An Extension of Asimov's Robotics Laws

by

Jan Nagler,[1,2,3] Jeroen van den Hoven[4], and Dirk Helbing[1,4,5]

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] Risk Center, ETH Zurich, Scheuchzerstrasse 7, 8092 Zurich, Switzerland
[4] TU Delft, The Netherlands
[5] Complexity Science Hub, Vienna, Austria

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

In a world, where Artificial Intelligence systems will decide about increasingly many issues, including life and death, how should autonomous systems faced with ethical dilemmas decide, and what is required from humans?

In the near future, autonomous vehicles are expected to substantially improve traffic flow and drastically reduce accidents (1). According to the Department of Energy (DOE), self-driving cars could reduce energy consumption in transportation by as much as 90 percent (2). A reduction of accidents by 90 percent seems possible as well, which would translate to saving millions of lives every year world-wide (3). However, many pressing questions remain regarding how to engineer autonomous cars and, generally, design artificially intelligent systems for safety and other moral values (4). What public policies and what regulations would be needed?

Should we design autonomous decision-making solely on the basis of moral values and moral principles, or should self-interests or company policies and market forces predominate? How can one design for moral values and develop an ethically aligned design of autonomous systems (5)? A number of proposals for codes and principles have been made (6). We suggest that rather than a grocery list of values we need some clarity in the form of a limited set of basic moral principles dealing with autonomous vehicles.

Electronic copy available at: https://ssrn.com/abstract=3110582

Asimov’s Laws of Robotics

Early on, the well-known science-fiction writer Isaac Asimov proposed “Three Laws of Robotics”, which he augmented by a fourth law later on. These are: *(i) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (ii) A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law. (iii) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. (iv) A robot may not harm humanity, or, by inaction, allow humanity to come to harm.*

These laws, however, may induce ethical dilemmas in certain critical situations. One of these, the ”trolley problem“, received a lot of attention in the autonomous driving literature, recently (7). Imagine a runaway railway trolley that is about to kill five people working on the trolley’s track. Furthermore, assume that you (or the robot) can save them only if a lever is pulled that diverts the trolley onto another track, where it will kill an innocent bystander with certainty. What should you do?

Analogously, if an autonomous car faces a situation where fatalities are inevitable, should it run over a crowd of people on the street, swerve into a smaller group of pedestrians on the walkway or sacrifice the lives of all car passengers by ramming into a concrete wall? Recently Bonnefon *et al*. studied the preference of test subjects for a number of similar trolley problems in autonomous driving (7). They found that even though participants approve of autonomous vehicles that might sacrifice passengers to save others, respondents would prefer not to ride in such vehicles. Respondents would also not approve regulations mandating self-sacrifice. Such regulations would make them less willing to buy or use an autonomous vehicle. Bonnefon *et al*. concluded that regulating for utilitarian algorithms (that minimize total harm) may paradoxically lead to more casualties by postponing the adoption of safer (autonomous vehicle) technology.

This may sound like an invitation to leave it to car manufacturers to determine the properties of their autonomous vehicles based on market criteria. However, this might result in different safety features. Luxury cars could offer a higher degree of self-protection as compared to cars in the lower price segment, in order to create incentives to buy a more expensive car. However, luxury cars cause more accidents, i.e. they already impose higher risks on others (8). Increasing the relative risk imposed on others by price-sensitive safety technology would be highly questionable from an ethical point of view.

So neither a laissez faire market solution, nor a government’s advertising utilitarian policy would work. Which policy is fair, given the fact that safety dilemmas of the trolley type are likely to occur, even if infrequent? Responsibility demands to prevent the occurrence of ethical dilemma situations in the first place, and if this is impossible, to take measures to reduce them as much as possible (9). This suggests the introduction of a fifth law of robotics as follows: *(v) Humanity and robots must do everything possible to reduce the occurrence of ethical dilemma situations*. Sometimes, the fifth law may only require better safety features of autonomous vehicles or stricter speed limits. Generally, however,

Electronic copy available at: https://ssrn.com/abstract=3110582

a re-design of systems – encompassing technology, infrastructure and organization – may have to be considered.

Yet, ethical dilemma situations may sometimes still occur. How should an autonomous vehicle or, more generally, an Artificial Intelligence (AI) system then decide? With better sensor and video technologies and powerful information systems, can we now take better decisions from the perspective of society? Such systems could identify objects and even individual people. Therefore, AI could distinguish between one person or 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? So, who should die, if an algorithm had to decide? Should people with higher status or life expectancy be protected, because they may contribute more to society?

Some may find it plausible and acceptable to value people and weigh their lives differently. In fact, not all organisations, leaders, policymakers, stakeholders and nations agree on the same fundamental moral principles to build a society on, or to prevent moral hazards and exploitation (10). In particular, age and health are sometimes used as factors to determine medical treatments and health risks. Explicit discussions have come from the EPA (U.S. Environmental Protection Agency) in its 2003 discussion of hazardous air pollutants: "*There is general agreement that the value to an individual of a reduction in mortality risk can vary based on several factors, including the age of the individual, (...) and the health status of the individual*.“ (11).

Recall, however, 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. This invariably shows that, even from a broadly utilitarian point of view, health, age, or social status are not suitable criteria to decide who should come to harm.

Consider also that each of us will get old or may fall ill, and that in 99.9% of cases somebody will be around with higher status than yourself. Should rich people be able to buy themselves a higher chance of survival? Should the person who pays more be saved and others sacrificed? This sounds like a profitable business model, but it would substantially harm our society, which – according to the United Nations' Universal Declaration of Human Rights – is built on equality.

In fact, the current rulings of many constitutional courts and ethical committees largely agree that people should *not* be valued differently, considering, for example, status, age, or health, but share a common humanity and human dignity. This is also a lesson learned from fascism and the Holocaust.

Today, robocops are being tested, drones are being used to kill dissidents, and a number of autonomous weapons are in the making (12,13). Even worse, some experts think about AI-based euthanasia (14). This includes computer-controlled implants that release drugs (or an overdose) to our bodies. Such devices could also be hacked (15,16). Cybersecurity is an arms race, not a problem that can be solved once and forever. Thus, the ethics of autonomous systems may soon affect all of our lives every day.

Furthermore, people with higher education may have a better chance to live long in prospering modern societies. In turbulent times, however, as we will encounter them in an unsustainable world that is increasingly faced with problems such as mass migration, war and terrorism, other skills may be important to survive. Therefore, whatever measure is taken to distinguish the value of people, there are always examples that show the inappropriateness of such a measure, also on purely utilitarian grounds. In particular, it is not well justified to attribute a different value to different individuals, such as the Citizen Score concept does (17).

Therefore, we formulate a sixth law of robotics: *(vi) If the application of rules (i)-(iv) leads to ethical dilemmas, which could not be avoided as required by rule (v), decisions should be randomized, giving each person the same weight*. The sixth law is compatible with the equality principle of many constitutions, the idea of human dignity and also with medical ethical imperatives. Namely, bioethics requires that a doctor's decision which patient to treat, in case two wounded appear equally threatened, should be random as seen by an observer – not based on sex, skin colour, wealth, or other features. A society with ubiquitous AI requires a social contract that is impartial, as proposed by the Harvard political philosopher John Rawls with his concept of "the veil of ignorance" (18). This implies that, in deciding about the normative principles of a society, one should ignore properties that serve self-interest.

In the following, we will substantiate the sixth robotic law further. Autonomous driving is challenged by ambiguity, uncertainty and complexity arising from incomplete information, imperfect classification, or just from unclear sensory input such as poor sight (4). The number of pedestrians on a road might be uncertain, the number of casualties may not be accurately estimated, or people may rather be injured than killed (see Figure 1). However, it appears unrealistic to specify appropriate deterministic decision rules for all probabilistic trolley problems, or just the majority of them.

Assume two cars that are about to collide on a bridge with deadly consequences for both. If one car decided to swerve and fall off the bridge, this would save the passengers of the other car. This, however, would not happen if both cars were in a "minimum self-harm" mode (in which passenger safety is the ultimate priority). Then, both cars would frontally crash into each other and kill all passengers. This dilemma is known in game theory as "chicken game", where paradoxically a "maximum harm" outcome is the expected to occur. A probabilistic decision, in contrast, would guarantee a fair chance of survival.

Next, let us discuss the example of manipulation or hacking. For example, someone may jump on the street and force the 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' lives at risk. In such cases, a probabilistic decision rule would make it less likely that an autonomous system could be successfully instrumentalized to harm people.

Finally, let us assume that a Citizen Score, attributing a certain value to everyone's life, would be used to determine who has to die and who will survive, when resources are scarce. Then, this would create a serious moral hazard. "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 the greatest effort possible to improve the opportunities and risks of all the others, if such an effort will not improve their own situation?

A Citizen-Score-based system would, hence, reduce the chance that the fifth law will be taken seriously and sufficient efforts will be made to avoid ethical dilemmas. A fair probabilistic decision rule, in contrast, would put everyone at the same level of risk, and hence everybody would have an incentive to reduce the number of ethical dilemmatic situations as much as possible.

To conclude, the fifth law demands artificial and human intelligence as well as creativity to avoid ethical dilemmas and critical situations. A successful implementation of the fifth law may require us to change the monetary, financial, and economic system, or even the organization of society altogether (10,19). Today's algorithms with the objective to minimize harm are deterministic (20) and, thus, may lead to unnecessary harm (as the above bridge problem has illustrated). Our extension of Asimov's Robotics Laws considers the limitations of optimization, Citizen Scores and utility maximization, and requires to spend more resources on systemic innovation. It is time to think about this.

References

1. B. van Arem, C. J. van Driel, R. Visser, *IEEE Trans. Intell. Transp. Syst.* 7, 429-436 (2006).

2. K. Bullis. *MIT Technology Review* (October 2012).

3. P. Gao, R. Hensley, A. Zielke, A roadmap to the future for the auto industry, McKinsey Quarterly (October 2014); available at www.mckinsey.com/industries/automotive-and-assembly/our-insights/a-road-map-to-the-future-for-the-auto-industry.

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

5. IEEE: Ethically Aligned Design Initiative (available at http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html).

6. Asilomar AI Principles (available at https://futureoflife.org/ai-principles/).

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

8. *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).

9. Van den Hoven et al., *Science and Engineering Ethics,* 18(1): 143-155 (2012)

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

11. E. Posner, and C. R. Sunstein, Univ. Chicago Law Review 72, 537-598 (2005).

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

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

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

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

16. 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.

17. 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

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

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

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

Acknowledgments
We acknowledge support from SNF (grant The Anatomy of systemic financial risk, No. 162776) and ERC Advanced Investigator Grant MOMENTUM (Grant No. 324247).

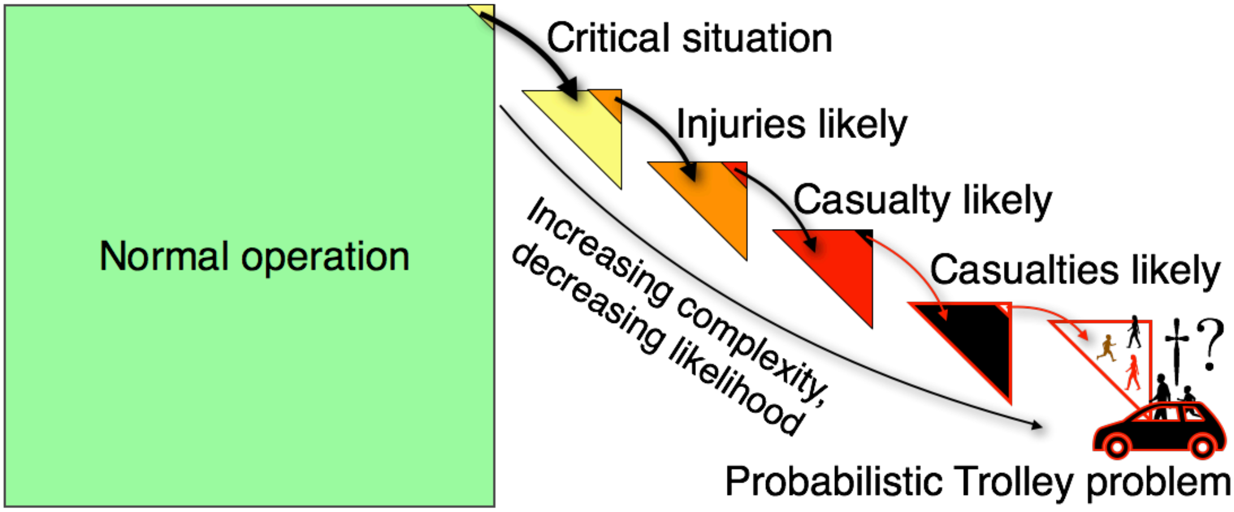

Figure 1: Complexity of the probabilistic Trolley problem