be more apt. *Artificial Throwing-Statistically-Related-Stuff-Together-and-Hoping-for-the-Best* is more of a mouthful but even closer to the truth.

My own attempts at building an Art Nouveau railway station out of Lego, designed from my knowledge and love of the genres. All this without any assistance from computer statistical image assembly.

Returning to process-based definitions, cognitive scientists have, unfortunately, still not advanced very far in terms of understanding the processes underlying human intelligence. However, the repeated failures of previous AI attempts to deliver the goods, from the 1950s to 1980s, have supplied some important clues. Humans are relatively proficient in open, unconstrained environments. They can use inductive reasoning to fill out the gaps, and draw upon common-sense knowledge to prevent really dumb inferences and actions. Computers are relatively proficient in closed, structured environments, can use statistical inferences to fill out the gaps, but even here may need *ad hoc* patching (programming cheats) to prevent really dumb inferences and actions.

Suppose you switch on a light and, simultaneously, a car backfires. In a fleeting moment you believe that your action caused the explosion outside. Then you collect yourself. Your common sense tells you that the light switch wasn't connected to the car so it cannot possibly have caused the explosion. This causal inductive reasoning, tempered with common sense, is really difficult to specify logically and implement in a computer. Either hundreds of thousands of individual rules about the basics of the world must be programmed, or else statistical relationships must be inferred from real world data (and intelligent humans will still need to patch the system manually to prevent inappropriate responses that result from spurious correlations). Think about the following two relationships, each one expresses a correlation:

quantity of ice cream sales at the seaside will correlate with numbers of jellyfish stings

quantity of ice cream sales at the seaside will correlate with numbers of wasp stings

Humans will have not too much difficulty understanding why the first correlation is a statistical coincidence and that the second encompasses a directly causal chain of events. In the first instance, ice cream sales increase in hot weather, and people are more likely to swim in the sea in hot weather, but there is not a direct causal connection for ice cream to attract jellyfish stings. In the second, sweet sugary ice-cream will attract wasps, and this will put the people eating ice cream at direct risk of being stung. Distinguishing these relationships by using statistical extrapolation is far harder for computers. Even if such a system demonstrates success, it would be pure fantasy to infer from correct output that the computer has any conceptual understanding of the subtleties of causal relationships.

Is human common-sense reasoning also a patch: an ad hoc intervention to override spurious inference? Analogously, perhaps, but the mechanism by which it operates is nothing like the ad-hoc kludges currently used to stop AI systems from falling over. There is something effortless about the way in which human knowledge is interconnected and easily associated, coupled with an inherent understanding about how

causality operates in the real world, that is nothing like clunky computer fixing. Of course, humans are not perfect but, on balance, their failings are generally far more forgivable than computer mishaps. Take the following example of how a computer might need to be patched if it fails to answer a simple question.

Q: If Jenny is in Egypt, is Jenny's left foot in Egypt?

Computer: ???

AI programmer adds this patch: If a person is in any location, and there is no reason to believe left foot is disconnected or missing, then assume left foot is in same location

Q: If Jenny is in Egypt, is Jenny's right hand in Egypt?

Computer: ???

AI programmer adds this patch: If a person is in any location, and there is no reason to believe that body part is disconnected or missing, then assume body part is in same location

Q: If Jenny is in Egypt, is Jenny's mobile phone in Egypt?

Computer: ???

Returning to how to conceptualise Artificial Intelligence, definitions are essential because researchers need to know what to aim for, and we need to be able to evaluate their success: How can we decide whether AI has been achieved if we do not know what AI should be? AI is big business, whether in the form of research grants or potential profits, so all claims need to be evaluated bearing in mind that practitioners of AI have clear vested interests in overstating their achievements and potential. This is why process-based definitions are so important: What is really happening inside that black box, and is it using plausibly intelligent methods to achieve its output? Furthermore, if we insist on process-based definitions, this might force improved transparency and honesty from people selling their products. The need to define, coupled with an awareness of the different types of definition, also (1) alerts us to wider issues concerning the efficacy of AI; (2) hints at methods we might use to stress-test the systems to see whether they are really as clever as is being claimed and (3) enables us to make more informed predictions of future developments. Hence, although AI is difficult to define, from the point of view of a sceptical observer this is not a serious stumbling block. We don't really need to worry about whether a behaviour-, task- or process-based definition is 'correct'. Provided we understand the different types of definition, and their strengths and weaknesses, our ability to evaluate AI for ourselves improves beyond recognition.

My lecture notes for my module on intelligent behaviour can be downloaded here: http://www.tubemapcentral.com/speaking/ib.html

For a thorough understanding of the philosophical, psychological, and practical issues underlying Artificial intelligence, try these books. Not much has changed since they were written. Computers might be bigger and faster today, but the differences then and now are merely quantitative, not qualitative.

Partridge, D. (1991). *A New Guide to Artificial Intelligence*. Ablex.

Copeland, B.J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Blackwell.

Ekbia, H.R. (2008). *Artificial Dreams: The Quest for Non-Biological Intelligence*. Cambridge University Press.

One of my favourite books on human intelligence is also one of the most readable.

Duncan, J. (2010). *How intelligence happens*. Yale University Press.

[**Note 1:** Once we can genuinely get away from the Von-Neumann architecture that is used for every single computer on every single desk, and fully functional quantum or biological computers become a reality, we might start to see **hardware-based definitions** of AI as part of the available options. This is still science-fiction territory at the time of writing.]

[**Note 2:** Chess is often taken to be a marker of human intelligence because, cognitively, we are ill-equipped to play it well, and we need to devise intelligent methods to overcome our own limitations. Computers are now able to store hundreds of thousands of reference games and predict so far ahead that they have no need for intelligent gameplay strategies. They can defeat humans by using brute force: dumb gameplay techniques that are applied so exhaustively that humans cannot compete.]