HUMAN TECHNOLOGY FOUNDATION

Ethics: Why We Should Focus on Privacy, Surveillance, Manipulation of Behaviour and AI "Intransparency"

Selected as one of the 25 experts of the new Global partnership on AI Working group, Vincent C. Müller is Professor of Ethics of Technology at the University of Eindhoven[1], University Academic Fellow at the University of Leeds[2] and Turing Fellow at the Alan Turing Institute[3]. Specialized on the ethics of disruptive technologies, he has just published the article "Ethics of Artificial Intelligence and Robotics"[4] in the *Stanford Encyclopedia of Philosophy* and is very well positioned to help us navigate the sometime confusing dynamic field of AI ethics and explain the crucial issues we must now address: opacity, the use of data and surveillance. He is now writing a book, "Can Machines Think? Fundamental Questions of Artificial Intelligence" which will be published next year by Oxford University Press, New York.

A lot of work is going on in AI and robot ethics, are we on the right track?
There indeed is a lot of action in the field, but we should be more specific. We all agree that AI should be fair and support human rights … but that is not very helpful. At the same time, there are real issues emerging. Take privacy: it is really hard to protect since our technologies have provided us with a huge surveillance system, and we just haven't really noticed what's going on. It was the same with the invention and adoption of cars on the streets: it is after the fact that we realized the impacts on urban planning, the pollution, the accidents… Technologies produce a dynamic in themselves, with benefits, of course, but also risks which we need to think about at the same time. And then we do something to mitigate those risks. It took decades to have seatbelts in cars and helmets for motorcycles. We are now trying to adjust to these new disruptive technologies.

You mentioned privacy and surveillance technology, is the current use of data one of the most urging questions to tackle?
As I argue in my Stanford Encyclopedia of Philosophy article, the debate on surveillance and the use of data is the most crucial and realistic one. Realistic because AI is already used for collecting the data, analyzing it and manipulate people, mostly for economic purposes. This is the most important and practical problem. (Editor's note: the sections of concern are 2.1 Privacy & Surveillance and 2.2 Manipulation of Behaviour)

In addition to AI use, what about AI opacity?
AI opacity, or "intransparency", is another crucial debate to address (Editor's note: the sections of concern are 2.3 Opacity of AI systems and 2.4 Bias in Decision Systems). AI systems are often not transparent, i.e., we don't know why it does what

[1] https://research.tue.nl/en/organisations/philosophy-ethics/persons/
[2] https://ahc.leeds.ac.uk/philosophy-6
[3] https://www.turing.ac.uk/people/researchers/vincent-c-muller
[4] https://plato.stanford.edu/entries/ethics-ai/

it does. So, it might do it because of the wrong reasons, which is the case with biased AI. Also, we might not be able to know how reliable a decision is, or on which assumptions it is based. It's a feature of the technology that is very hard to get rid of, especially in machine learning.

Generally, technology often generates opacity. This opacity will be different for different people: end users, the engineers who designed it, the operators, etc. Opacity can even be then a design and a political feature. I have a personal experience related to credit card applications. When in Greece, the bank refused my application, I ask a bank employee why and even he did not know: he had just sent the data to the centralized computer, and then received the result: "No." When I settled in the Netherlands, I also applied online to get a credit card, the answer was also "No." I then went to the branch office to talk to an employee and she could explain that the application needed a signature from someone at the branch to be approved, and normally the customer would have to be at the bank for more than six months. Since I could have had this conversation with her and explain my situation, she could check my bank account and agreed to sign for my credit card application … which I eventually got! In that latter case, a human could resolve the absurd situation. This was just an unimportant issue and we now have systems that decide on life or death or "support" political decisions. We have a real problem in our democracy if we make the decisions that are not properly transparent.

But then Machine Learning (ML) has a different sort of opacity: the models can be **opaque even to the experts**, which is worse. Some people then think we should not use ML for safety-critical systems, since we will never be confident it will properly work in new situations.
Ethics is often brought to technology from the outside, but with ML, the experts recognized the problem and brought it up. And then one can ask: "Are we humans able to completely explain the process of our own decision-making?" Certainly not. It illustrates the importance to understand the difference between the *explanation* of a decision (e.g., the Human-Resource AI system recommended hiring this person because of this and that parameters) and the *justification* of a decision (e.g., it's justified to hire a person who has the proper qualifications required, who is motivated to do the job, etc.). To overcome this pressing AI opacity issue, people are working on the "Why" and justification of conclusions and decisions made by AI, a field also called "explainable AI," (which is not a good term).

Does that mean that the other issues are less important?
The questions raised by robots and interactions with robots are far less crucial and monopolize too much of AI ethics effort (Editor's note: the sections of concern are 2.5 Human-Robot Interaction, 2.6 Automation and Employment and 2.7 Autonomous Systems). In my view, people overstate the ethical problems here. People see robots as a dooming threat, imagining we will have sex with robots, wage war to robots, etc. First, I think that it is an overstatement since robots are not as powerful now or in the near future. And second, the issues raised are standard questions we encounter with every piece of technology. For instance, even with an ordinary car the responsibility of an accident is hard to establish between the driver, the car manufacturer, the car designer, the repair shop, the companies that

produced the spare parts… The questions are not fundamentally different with autonomous vehicles.

Then I expose problems that are interesting mostly for philosophers, since AI and robotics ethics provide with material to reflect on what we mean by "ethical," by "agency," by "responsibility." They are referred to as metaethical problems. (Editor's note: the sections of concern are 2.8 Machine Ethics and 2.9 Artificial Moral Agents)

What do AI and robotics tell us about humanity?
They allow us to reflect on how we see ourselves.

A first question is: Are we inherently different from machines? There is the computer metaphor for human beings "I take data in, I process and I produce action." And we now have the machines that do the computation as well, but on different hardware. Therefore, if what humans are doing is computing, then something like "feeling embarrassed for what we do" may be attainable for other computing devices. Then there is this vision of human beings as the sum of a physical body and a "soul," a "spirit" that does not die. A position which is hard to sustain. We tend to think we are different from the rest of the world. What happens if we throw away this view? La Mettrie, a French Enlightenment philosopher, did exactly that in his book "L'Homme machine" published in 1747, arguing that humans are some kinds of mechanisms. Then philosophers must work on a proper definition for mechanisms and computing mechanisms.

Another point is raised by the idea that everything is intertwined and interdependent; thus, there is no entirely separate entities. For instance, my brain cells are entwined in a very complex neural structure, meaning my thinking is the result of the implication of all my brain cells, which are not directly in my control, and is influenced by many things (the amount of sleep, of sugar, etc.). Furthermore, the context shapes the individual. I learnt German because I was raised in Germany. I cannot do anything alone, everything I accomplish is thanks to me AND my environment. It's absurd to think, "I'm this one thing, and I am in control." It thus threatens the very notion of single responsible agent: the fact that a human system involves part of the mental states, physical states and social states means you cannot identify one particular agent that would be responsible for the particular act.
We often make the mistake to laugh about people in the past who made terrible mistakes in judgment (e.g., failed to condemn the enslavement of Africans). But people will laugh at us in the future. In fact, people from other cultures are laughing at us now. Not taking into account how our environment shapes our thinking leads to a misunderstanding. We should therefore be more careful about attributing responsibility to people, and rethink responsibility altogether.

Do you have operational advice for citizens, researchers, public and private leaders who are reading these words?
Everybody who is making or using AI system is a responsible human and is responsible for her or his actions. They should ask themselves, "Is what I am designing going to generate good outcomes? The question can be hard to answer,

and that's OK, but it's important to keep it in mind. It is very easy to lose focus on that question. For instance, you are an AI engineer and your manager briefs you to do an AI that maximizes something, but it can very well ignore privacy at the same time. People who make decisions should realize that they have an impact and weigh the consequence of their actions and inactions. It applies to everybody, and in particular in technology. The higher the responsibility, the greater the impact, so we can say that Trump and Johnson are responsible for a lot of deaths by not wearing masks and not protecting others. —So, we are all responsible for what we do. We must be alert.

Interview by Lauriane Gorce, Scientific director of *Institut de la technologie pour l'humain — Montréal*