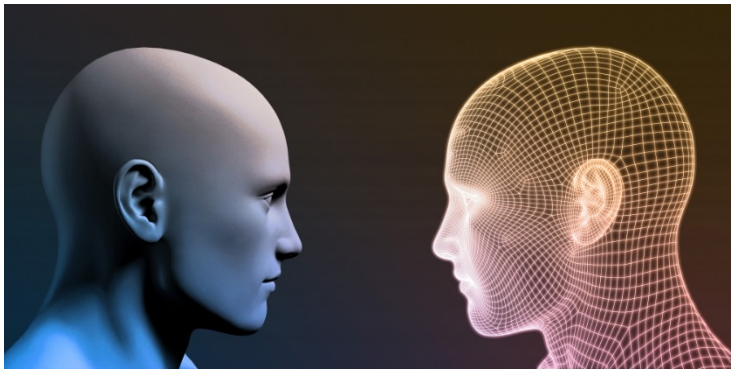
# What if technologies had their own ethical standards?

Technologies are often seen either as objects of ethical scrutiny or as challenging traditional ethical norms. The advent of autonomous machines, deep learning and big data techniques, blockchain applications and 'smart' technological products raises the need to introduce ethical norms into these devices. The very act of building new and emerging technologies has also become the act of creating specific moral systems within which human and artificial agents will interact through transactions with moral implications. But what if technologies introduced and defined their own ethical standards?



© Kentoh / Shutterstock.com.

Self-driving cars might need to make decisions about the life and death of their passengers and other people on the road in the case of unexpected events. Increasingly autonomous personal robots could serve people loyally, but could also inflict harm on them. Blockchain applications, with their society-transforming power, could improve access to finance and banking services for underserved populations, and distribute aid to refugees in a more transparent and efficient manner. All these technologies could have resounding ethical implications for people's lives. Human-computer interaction (HCI), including social networking and gaming, modifies the way humans think and changes human motivation and personality by shifting moral values. These moral influences are summed up by the concept of Homo informaticus.

Equipping machines with ethical principles or procedures, and strengthening the moral potential of machines for resolving ethical dilemmas in order to function in an ethically responsible way, has in fact become a dominant technological trend. Designing technology necessarily involves embedding answers to ethical questions by selecting moral frameworks, such as the consequentialist utilitarian or the Kantian deontological rights-based approach, or by creating specific moral systems within which human and artificial agents will interact through moral transactions and prevent possible harmful behaviour. Selecting among several competing algorithms or programming a computer to be ethical is a complicated process, as ethics operates in a complex domain that requires familiarity with moral principles and the ability to recognise and evaluate a vast array of parameters concerning humans and their environments.

The embedding of ethical values and moral agency at the design phase of technology development raises questions about the moral status of these technologies. Will robots ultimately become morally valuable, beyond their instrumental value as devices merely manufactured to perform specific tasks, and have moral duties and rights? Is it possible to construct some kind of 'artificial moral agents' that could have 'moral knowledge' and apply it in a range of different and possibly complex moral dilemmas? If so: which moral code should they be programmed with? Do Asimov's Three Laws of Robotics constitute an adequate moral point of reference or is there a need to develop 'robot virtue ethics'? Could robots ultimately have not only obligations and duties, but also moral rights?

Such questions have led to the emergence of initiatives such as the IEEE global initiative on ethics of autonomous and intelligence systems, the recommendations of Germany's Ethics Commission on automated and connected driving to specifically programme ethical values into self-driving cars and



Author: Mihalis Kritikos, Scientific Foresight Unit (STOA)

PE 624.247 – September 2018

prioritise the protection of human life above all else, as well as the blockchain ethical design framework. Moreover, as robots inevitably take on more responsibility for human life and safety, there are intensive discussions about whether they should be programmed as 'artificial moral agents' (AMAs) that could actually aid or even replace humans when it comes to difficult moral decision-making in specific and limited contexts. Two classes of AMAs have been described: situation-action machines with rules specifying actions to perform in response to particular stimuli, and choice machines, which possess utility functions over outcomes, and can select actions that maximise expected utility.

Beyond the ethical programming of smart devices, machine-learning algorithms could automate complex ethical decision-making themselves. Their capacity to tweak operational parameters and decision-making rules in the wild, and to define themselves best-fit models to make sense of a set of inputs, mean that machines may be able to write their own morally sophisticated algorithms, much as they might 'reproduce' themselves. Machine learning is expected to come to its own conclusions about ethics and adapt them to unpredictable situations, by programming core principles and relying on iterative learning and experience to guide the results. If artificial intelligence (AI) is empowered to adapt its own ethics, technological codes and algorithms will be expected to become the ultimate curators and gatekeepers in our quest for ethical safeguards, to determine behavioural options and the limits of interaction in virtual spaces and to define the terms of access to information in cyberspace. Today's autonomous-vehicle designers are grappling with a series of ethical quandaries in their design work: How should the many competing values be weighed? Where should the line between machines and humans be drawn with regard to the power of control?

## What do the disruptive effects of technology upon ethics mean for European policy-making?

Despite the short time cycle of technology innovations, especially in foundational technological systems, even the most powerful algorithms being used today 'have not been optimised for any definition of fairness'. Therefore, in view of recent EU policy and legal initiatives in the field of algorithms, the EU legislator may have to consider the need for ethical standards to be built into algorithms that guide eligibility decisions in areas such as credit, insurance, employment and school admissions. The EU may need to consider establishing common user-centric standards or guidelines for embedding human norms and values into autonomous and intelligent systems that are contextual and application-specific. As an algorithm is only as ethical as the data and goals fed into it, EU policy-makers need also to consider seriously whether machine moral agency or autonomy, including the desirability of building AMAs that are full ethical agents that, like ethical human decision-makers, would be capable of making explicit moral judgements, should be 'designed out'.

Moreover, policy-makers should collaborate closely with technologists to investigate, prevent and mitigate potential malicious uses of AI allowing dual-use- and misuse-related considerations to influence research priorities and norms, including the potential of technologies to develop their own moral agency. Given that moral agency comes in degrees, and that it is difficult to draw a line between 'implicit' and 'explicit' ethical agents, EU norms in this area need to remain dynamic and flexible. Any EU recommendation or guideline in the domain of the ethics of programming and automating ethical decision-making means designing the management of scenarios such as trolley and tunnel problems in ways that avoid reinforcing patterns of social inequality.

Getting appropriate data on explicit ethical measures to train AI algorithms appropriately is challenging, because ethical norms cannot always be clearly standardised. The Massachusetts Institute of Technology's Moral Machine project shows how crowd-sourced data can be used to train machines effectively to make better moral decisions in the context of self-driving cars, on the basis of a commonly agreed set of definitions. Extending the use of moral proxy models and using machine learning on human-labelled instances may also prove to be of added value. There is an urgent need for a public debate among affected stakeholders about what a just distribution of the risk burdens and benefits of self-learning systems should look like and how to secure meaningful human oversight and control of AI and alignment with human values, as well as about how to minimise algorithmic opacity, hidden machine bias and automation bias.