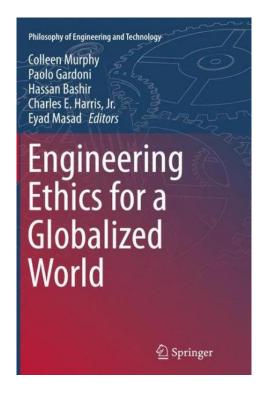
# ETHICAL AI? The Design of Moral Machines

## A GLOBAL SITUATION

For 200 years, engineering has increased its usefulness for, and impact on, societies, industries, businesses, work and jobs. Consequently, engineering ethics has emerged as a field of applied ethics. Engineering ethics covers the obligations of engineers to clients, diverse human societies, the natural environment, and the engineering profession.

Propelled by the Industrial Revolution – which has gone from local to regional first, and to global lately – engineers are now changing the entire world for better or worse. Thus,



*Engineering Ethics for a Globalized World* has become an issue. The rise of a global economy and the approach of Western or Western-trained engineers have created this new ethical frontier. Global engineering ethics must handle cultural, social, religious, curricular and pedagogical challenges. It must provide guidance to the engineering profession in how to work responsibly in the non-Western contexts of developing countries. And it is expected to steer the myriad applications of technology to the millennium goal of poverty reduction and ultimately eradication.

#### A LOCAL DEPARTMENT

Now, if this is the daunting global situation, what can a small interdisciplinary department in the College of Engineering and Applied Sciences, such as our Department of Technology and Society (DTS), do?



#### Actually, a lot. DTS

- combines technical expertise with social science research and humanistic values;
- it can freely say, *engineering has become much too important to be left to the engineers*, act accordingly, and, therefore,
- helps the engineering profession to respond sensibly to the growing onslaught of external demands.

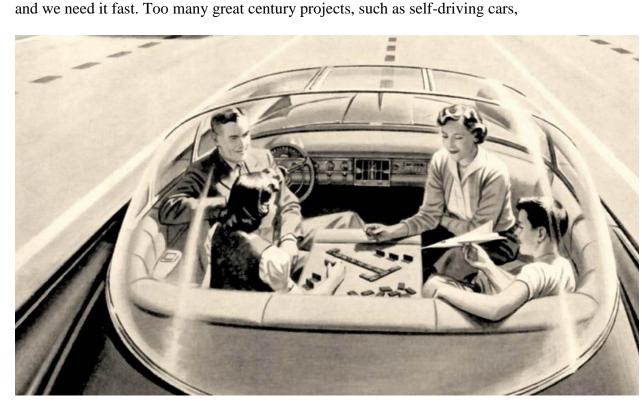
Specifically, DTS contributes to the development of **Smart Engineering Education** in four critical intersections of technology and society: *smart development, smart communities, smart ethics*, and *inclusive STEM education*.

DTS is a resource, because mere technical solutions cannot solve the big challenges of our civilization. For example, the threat of climate change cannot be solved by SRM (solar radiation management) techniques alone, such as stratospheric aerosol injection to block sunlight from reaching earth's atmosphere.

Humanity's biggest problems radiate out into billions of lives, numerous disciplines, and all politics. The urgency to expand our technoscientific approaches toward the societal, ecological, moral and political realms is growing quickly. Why? Because the speed of disruptive technologies is increasing rapidly and the window for responsible innovation is shrinking constantly. Humanity's time to adapt to its own creations is running shorter and shorter. From the domestication of fire to the domestication of plants and animals during the Neolithic Revolution, the domestication of fossil fuels during the Industrial Revolution up to the domestication of nuclear energy in the 1940s, our time to respond has shrunk from hundreds of thousands of years, to ten thousand years, a few hundred years and now decades. The development of holistic engineering design must quickly speed up locally and globally.

# ETHICAL AI – QUESTION MARK

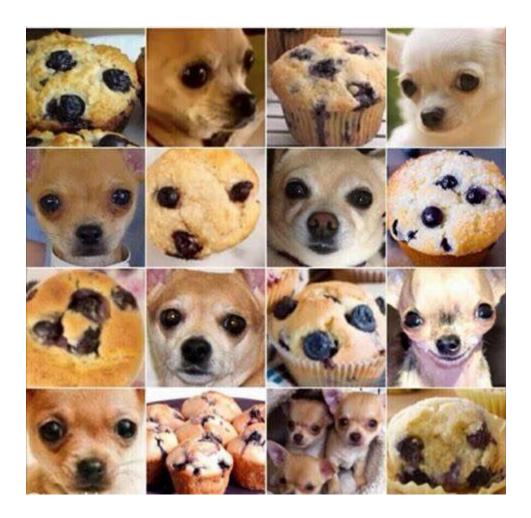
Our topic today is a case in point. We need a *computational ethics* for the twenty-first century,



are racing ahead without guidance and support of a computational ethics. Of course, this new branch of engineering ethics is neither readymade nor easy to grow. "Embedding ethical principles in an AI system" sounds good, but how to do it, is far from clear. Even if "moral machines" can be built, we cannot assume that their decisions will be uncontroversial or at least acceptable like a national Supreme Court ruling.

### HOW SMART IS AI?

To see the current AI situation plainly we must peel away the hype that surrounds AI and ask: *How smart is Artificial Intelligence right now?* And the answer is: *Not as smart as a 3- to 4year-old preschooler*. Does a preschooler see what you see?



Of course, they are not mistaking a puppy for a muffin. Yet Google's image recognition technology has proven less smart and even truly repugnant. It has labelled a black software developer and his friend "gorillas."<sup>1</sup> This mistake and embarrassment happened in 2015. It was quickly fixed, yet not by making the offensive algorithm smarter, but by erasing the "gorilla" tag from Google Photos. Over two years later, Google's primitive stopgap was still blocking the service from recognizing actual gorillas.<sup>2</sup>

AI image recognition will eventually catch up with 3- to 4-year-old children and spot humans, primates, puppies, and muffins fast and accurately. This improvement will certainly come. Yet it is also urgently needed, because misidentifications that are "only" offensive, when based on biased data, can be deadly, when they happen at 60 miles-per-hour in a self-driving car.

### AI – THE BEST OR WORST THING EVER?

Next to the hype stands the chorus of prominent AI Cassandras ranging from Elon Musk



over Prince Charles (AI is "CRAZY") to Stephen Hawking. What shall we make of that? How

safe is Artificial Intelligence? Is AI the best or worst thing ever?

A year ago, in November 2017, at the Web Summit in Lisbon, Portugal,

<sup>&</sup>lt;sup>1</sup> See twitter.com/jackyalcine/status/615329515909156865.

<sup>&</sup>lt;sup>2</sup> See Tom Simonite. "When It Comes to Gorillas, Google Photos Remains Blind." *Wired*, January 11, 2018, wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind.



the late Professor Hawking saw our minds *amplified* by AI and our computers *emulating* human intelligence and exceeding it; he imagined AI reversing the environmental damage done by the Industrial Revolution and eradicating poverty and disease.<sup>3</sup> Yet he also warned AI could "sideline" us, destroy "millions of jobs," and even humanity itself. Hawking said,

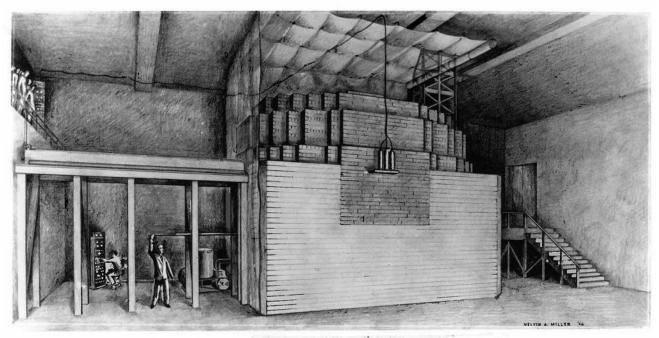
the rise of powerful AI will be either the best, or the worst thing ever to happen to humanity. ... What is the answer? To control AI and to make it work for us, and to eliminate as far as possible its very real dangers...

As a historian, I am not so sure about the best or worst thing "ever," but as a contemporary, I think Hawking is correct on both counts: *One*, AI hits home as our most profound challenge.<sup>4</sup> *Two*, our technocivilization must indeed learn *to control AI, make it work for humanity, and manage its dangers*. The good news is, something similar has been achieved

<sup>&</sup>lt;sup>3</sup> Delivered on November 6, 2017, via video-link. See Stephen Hawking, *Speech at Websummit 2017*. The

UNIVERSE YouTube Channel. Accessed August 31, 2018. <u>https://www.youtube.com/watch?v=vxwD9VYCPZo</u>. <sup>4</sup> Elon Musk, another prominent AI warner, goes further than Hawking. He thinks AI is beyond our control, that "the percentage of intelligence that is not human is increasing," so that eventually humanity "will represent a very small percentage of intelligence"; see Think Club. *ELON MUSK ON THE DANGERS OF ARTIFICIAL INTELLIGENCE*. Accessed September 10, 2018, <u>youtube.com/watch?v=Ufqn4XSH7bo</u>. A transcript (riddled with errors) is available at "#1169 - Elon Musk." JRE Scribe. Accessed September 10, 2018, <u>https://jrescribe.com/transcripts/p1169.html</u>.

once before, in December 1942 in a squash court underneath the football grandstand of the University of Chicago's Stagg Field.



Chicago Pile I (CP-I), World's First Reactor

In this circumscribed space, Enrico Fermi and colleagues demonstrated how to control nuclear fission. The bad news is, the chain reaction AI has triggered is not localized in a discrete place, but an explosion of machine learning all around the world, driven by all sorts of actors in all kinds of areas, governmental, military, industrial, and academic. Considering this unprecedented ubiquity of the AI Revolution, Tech and Society has resolved to research "smart ethics," that is, a universal computational ethics that could theoretically apply to all algorithms.

#### AI FOR GOOD?

We observe growing student interest in what is called "Computing-for-Good" (C4G).<sup>5</sup>

Husband and wife team Toni and Lori Scarlatos from Computer Science and Tech and Society, respectively, have responded to this demand for ethical computing in our college. They have created a cross-listed undergraduate course for Benevolent Computing, ISE 339.

But here's the rub. Voluntary forays into C4G, which excite the students that are drawn to it, are much too little and do not scale up to meet the great civilizational challenge that Hawking and others have diagnosed. Computer science needs *universal* guidance and *mandatory* classes in, and practical tools for, ethical programming. *All* computer science students must learn new how-tos:

• How to build moral machines with algorithmic ethics.

<sup>&</sup>lt;sup>5</sup> See Computing for Good at Carnegie Mellon University, <u>hcii.cmu.edu/courses/special-topics-computing-good-c4g</u>, and at SBU, <u>stonybrook.edu/magazine/2018-winter/computing-for-social-good</u>.

- How to overcome moral Taylorism, which breaks complex projects up into small and seemingly ethically neutral tasks.
- How to decline "sweet" offers of work and money.
- And how to team up with philosophers, sociologists, linguists, and other strangers.

In short, technology-driven societies must build computational ethics across the governmental-

military-industrial-academic complex.

# AI CROSSROADS



Humanity has arrived at the crossroads again. Pure science is taking a backseat in our brave new AI-enabled world. The physical universe, humanity's first all-around nature, is receding, and our second nature is rushing into the foreground. Now we are surrounded by all the new things humanity has created and the Internet has connected, from light bulbs to cell phones. The physicists, who reigned supreme during the grand nuclear challenge in the 1940s, have been

overtaken by AI engineers in terms of pay scale and societal relevance. Technoscience (which we have analyzed as "finalized science"<sup>6</sup>) is leading us into the twenty-first century.

The wisdom of the so-called "deep learning" of neural networks and the discoveries of big data science projects propel our AI machinery toward intelligence and smartness. More and more of our gadgets and machines acquire good-enough knowledge, begin to sense, and act. This progress may be slow and rocky on its own terms, as the example of Google's biased image recognition has shown, but it is blindingly fast on the larger background of global human history.

The demand for a computational ethics has been recognized as a *foundational research challenge* in a white paper for our new *Institute for AI-driven Discovery and Innovation*:

As AI systems continue to expand the range of tasks they can learn to do on their own, there is increasing likelihood that they will confront ethical dilemmas. For example, should the self-driving car run over a deer or steer the car around the deer instead and crash into the embankment on the road, possibly injuring the car's occupants? Thus it will become increasingly important to codify ethical principles in AI systems. Embedding ethical principles in an AI system can also possibly thwart these machines from turning malevolent. Becoming malevolent is not outside the realm of possibility as these machines train themselves through volumes of data, increasing the chances of becoming exposed to malevolent behavior. Building ethical AI systems is a novel and unique challenge. It calls for developing the new area of computational ethics – a framework for codifying ethical reasoning in machines.<sup>7</sup>

Again, our historical experience is promising. Medical students have pledged the Hippocratic Oath to do no harm (*primum non nocere*) in various forms over the centuries. Their example shows that a professional group can rise to the challenge of risky and dangerous knowledge. A similar "golden rule" was proposed by Karl Popper, Joseph Rotblat and others for natural scientists in the second half of the twentieth century, but that oath did not materialize.

<sup>&</sup>lt;sup>6</sup> See Wolf Schäfer, ed. *Finalization in Science: The Social Orientation of Scientific Progress*. Vol. 77, Boston Studies in the Philosophy of Science. Dordrecht, Boston, Lancaster: D. Reidel, 1983.

<sup>&</sup>lt;sup>7</sup> "Stony Brook University Institute for AI-Driven Discovery and Innovation." Unsigned and undated MS.

Chemists, who weaponized gas in the First World War, and physicists, who produced nuclear bombs in the Second, had ample reason to embrace an overarching code of conduct, but they could not bring themselves to limit their "sweet" (Robert Oppenheimer) research. Facing the possibilities of malevolent AI, computer scientists could join the medical profession by lifting engineering onto the track of ethics. Yet they can also follow the other beckoning track that promises untrammeled freedom for unlimited research.

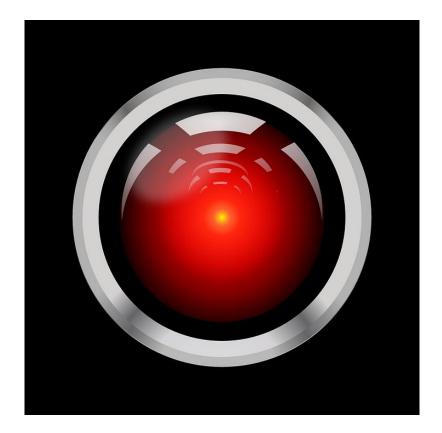
The image behind me is misleading. Our pathways into the future are not laid out like railroad tracks with well-defined stations in the years 2024, 2042, and 2081. Our future looks more like this:



Cloudy with few details and no certainty that green turf, or anything else now visible, will continue on either side.

#### MALEVOLENT AI

Ethical AI – Question Mark can be understood in many ways. For instance, it could be read as a skeptical question, an inquiry that would ask: Isn't unethical AI more likely than ethical AI? Will moral machines take over that exhibit tribal biases and other ethical failures? Can malevolent AI win the race?

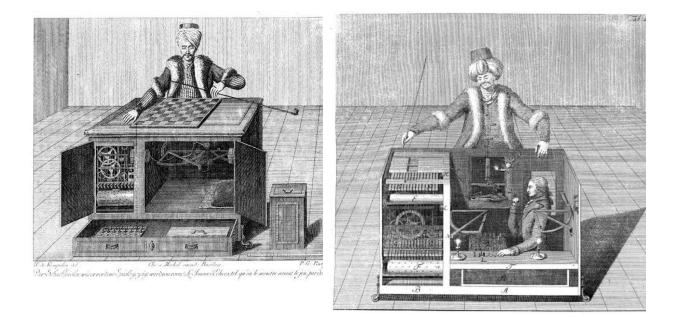


I think we should consider that prospect. The victory of malevolent AI is perhaps not likely, but utterly possible. Humanity could be swamped with ethically questionable AI engines – algorithms that will elevate economic inequalities and exacerbate societal prejudices about races, genders, religions, and other relational issues. Unethical AI could put discrimination on a quasi-scientific level of excellence and malevolent AI could turn against humanity. I don't want to elaborate this scenario, but would like to warn Computer Science against stopgap remedies, such as outsourcing the ethics question to the Philosophy Department. What else than a Reader's Digest course in ethics can come of that? More serious work is needed to equip computer science students with the crucial new how-tos mentioned earlier.

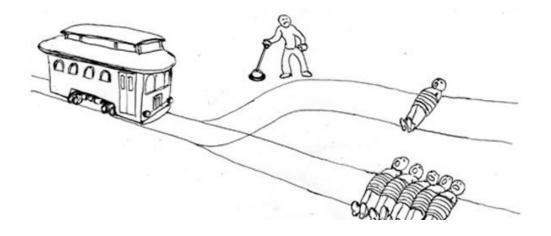
#### **BENEVOLENT AI?**

Ethical AI – Question Mark could be understood positively and answered affirmatively as well. And why not? Ethical AI is entirely possible. Moral machines could be designed that would handle moral problems not only faster, but also better and more consistently than human brains. Autonomous automobiles could be built that will cause far fewer accidents than human-driven cars. Furthermore, the smart ethics of AI-driven vehicles could execute moral reasoning on a much higher and more reasonable level than normal humans normally exhibit.

I don't want to elaborate this scenario either, but would like to remind the skeptics among you about the career of chess-playing engines. They have come a long way since Wolfgang von Kempelen's fake "Mechanical Turk" from 1770:



Chess algorithms had to wait for Alan Turing<sup>8</sup> and Claude Shannon,<sup>9</sup> who developed the first true chess engines independently in the late 1940s. From then on, chess-playing algorithms evolved quickly. IBM's Deep Blue defeated world chess champion Gary Kasparov in 1997. We might expect a similar outcome for ethics engines. However, let's complicate this optimistic perspective with the "Trolley Problem."



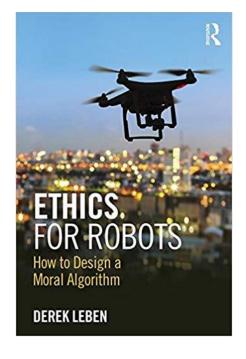
<sup>&</sup>lt;sup>8</sup> See https://docs.google.com/file/d/0B0xb4crOvCgTNmEtRXFBQUIxQWs/edit

<sup>&</sup>lt;sup>9</sup> See https://vision.unipv.it/IA1/aa2009-2010/ProgrammingaComputerforPlayingChess.pdf

Automotive AI engineers are disinclined to engage with the many variations and difficulties of this famous thought experiment.<sup>10</sup> Given the supreme engineering goal of total safety, their reluctance is understandable. Autonomous cars will seek collision-free paths and avoid the danger cones of all potential obstacles; they will not drive into dicey situations, but steer away from them. Eventually, car accidents will disappear and become a thing of the past.

Christoph von Hugo, Mercedes-Benz's manager of driver assistance systems, expressed the total safety vision perfectly. "After sorting through thousands of terabytes of real-world test data and multiples of that in simulations," he said, "answers to the Trolley Problem will become less important as the advantages of ever-alert autonomous cars become apparent."<sup>11</sup>

Nevertheless, enlightened by the arguments of Derek Leben, a philosopher,



<sup>&</sup>lt;sup>10</sup> Posed first by two philosophers, Philippa Foot in 1967 and Judith Thompson in 1976; see "The Problem of Abortion and the Doctrine of the Double Effect" at <u>https://philpapers.org/archive/FOOTPO-2.pdf</u> and "Killing, Letting Die, and the Trolley Problem" at <u>https://laurenralpert.files.wordpress.com/2014/08/thomson-killing-letting-die-and-the-trolley-problem.pdf</u>.

<sup>&</sup>lt;sup>11</sup> Michael Taylor, "Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians." *Car and Driver*, October 7, 2016. <u>https://www.caranddriver.com/news/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians</u>.

who has wrestled with the design of ethics algorithms for autonomous machines, I would still uphold the importance of the Trolley Problem. *First*, because an automotive AI system, which is constantly trying to avoid accidents, is always juggling *better* or *worse* trajectories, hence, it is "implicitly making moral judgments."<sup>12</sup> *Second*, it is implausible that dilemma-style collisions will completely disappear. They could become extremely rare, but hardly shrink to zero. Hence, the venerable practice of bioethics "to consider extreme and unusual situations" (ibid.) should be emulated in the Automotive Ethics Lab.

<sup>&</sup>lt;sup>12</sup> Derek Leben, *Ethics for Robots: How to Design a Moral Algorithm*. Abingdon, Oxon; New York, NY: Routledge, 2018, p. 100.