Why we need 'Conscious Artificial Iintelligence'

Hans Peter Willems
MIND|CONSTRUCT
May 2012

Abstract

In this paper we will look at some of the reasons for needing the development of 'conscious' or 'self-aware' Artificial Intelligence, also known as 'Strong-AI' or Artificial General Intelligence (AGI), and the apparent risks involved in doing so. Throughout this paper it will become apparent that the need for this development is closely related to taking control of the risks. I will present a practical definition of 'conscious AI', that will suffice within the scope of this paper, without going into the debate about what 'consciousness' actually is or what constitutes consciousness.

Keywords: Artificial Intelligence, Consciousness, Singularity, Coherent Extrapolated Volition.

Introduction

Before we can delve into needs and risks, it is necessary to have a working definition of 'conscious AI', at least within the scope of this paper. The discussion on what 'consciousness' specifically constitutes still rages on today, as it has been since the philosophers took hold of the topic. However, for this paper we will refrain from trying to give a universal definition of consciousness; for AI-development the definition does not have to be universal, nor applicable to humans for explanation of human consciousness: '*in this case, simulation of internal processes in enough detail to replicate approximate patterns of the system's behavior*' [David Chalmers, 2010]. What we do need is a definition that can be used to model conscious AI towards 'human-

like' conscious behavior, as far as aspects of human consciousness are useful and applicable in Artificial Intelligence.

Towards a practical definition of human-like Artificial Consciousness

Within the scope of this paper we define consciousness as the ability to 'rise above programming'. We can say that humans have initial programming (instinct) and are capable of rising above it. When we talk about conscious artificial intelligence we therefore talk about artificial intelligence that can rise above it's (initial) programming, or even preferable, that can develop useful behavior without specifically being programmed for that behavior. We want it to be 'human-like' so we can identify with it [Becker, et al, 2007], and the conscious AI can identify with us humans. This means that a 'certain approximation' of human-likeness will be enough to consider this conscious AI to be 'human-like'. Going from here we can now define consciousness in Artificial Intelligence to be human-like when it instantiates behavior that gives us a recognizable human-like experience. Without the need to specifically define the traits that make us human, we can say that we experience consciousness as that which makes us human in our overall behavior.

I acknowledge beforehand that this definition is a narrow one, and doesn't take into account all the complexity and inevitable problems that needs to be solved to create consciousness in Artificial Intelligence. However, for sake of argumentation towards the need for conscious AI, we choose to use such narrow definition.

So for the scope of this paper we can now define artificial consciousness as that which makes AI appear human-like in it's behavior. However, the above given definition obviously leads to the implementation of human-like behaviors and abilities like reasoning, planning, ambition and even free will. This is already hard to 'control' in humans, so the risks in implementing such behavior and abilities into Artificial Intelligence should be evident.

Risks involved in Conscious AI

We only have to take a short look at humans to see what problems might arise from implementing human-like abilities and behavior into Artificial Intelligent systems. If we add to that the possibility of Artificial Intelligent systems becoming much more intelligent than humans [David Chalmers, 2010], and a doomsday scenario is starting to take form. Let's take a look at the most obvious threats that might emerge from the advent of 'conscious machines':

Competing for resources

Like Humans, machines need resources to be able to operate, and conscious machines will be no exception. Currently, when we don't have the resources for a machine to run, we can decide to switch it off or unplug it. But will a conscious machine allow us to do so. As consciousness is linked to free will, reasoning, planning and ambition, a conscious machine might demand it's needed resources, maybe even fight for it. In any form it is imaginable that conscious machines could challenge humans for the available resources.

Will conscious machines be able to challenge us? That remains to be seen and is obviously closely coupled to the 'utility' that we would give these machines. However, the roles where conscious machines would be useful or even excel, could very well be the roles that would give them this utility, and therefore ability, to challenge us. It is hard to predict where this could go because we can hardly imagine all possible roles that conscious machines could fill. From this we can at least conclude that a certain risk is foreseeable.

Unfriendly Artificial Intelligence

Let's take the 'compete for resources' problem one step further; it could very well come to the point that the conscious machines see humans, or even all of humanity, as a threat to their own existence [Eliezer Yudkowsky, 2006]. Albeit for lack of

resources, 'just the idea' that we humans pose a risk, or any other yet unimaginable reason for the conscious machines to really 'not like humans'; we are now talking about a threat to humanity.

If a conscious machine has free will, can plan it's actions, has desires and ambitions, all these things might actually turn against us. And as we made the machine conscious, and in doing so gave it control over it's own actions, we will have in all probability given away the key that could 'switch it off'.

Super intelligence & human's last great invention

It has been stated that strong-AI or AGI will be the last great invention of humankind [Nick Bostrom, 2003]. As soon as we build a computer that is as intelligent as humans, the advancement of computer processing power will (very) soon afterward result in that human-like computer to design a superhuman-like computer [David Chalmers, 2010]. Humanity will then be faced by a 'super intelligence'. This super intelligence will possibly think every original thought before a human can think it, it will explore our world, science, the universe faster then we humans can keep up with. Eventually, we humans will simply be obsolete, no longer able to contribute to our own world, surroundings or just anything that gives 'reason' to humanity.

Solutions to the risks and other needs for Conscious AI

The 'simple' solution to the aforementioned problems seems to be just not to build it. Let's just stay away from implementing consciousness in Artificial Intelligence and all will be fine. This is the 'prohibition defense' and is easy to falsify; Both nature and human history has shown that development (in any area) will occur anyway. This can be the result of evolutionary systems, ill-informed human interaction or just plain stupidity. Nevertheless, it will happen. So if we don't do this development in a controlled manner and with predefined goals and intentions, it will happen in an uncontrolled manner without (well thought out) goals and possibly very wrong

intentions [Eliezer Yudkowsky, 2004]. So it should be clear that we need to take this on with controlled development instead of uncontrollable prohibition.

Coherent Extrapolated Volition (CEV)

Several ideas, like the three laws of robotics and similar (more serious) efforts have been stated [Robin Murphy, et al, 2009] as a solution to keep the behavior of Artificial Intelligence under control. However, a strong case has been made for the probable failure of such scenarios [Eliezer Yudkowsky, 2004]. Instead we should build Artificial Intelligence with the sense of 'aiming to be good or correct' within any frame of reference that is applicable within the current or (future) actual environmental situation [Eliezer Yudkowsky, 2004]. But when we talk about 'the sense of...', we are automatically introducing experienced phenomena into the equation. And for that to be possible in a Artificial Intelligence, it has to be conscious. Consciousness, therefore, is the key ingredient for actually being able to implement CEV in the first place. This seems a great paradox: we need conscious AI to be able to implement CEV, while we also need CEV to be able to implement 'safe' conscious AI. But instead of a paradox, I suggest it is a harmonious coincidence. The solution for safe conscious AI is directly linked to having conscious AI in the first place. And because we already established that AI will sooner or later turn into conscious AI, this means that to have safe AI, it needs to be conscious AI.

Autonomous operation

One of the main problems that is seemingly hindering serious applications of Artificial Intelligence, is the simple fact that those existing systems can only perform the exact feat that they are programmed for. Even systems that have some sort of capability to self-learn, are doing so based on previously programmed functionality to learn within a specific, and again previously determined, domain of knowledge and/or application. What sets strong-AI apart is the 'general' part in the term

'Artificial General Intelligence'; the ability not only to operate autonomous within a predefined area of application, but to be completely autonomous within any general field of application and handle *domain-independent skills necessary for acquiring a wide range of domain-specific knowledge* [Peter Voss, 2002]. To reach this goal, we need a system that is capable of defining it's own rules towards it's own development in these general areas of application. Only real (self-)conscious systems are capable of (re)adjusting their own frame of reference and tune their own rules within this inner frame of reference. Therefore the implementation of consciousness is the only viable way towards 'general intelligence'. Artificial Consciousness is the only predefined framework that is capable of adapting towards new goals, new knowledge, new ways of reasoning about possibilities, have actual insights and finally being able to self-implement the needed adaptions and actions to follow up on those insights. Only consciousness can grow beyond it's initial programming.

Integration into society

Our society is a human society; it is ultimately tuned towards humans. For any non-human intelligence to be able to integrate effortlessly into this society, this non-human intelligence should be adapted as close as possible to this human-oriented tuning of our society [Christian Becker, et al, 2007]. It should therefore be obvious that the more human-like this intelligence will be, the more adapted it will be to our society.

This leads us to the role of consciousness in our society. Human society is what we, as humans, perceive as our current reality. We interact with this reality in ways that form, tune and readjust that what we perceive as our society. This, in itself, is a process that is totally driven by consciousness. So consciousness leads to perception of reality, which in turn leads to possible interaction with that perceived reality. For any non-human intelligence to be successfully integrated into our society, it must be capable of interaction with our perceived reality. Therefore it must be able to perceive

this reality by itself and inevitably be 'conscious' to be able to do so.

Specific applications in need of consciousness

Several things that humans are capable of doing, involve consciousness. It eventually comes down to pure 'understanding'; not the way that understanding is modeled into ontologies or other forms of semantic databases, but the way we understand something based on experiences, feelings and both objective and subjective perception of the world around us [Sidney K. D'Mello, et al, 2007]. Any task that needs human-like understanding, needs therefore conscious behavior. So if we want Artificial Intelligence to be applicable in these specific areas, it need to be conscious to be able to understand in this way.

One obvious example is machine translation. It is, to a certain extend, possible to implement grammatical rules that can tackle sentence construction, word sense disambiguation and even semantic meaning in different languages. However, it is totally unfeasible to harness the complexity of folklore based meanings of words and sentences, or the choosing of wording based on the current emotional state of a person that is using those words, into previously defined rules. To have human-like translations, we need human-like understanding. No formal description of language, no matter how elaborate and/or complex it is defined, will be a viable substitute for conscious understanding.

Assessment of the level of intelligence

Probably of minor importance, but nevertheless relevant, is the way we measure human intelligence and how this measurement applies to Artificial Intelligence. It has been argued that the 'Turing test' [Alan Turing, 1950] is insufficient for determining the real level of 'intelligence' in a machine [Blay Whitby, 1997]. It seems that we need a better way to assess the level of intelligence in artificial implementations.

The obvious test would be to see if the machine measures up to our own perception

of human intelligence. Such a test should go beyond a simple measurement of available knowledge, as by that standard a public library is much more intelligent than a human; it simply holds more knowledge than an average human. In the end we would not be able to keep consciousness out of the equation, as conscious interaction with both previously processed knowledge and, yet to be interpreted, new knowledge and perceptions of our reality, is the basis of our 'intelligence'.

Conclusion

I suggest that consciousness in Artificial Intelligence is not only desirable, but a direct necessity, if we want Artificial Intelligence to succeed as human-like participants in our society. Without consciousness, AI will remain 'just a machine' and therefore never integrate into our society on any higher level than machines currently do. And if we don't develop this ourselves, in a controlled manner, it will manifest itself sooner or later in a form that integrates into our society in ways that might be very undesirable for humanity. Implementing consciousness into the machine might seem like 'giving up control', but in the end it might be the only way to have any control or even 'at least some control'.

References

- David J. Chalmers (2010). The Singularity, A Philosophical Analysis. Journal of Consciousness Studies, 17, No. 9–10, 2010, pp. 7–65. http://www.imprint.co.uk/singularity.pdf
- Christian Becker, Stefan Kopp, and Ipke Wachsmuth. (2007). Why emotions should be integrated into conversational agents. In Toyoaki Nishida (Ed.), Conversational Informatics: An Engineering Approach (pp. 49-68). Wiley. http://www.techfak.uni-bielefeld.de/a ... oc/BeckerEtAl_ConvInf.pdf
- Eliezer Yudkowsky (2006). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In Global Catastrophic Risks, eds. Nick Bostrom and Milan

Cirkovic.
http://singinst.org/upload/artificial-intelligence-risk.pdf

- Nick Bostrom (2003). Ethical Issues in Advanced Artificial Intelligence. In Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, pp. 12-17. http://www.nickbostrom.com/ethics/ai.html
- Eliezer Yudkowsky (2004). Coherent Extrapolated Volition. The singularity institute, singinst.org. http://singinst.org/upload/CEV.html
- Robin R. Murphy, David D. Woods (2009). Beyond Asimov: The Three Laws of Responsible Robotics. EEE Intelligent Systems Volume: 24, Issue: 4, Publisher: IEEE Computer Society, Pages: 14-20. http://ts-si.org/files/IEEEIS_WebExtra-0709.pdf
- Peter Voss (2002). Essentials of General Intelligence: The direct path to AGI. In Artificial General Intelligence, Goertzel, Ben; Pennachin, Cassio (Eds.), Springer, 509 p. http://www.adaptiveai.com/research/index.htm
- Sidney K. D'Mello and Stan Franklin (2007). Exploring the Complex Interplay between AI and Consciousness. AAAI Fall Symposium on AI and Consciousness: Theoretical Foundations and Current Approaches. http://www.aaai.org/Papers/Symposia/F ... /FS-07-01/FS07-01-009.pdf
- Turing, A.M. (1950). Computing machinery and intelligence. Mind, 59, 433-460. http://www.loebner.net/Prizef/TuringArticle.html
- Blay Whitby (1997). The Turing Test: AI's Biggest Blind Alley? In: Machines and Thought: The Legacy of Alan Turing. Mind Association Occasional Series, 1 . Oxford University Press, USA, pp. 53-62. http://www.sussex.ac.uk/Users/blayw/tt.html