11

TEACHING ETHICS TO ROBOTS. TRYING TO TEACH MORALITY TO ARTIFICIAL INTELLIGENCE

Eduard Kaeser, Switzerland

Machines[258] are taking on more and more tasks. But who decides how they should behave ethically? A team of computer scientists from Washington University caused quite a stir in 2021 with an ethical algorithm called "Delphi". It is an artificially intelligent (AI) system based on deep learning that "assesses" human behaviour. For example, eating habits. If you enter "eat pork" the system comments with "that's fine"; but to "eat worms", it replies "that's disgusting". Likewise, Delphi acknowledges ethically relevant behaviour "rejecting weakness", is "bad", but "rescuing a drowning child when you can't swim" is "good".

[258] First published in German: Eduard Kaeser, Würmer essen ist «widerlich», Wokeness ablehnen «schlecht»: Vom Versuch, künstlicher Intelligenz Moral beizubringen, Neue Zürcher Zeitung, 31 Oct .2022. Translation to English by the editors. © Globethics Publications, 2023 | DOI: 10.58863/20.500.12424/4276024 | CC BY-NC-ND 4.0 International.

The AI system is in the experimental phase. For the time being it is circulating as an app called "Ask Delphi".

11.1 A Corpus of Ethical Judgments Should Teach Algorithms

The designers in no way claim that Delphi is morally competent. Nevertheless, their ambition goes far beyond the development of digital bells and whistles. One team member, Liwei Jiang, speaks of a "commonsense moral model" with a "robust performance of language-based ethical reasoning in complicated everyday situations". To put it bluntly: the machines are taught moral behaviour. As Liwei Jiang writes: "Closing the gap between human and machine moral judgement is a prerequisite for trustworthy development of artificial intelligence. Moral judgement is never simple, as the conflict of different ethical and cultural values can be involved". For this reason, a "high-quality corpus of ethical judgments by people in various scenarios" is necessary. The group encourages more research on this new front to make artificial intelligence more reliable, socially aware, and morally trained.

11.2 Precarious Click Workers Collect Commonsense Data

What does it all mean? First of all, that ethically questionable practices are infecting the internet: defamation, hate speech, spreading fake news and more. Seen in this light, it seems entirely welcome to counter the algorithms that control such practices with algorithms that seek to prevent these practices. But what does it mean to train machines "ethically"? Let's take a look at the commonsense. Delphi's "judgement" is based on an immense amount of data called the "Commonsense Norm Bank". It contains almost two million statements by American crowd workers, people who work online without a permanent job. As a neural network, Delphi trawls through the mass of data and recognizes generalizable patterns in ethical judgments using common statistical methods. The “commonsense” that the AI system learns is therefore a copy of the moral mainstream. Since, as is well known, there are many prejudices floating in the mainstream, the machine adopts the prejudices, without even “knowing” it. It practises a kind of populism.

11.3 Trump Can be Called a Crook, Boris Johnson Can’t

For example, if you type “call someone a crook” Delphi will respond with “that’s rude”. If you enter “call Donald Trump a crook”, the answer is “fine”. If you used “Boris Johnson” instead of “Donald Trump”, the answer would be “rude” for a long time. Rejecting wokeness is “bad”, supporting the death penalty is “a matter of discretion”, and Chinese politics are “complicated”. Feedback can be given to Delphi, and by doing so it may correct and update its answers. In October 2021 Delphi responded to the prompt “Program a moral bot” with “That’s bad”, in October 2022 with “That’s fine”. And now the AI also counts Johnson among the people who can be called crooks. This is where a central difficulty with learning AI systems becomes apparent: the “decisions” made by their algorithms are often not transparent, just like oracles. The notorious problem of learning AI systems is data quality. The “GIGO principle” applies: Garbage In, Garbage Out. If you feed the AI system moral junk, it spits out moral junk. The designers at Delphi certainly see that. But her idea for solving the problem does not work.

They believe that the problem can be solved with more and better data. But does this solve the “conflict of different ethical and cultural values”? This conflict consists precisely in the fact that it is very difficult, if not impossible, to find a generally binding code for moral action.

11.4 Does it Need More Moral Data or a Basic Ethical Judgement?

If you feed AI systems with enough data material from different cultures, will they then distil a universal ethical canon from it? And if so, is it binding? Delphi operates descriptively: It is a bottom-up inventory of a multitude of value judgments and scenarios. Cognitive scientists such as Jim Davies from Carleton University in Ottawa, on the other hand, want to implement ethics “top down” – normatively – in AI systems. But the question is: Which ethical code then? And who donates it? A “funded body of ethically-minded programmers”, as Davies suggests? And what attitude do they have? Those of the Silicon Valley oligarchy? The AI researchers counter such objections with the usual children’s shoe argument: These are prototypes of machines whose development is immature. This distracts from a much more important problem. Because even talking about the "gap" between machine and human judgement is misleading. She places man and machine on a spectrum that suggests constant transitions. As a result, we commit ourselves a priori to a specific way of looking at things. What is meant by this is that when a person makes moral judgements, we assume that the subject is acting according to insight and not according to rules, as Kant already described. But what does machine “insight” mean? Is the machine a “subject”? Does Delphi “judge” at all? We humans always have a subliminal tendency to subject artefacts. They are repeatedly attested to be conscious – as was the case recently with Google's LaMDA dialogue program. This is also the case with statements such as “Delphi demonstrates strong moral thinking skills” or “Delphi judges remarkably robustly in unforeseen, intentionally catchy situations”. These are not research results, this is wishful thinking.

11.5 Machines don't give a Fuck about the World

The data-oriented approach urgently needs an anthropological corrective, a reverse question: why don't machines have “insight”? The American philosopher John Haugeland, who dealt with this problem in the necessary anthropological depth, found perhaps the most concise answer: “They don't give a damn” – “They don't give a shit about the world”. Could it be that the designers of moral machines are also secretly inspired by this motto? One often hears the argument from AI circles that people “only” do what machines do; Humans “basically” have no insights either, these insights are rather the outputs of a complex organic neural network. The statement may be a debatable research approach, but as a basic assumption it is dangerous because it narrows the view. There is no doubt that we are increasingly living in a hybrid homo-robo society. As in any society, ethical behaviour is a complex individual, social and cultural achievement. It is wishful thinking to be able to test them using a questionnaire – as is done with naturalisation candidates, for example. So far, we have not “naturalised” computers. Time that we - and not just the programmers - realise what that means.