Ethics for Times of Crisis

by

Jan Nagler,[1,2] Jeroen van den Hoven[3], and Dirk Helbing[1,3,4]

Affiliations:

[1] Computational Social Science, Department of Humanities, Social and Political Sciences, ETH Zurich, Clausiusstrasse 50, CH 8092 Zurich, Switzerland
[2] Computational Physics for Engineering Materials, IfB, ETH Zurich, Wolfgang-Pauli-Strasse 27, CH 8093 Zurich, Switzerland
[3] TU Delft, The Netherlands
[4] Complexity Science Hub, Vienna, Austria

E-mail addresses: jnagler@ethz.ch; m.j.vandenhoven@tudelft.nl; dhelbing@ethz.ch

What will happen in a crisis when Artificial Intelligence systems will decide about increasingly many issues, including life and death? Will a Citizen Score based on Big Data determine our chances of survival? How should autonomous systems make decisions that are ethically aligned with what is morally required from humans? We argue that in times of permanent crisis the dominant approach should be innovation instead of optimization.

These days, everyone is talking about artificial intelligence (AI), robots and self-driving cars. We absolutely agree that these technologies have promising applications and perspectives. So, why then are people like Elon Musk warning us that AI may pose the biggest existential threat to humanity? (1) And what does this have to do with mundane things such as autonomous cars, if anything at all?

Autonomous cars will certainly be involved in accidents where people may die. Then the question is (2-4) when fatalities cannot be prevented – should an autonomous vehicle be programmed to run over a crowd of people on the street, swerve into a smaller group of pedestrians on the walkway or sacrifice the lives of the car passengers by ramming into a concrete wall? Should particular kinds of people be privileged, giving them higher chances of survival? For example, should luxury cars be allowed to offer a higher degree of self-protection as compared to cars in the lower price segment, to create incentives to buy a more expensive car? Would this be ethically justified, given that expensive cars cause more accidents and already impose higher risks on others (5)?

These are the kind of ethical dilemmas that are now frequently discussed. But there is a far bigger problem nobody is openly talking about: In the non-sustainable world we are

living in, when there are not enough resources left for everyone, will autonomous systems be used to decide about life and death? If yes, how should they decide? Therefore, when we talk about ethical principles governing how autonomous vehicles deal with matters of life and death, one should always keep in mind the implications for scaled up and generic applications in times of crises. Or to put it in a Kantian form: what if the maxims or policies of these types of machines were to be become the universal principles for all machines?

AI systems knowing "who is who"

With better sensor and video technologies and powerful information systems, artificial intelligence (AI) systems are increasingly capable of distinguishing between one person and many, a child and an elderly person, an average person and a famous politician, a white person and a person of colour, a person with a job or without, a rich and a poor person, a convicted criminal and a saint, a healthy person and one who may die soon, a person with health or life insurance and without?

Should people with higher status or life expectancy be protected, because they may contribute more to society? Should a Citizen Score decide, which represents the value of a person from the point of view of the government, as it is currently being tested in China in other areas of life? (6) Should a person who pays a higher insurance premium have a higher chance of survival and others be sacrificed? This sounds like a profitable business model, but it would fundamentally contradict the principle of equality and human dignity, on which the United Nations' Universal Declaration of Human Rights is built.

Criteria such as health, age, or social status are not suitable criteria to decide who should come to harm, live or die, not even from a narrow utilitarian perspective. Recall that Kant, the father of Enlightenment, who inspired modern democratic constitutions, wrote his masterpieces at old age. Van Gogh had a very low social status during his lifetime. Mozart died poor. Beethoven was almost deaf when he wrote his 9th symphony. Degas and Toulouse-Lautrec were handicapped and Monet had impaired sight, but they became three of the most important painters of Impressionism. These individuals have created some of the greatest cultural achievements in the history of humankind. Whatever measure is taken to distinguish the value of people, there are always examples that show the inappropriateness of such a measure.

Utilitarian thinking can be inappropriate

The current rulings of many constitutional courts and ethical committees largely agree that people should *not* be valued differently, but share a common humanity and human dignity. This is also a lesson learned from the history of fascism and the Holocaust. Utilitarian "optimization" appears to be highly immoral, if it intentionally exposes different kinds of people to different life-threatening risks. For example, the new Hippocratic oath, https://www.wma.net/policies-post/wma-declaration-of-geneva/ requires doctors to swear: "I will not permit considerations of age, disease or disability,

creed, ethnic origin, gender, nationality, political affiliation, race, sexual orientation, social standing or any other factor to intervene between my duty and my patient." Perhaps, given that "code is law", we should require a similar oath from computer scientists and software engineers to ensure *design for values* (7).

Also note that autonomous systems based on utilitarian principles would be easily exploitable – both by criminals and authorities. Deterministic decisions could be successfully instrumentalized to harm people in cases of manipulation or hacking. For example, someone may jump on the street and force a deterministically deciding vehicle to crash into a concrete wall, or someone may trigger or hack the vehicle sensors to trick it into a dangerous manoeuvre that might put passengers at risk. Other drivers may anticipate the safety behaviour of the type of car you are driving and exploit its response repertoire for personal advantage. In such cases, a probabilistic decision rule would make it less likely that an autonomous system could be successfully instrumentalized against other people.

Overall, if ethical dilemmas cannot be avoided, decisions should be randomized, giving each person the same weight. A society with ubiquitous AI requires a framework that is impartial, as proposed by the Harvard political philosopher John Rawls with his concept of "the veil of ignorance" (8). This implies that, in deciding about the basic normative principles of a society, one should ignore properties that could be tailored to serve self-interests. This, again, suggests that humans should not be treated differently in a critical situation and solutions based on utilitarian grounds should be rejected.

Killing algorithms are not science-fiction

Very soon, the ethics of autonomous systems may affect all of our lives every day. In turbulent times, as we may encounter them in an unsustainable world, decisions about life and death may become commonplace. Today, robocops are being tested, drones are being used to kill dissidents, and a number of autonomous weapons are in the making (9, 10). Some experts even think about AI-based euthanasia (11) and the use of palliative means. Soon, computer-controlled implants may be used to release drugs to our bodies, but such devices would be vulnerable to hacking and may cause overdoses (12, 13).

Let us assume or a moment, one would apply the Citizen Score, as it is currently tested in China, or the United Kingdom's KARMA POLICE program (14) to make decisions about life and death – saving those who have a higher score. Such a "digital judgment day" approach would create one of the most serious moral hazards imaginable. "The elite", i.e. the people with the highest scores, would always have the lowest risks and the greatest opportunities. Therefore, why should they make a serious effort to improve the opportunities and risks of all the others, if it will not improve their own lives? In contrast, an unbiased probabilistic decision rule in combination with a fair veil of ignorance, would put everyone at the same level of risk, and hence everybody would have an incentive to reduce the number of ethical dilemmas as much as possible.

Humanity has a moral obligation to prevent the occurrence of ethical dilemmas, i.e. choice situations where one cannot fulfil all moral norms at the same time. Furthermore, if ethical dilemmas occur nevertheless, there is an obligation to transform them into situations which expand the set of obligations one can satisfy, whenever possible. A Citizen-Score-based system would certainly miss this goal. It provides a framework suggesting we can fulfil our moral duties by optimization that weighs and counts lives and deaths, and that such a decision can be automated and dealt with by a machine.

To minimize the number of critical situations, we do not only need the best use of human and artificial intelligence, but creativity as well. In a crisis, innovation may be more important than optimization. To successfully address the sustainability challenges of our planet, we may have to fundamentally change the monetary, financial, and economic system, or even the organization of society altogether (15,16). Given the limitations of optimization, Citizen-Scores-based systems and utility maximization, we should spend much more resources on systemic innovation, e.g. on participatory resilience and City Olympics, where cities all over the world and the regions around them would regularly compete for the best environmental-friendly, energy-efficient, resource-saving, crisis-proof, and peace-promoting solutions. https://www.theglobalist.com/technology-big-data-artificial-intelligence-future-peace-rooms/ Such approaches could dramatically improve the future prospects of humanity within a short period of time.

References

1. S. Gibbs, The Guardian, Monday 27 October 2014, available at https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat

2. B. Deng, *Nature* 523, 24-26 (2015).

3. J.-F. Bonnefon et al., *Science* 352, 1573-1576 (2016).

4. D. Leben, *Ethics Inf. Technol.,* 19:107-115 (2017).

5. *The Telegraph*, Nov 18, 2015 (available at http://www.telegraph.co.uk/finance/personalfinance/insurance/motorinsurance/11993627/Its-official-drivers-of-luxury-cars-cause-more-accidents-insurers-say.html).

6. D. Storm, ACLU: Orwellian Citizen Score, China's credit score system, is a warning for Americans. Computerworld (7 October 2015); available at http://go.nature.com/3pq8b4, and at http://www.independent.co.uk/news/world/asia/china-surveillance-big-data-score-censorship-a7375221.html

7. J. van den Hoven, P. E. Vermaas, I. van de Poel, Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains, Springer (2015).

8. K. B. Rasmussen, *Philos. Stud.* 159:205-218 (2012).

9. S. Russell, S. Hauert, R. Altman, and M. Veloso, Nature 521, 415-418 (2015).

10. IJCAI conference, July 28 (2015). Open letter initiative; available at https://futureoflife.org/open-letter-autonomous-weapons/

11. F. Hamburg, Een computermodel voor het ondersteunen van euthanasiebeslissingen (E.M. Meijers Reeks)

12. Hackers remotely kill jeep highway, Wired, July 24, 2015; available at https://www.wired.com/2015/07/hackers-remotely- kill-jeep-highway/

13. Hackers reveal nasty new car attacks, Forbes, July 24, 2013; available at https://www.forbes.com/sites/andygreenberg/2013/ 07/24/hackers-reveal-nasty-new-car-attacks-with-me-behind-the-wheel-video/#45c73d7b228c.

14. https://www.theverge.com/2015/9/25/9397119/gchq-karma-police-web-surveillance

http://www.dailymail.co.uk/news/article-3249568/GCHQ-spooks-spied-internet-user-operation-called-Karma-Police-according-leaked-documents.html

15. D. Helbing, *Nature* 497, 51-59 (2013).

16. D. Helbing and E. Pournaras, *Nature* 527, 33-34 (2015)