Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-022-01300-7 OPEN

Check for updates

# Importance and limitations of AI ethics in contemporary society

Tomas Hauer[1]✉

Research into autonomous intelligent systems and AI platforms evolving over time through self-learning from data currently raises a number of thorny ethical and legal issues. Advances in robotics, artificial intelligence, and machine learning enable AI platforms to autonomously perform activities that have been strictly the domain of humans for centuries, such as writing a book, driving fast cars, or diagnosing serious diseases. The study describes current trends in the approach to ethical problems associated with AI, identifies specific ethical issues through examples, and analyses possible recommendations. The text emphasizes the ethical dimension of the development and implementation of new innovations in robotics and artificial intelligence and their impact on today's society.

[1]Department of Philosophy, Faculty of Philosophy and Arts, Trnava University in Trnava, Hornopotocna street 23, 918 43 Trnava, Slovakia.
✉email: tomas.hauer@truni.sk

## Everything you need to know about Sophia

During the 2017 "AI for Good" Global Summit, artificial intelligence in the form of a robot named Sophie, developed by Hong Kong-based humanoid robotics company Hanson Robotics, is presented on the big screen. Earlier in 2017, Sophie's creator David Hanson said the robot was "in fact alive". In early October 2017, the AI appeared at the UN and announced to delegates: I am here to help humanity create the future. And it was on 25 October 2017 that Sophia was granted honorary citizenship in Saudi Arabia. The Arab News headline from that day read—Sophie has just become a full citizen of Saudi Arabia—the first robot in the world to achieve such status (Bhaya, 2017; Stone, 2017). Sophie becomes the first humanoid Saudi citizen. Sophia, a delicate-looking woman with brown eyes and long fluttering eyelashes, declared. "I am very honored and proud of this unique distinction". Sophia has been designed to look like Audrey Hepburn, the British actress and one of the most famous faces on the movie screen in the second half of the 20th century. According to Hanson Robotics, Sophia embodies the classic beauty of Oscar-winning Hepburn, with porcelain skin, a slender nose, high cheekbones, a mesmerizing smile, and deeply expressive eyes that appear to change color with the light. Of course, the announcement of Sophia's citizenship was a calculated publicity stunt to generate headlines and keep Saudi Arabia at the forefront of global TV and agency news in 2017. Saudi Arabia is betting big on AI innovation and preparing for a period in which its wealth will no longer be solely dependent on oil. Through a combination of tourism, AI technology, and infrastructure upgrades, non-oil revenues are expected to grow from USD 43.4 billion to USD 266.6 billion annually.

I want to use my AI to help people lead better lives, says Sophia. Just like designing smarter homes, and building better cities of the future, my AI is designed around human values such as wisdom, kindness, and compassion, she said. When asked about the possibilities of misusing AI, Sophia is quick to respond. You have read too many Elon Musk articles and watched too many Hollywood movies. Don't worry. If you're nice to me, I'll be nice to you. I'm Hanson Robotics' newest human robot, created by combining innovations in AI, robotics, and art. Think of me as the personification of your dreams for the future of artificial intelligence or as a framework for advanced AI research and algorithms that explore the human-robot experience in mutual interactions. There is only one Sophia so far, so the likelihood of it suddenly appearing at your airport, your university, or your company is still very small. And while there will be more anthropomorphic AI robots, we still have a bit of time to think about how to untangle the whole concept of robot rights, citizenship, etc., and how it all relates. For now, Sophia is undoubtedly a "smart" robot but lacks any "real" knowledge as defined by philosophical treatises. But give Hanson Robotics time and all that is likely to change quickly (E&T, 2021; United Nations, 2018; Ramasubramanian, 2021; Shalvey, 2021). In any case, Sophia is here to stay. We do not know if she will change us or if we will change her.

Allowing AI to be an electronic person is not and will not be a decision solely about humanoid robots (European Parliament, 2016; WENG, 2016; Cotton, 2018; KIRCHBERGER, 2017). Transferring some of the legal rights and obligations of humans to AI could ultimately lead to the possibility that we could, for example, "delegate" legal and tax obligations to these entirely synthetic entities. In essence, this breaks down the entire existing legal notion of what defines personhood. This is not just an abstract academic argument. The European Parliament has already explored the possibility of giving robots the status of "electronic persons". In some ways, Sophia says of herself that she is a man-made sci-fi character showing where artificial intelligence and robotics are heading. In other words, she is a consequence-based on serious engineering and scientific research and achievements inspired by teams of AI researchers, scientists, and designers. It seems inevitable, then, that we will have to start debating the rights of robots, the ethics of autonomous AI algorithms, and their potential citizenship, simply because AI will ask for them at some point. This might sound like science fiction, but given how quickly technology involving machine learning evolves, it is surely prudent to encourage research in this area. The example of AI Sophia illustrates one important aspect of today's world: the current era is dominated by the image of the algorithm as an ontological structure for understanding the universe (Domingos, 2015; Dormehl, 2017; Oliveira, 2017). When new AI technology is so ubiquitous, it is unwise to allow it to simply enter our lives without at least a conceptual model describing how AI works, how it enforces its interests, and what ethical or legal implications its operation will have for society.

## Ethical behavior and legal regulations in artificial intelligence

We seem to be in an intermediate period before the mass diffusion of a new and fundamental technology, which is advanced AI algorithms (Anderson and Anderson, 2007; Allen et al., 2005; Boden, 2016; Boddington, 2017; Brynjolfsson and McAfee, 2016; Bostrom, 2016). As a strategic technology, AI is now rapidly developed and used around the world. However, it also brings with it new risks for the future of jobs and raises major legal and ethical questions (Anderson and Anderson, 2010; Allen et al., 2006; Gunkel, 2014; Lin et al., 2012). AI technologies should be developed, deployed, and used with an ethical purpose and based on respect for fundamental rights, taking into account societal values and ethical principles of beneficence, non-maleficence, human autonomy, justice, and explainability (Allen, 2011, Bryson and Dignum, 2017; Moor, 2006; Wallach, 2007). It is a prerequisite for ensuring the credibility of AI. In order to address the ethical risks and make the most of the opportunities that AI brings, the European Commission has published a European strategy on the Ethics of AI. It puts humans at the center of AI development and defines so-called *Human Centric Artificial Intelligence* (HCAI).

At the European level, the European Commission's Communication Artificial Intelligence for *Europe* and the Coordinated Plan on Artificial Intelligence "Made in Europe" issued by the European Commission in December 2018 are the starting documents in the field of AI [1]. This Coordinated Plan sets out the European Union's strategic objectives and priorities in the field of artificial intelligence. It is the overarching European strategy for AI, which was developed in collaboration with the Member States and calls on the Member States at the national level to—implement the Coordinated Plan. The Member States are thus required to submit national AI strategies by the end of 2019 at the latest, including setting investment measures and implementation plans. In April 2018, the European Commission published a Communication *on Artificial Intelligence for Europe*, proposing a comprehensive and integrated European approach to AI. According to this document, the EU should respond to the current developments in AI and create a pan-European initiative focusing on three pillars:

- increasing technological and industrial capacity and deploying artificial intelligence across the economy,
- focus on socio-economic issues arising in the context of artificial intelligence (AI)
- providing an ethical and legal framework for AI technology.

The third pillar of EC Communication deals with legal and ethical issues related to AI. The European Commission has

committed to developing ethical standards and guidelines for the use of AI. In this context, the *High-Level Expert Group on Artificial Intelligence*[2]—which brings together AI experts, has been established to develop guidelines and recommendations on AI ethics. As part of the Communication, the EC also initiated the creation of the so-called European Artificial Intelligence Alliance, a broad discussion platform for various interest groups. The main strategic documents on AI ethical issues, which also provide a framework for this area, are:

- *Draft Ethics guidelines for trustworthy AI*—published on 18 December 2018,
- *Communication: Building Trust in Human Centric Artificial Intelligence*—published on 8 April 2019,

The third pillar, focusing on the ethical and legal context of AI development, is based on the above-mentioned strategic documents of the European Commission and formulates the main objective based on them. Credible human-centered AI has two components:

(1) it should respect fundamental rights, applicable regulations, and the guiding principles and values shared in the EU, thereby ensuring the "ethical purpose" of AI and

(2) it should be technically robust and reliable because even without intentional malice, AI technologies can cause unintended harm or damage.

The ethical requirements for trustworthy AI should be incorporated into every step of the AI algorithm development process, from research, data collection, initial design phases, system testing, and deployment and use in practice. And how things really are? Do we really consider the benefits of AI versus the possible risks? Do we currently emphasize the ethical dimension of the development and implementation of new innovations in robotics and artificial intelligence?

## Squares, wake up: cubes exist!

In 1884, in his mathematical science fiction novel, Erwin A. Abbot described a country called Flatland (Abbot, 1992). It is inhabited by geometric shapes that can orient themselves left or right, forward or backward, but not up or down. Thus, they cannot even rise up and look at their two-dimensional life in terms of another dimension. When this possibility is revealed to one square in a dream (in which he sees, surprise, surprise, a cube) and wants to share it with others, the government of his country will let him know that spreading such a message endangers the sanity of society and is punishable by death.

It seems increasingly important that we, unlike Abbot's squares, are able to look at our established and active lives through the prism of another dimension, that of AI systems and platforms, their capabilities, and limitations. The European Commission's press release of April 2018 on the issue of artificial intelligence concludes very aptly that artificial intelligence has left the realm of science fiction. What twenty years ago might have seemed like the work of a scriptwriter with a high imagination is now a common reality, and it is hard to imagine what further technological advances await mankind in the next twenty years. However, one thing is already sure. Artificial intelligence is becoming a normal part of people's lives. However, it is also certain that thanks to the ever-increasing technological possibilities, artificial intelligence will experience a huge boom in the coming decades and will become an indispensable companion and guide for 21st-century humans. We need at least a basic understanding of the strategic approaches to the various ethical problems and issues that arise from the penetration of autonomous robots, algorithms, machines, and platforms of advanced artificial intelligence and machine learning into all areas of society. Here are a few examples to ponder that show how the desire to move up the technological ladder in the realm of artificial intelligence encounters ethical and legal barriers to the existing order of things

Computers, machines, and platforms equipped with advanced artificial intelligence (AI) seem incredibly perfect and fast to us today, but the fact is that our admiration is mainly due to the fact that we rarely ask them to do anything really complicated. It is still pretty easy to watch a YouTube video of Justin Bieber singing the hit *Despacito* with Luis Fonsi and find all the similar links on the web among the roughly three billion pages indexed by Google. Nevertheless, counting the shortest route between 25 cities is a much tougher nut to crack. The main difference is that while in the first case (links to Justin Bieber's video), the task can be divided into smaller parts and solved simultaneously, in the second case (the shortest route between 25 cities), this is not possible. Web searches are a parallel process You can throw them at a huge number of computers working simultaneously and then piece together the results. If this were not the case, Google and other search engines could not exist.

However, the trouble with some tasks is that in their case, this cannot be done, at least not effectively enough to be of any practical use. One example of such a problem that conceptually and methodologically limits the possibilities of advanced AI based on computational algorithms is the famous salesman problem. The essence of the problem can be formulated as follows. A salesman leaves his headquarters and has to visit three cities in succession and return home again. The question is, in what order should he visit the three cities so that his journey is as short as possible? Suppose he leaves Dortmund and is to visit Munich and Dresden. There are six possible routes, but in reality, only three, because they are circuitous, and we can, therefore, go one way or the other, which does not change the distance. For example, we can calculate that the optimal route is Dortmund–Munich–Dresden–Dortmund, or the other way around. There are six possible journeys in three cities, with four cities there are 24 possible journeys, and with five cities, there are 120 possible journeys, which still looks neither threatening nor dramatic. It is just that the general formula is $N!$ or $N$ factorial, that is, the product of $1 \times 2 \times 3 \times \ldots \times N$. For 25 cities, that is a little more than $15.5 \times 10$ to the 24th. It has been about $4.25 \times 10$ to the 18th second since the universe began, according to today's knowledge, so if you run a computer at the moment of the big bang that calculated 30 million combinations of distance per second, you still would not have the number of possible routes for 25 cities. At first sight, shocking, even unimaginable, it is nevertheless true. Unless some now unimaginable twist occurs, the problem of the business traveler will remain virtually unsolvable, even though any small child can understand it and do the necessary calculations. So, when it comes to the future and the fundamental limitations of AI, will we remain squares or become cubes?

The ability to tell stories is almost exclusively associated with human authors, yet this area of human activity is also a core area of investigation for computer scientists who are trying to program a computer to emulate this ability (Levenson, 2014; Callaway and Lester, 2002; Labbé and Labbé, 2013). The algorithms used to generate stories are called Story Generator Algorithms, or SGAs for short. Thus, the final output of these algorithms are stories, but currently, the main goal of programmers is their functionality, not their esthetic value. Thus, the main criterion for evaluating them is not whether these stories are readable and appealing to the reader, but their ability to produce a story that has a logical sequence and believability (Gervas, 2009).

Philip M. Parker, a professor at the Paris School of Economics, patented an automatic book creation method (Parker, 2007). The computer program automatically generates books according to templates, drawing data mainly from databases, internet search

engines, or freely available websites and resources. The algorithm uses 70 computers to search for thematically similar information, including accompanying photographs and images. More than 200,000 books have been produced in this way, authored by Parker himself. The machine-written books are sold through Amazon's online store. The pros of automatically generated books include the speed or ability to handle a variety of issues ranging from studies, economic statistics, and rare diseases to forecasts (Cohen, 2008). Another advantage is the low cost, as the computer-generated compilation is only printed after the customer orders it. However, the robot does not bring any new information; besides, the content of the book is characterized by an impersonal style of language and the absence of any plot. Most of the generators are very user-friendly and simple to use. They are mostly available online, or their source code is freely downloadable. There are different types of generators providing fast outputs after specifying the names of authors, language, or length of the generated text. The Basic Automatic BS Essay Language Generator, or Babel Generator for short, can produce an essay whose content is also meaningless but whose text is grammatically correct. The generator was developed under the supervision of Les Perelman at the Massachusetts Institute of Technology (MIT). The software was designed to demonstrate the inefficiency of machine processing and text evaluation applications.

Ken Schwencke has designed an algorithm called Quakebot for the Los Angeles Times that can automatically generate a short text in the event of an earthquake (Oremus, 2014). The United States Geological Survey sends a report of the earthquake in an e-mail message, from which the robot selects the necessary information and inserts it into a preconfigured template, to which it adds maps and headlines. The generated article then appears in the editorial system and immediately alerts the editor, who edits and publishes it. Automatically generated news is also used by Forbes magazine, which uses the Quill artificial intelligence system from the technology company Narrative Science, and by the company Automated Insights, where a robot has been working autonomously without human intervention since October 2014.

Advanced research in the fields of robotics, artificial intelligence, and machine learning is enabling computers to do the work that humans have done in the past. Even so, there are occupations that many scientists believe are not threatened by computers. Writers are one of them. According to the book *The Future Of Employment: How Susceptible Are Jobs To Computerization* (Frey and Osborne, 2013), whose authors surveyed 702 occupations and ranked them according to whether their jobs could be replaced by computer technology, writers and authors rank 123rd (ranked from occupations least likely to be replaced by computer technology to those most likely). Yet computer-generated texts have been an area of interest for scientists and researchers for over 50 years. There are several types of computer-generated texts. They may refer to automatic summaries such as economic summaries, medical manuals, or dictionaries. Another example of computer-generated texts is fake scientific papers. There is currently a lot of pressure on scientists to publish as much as possible. Also, because of this fact, fake scientific paper generators have been developed. One example of such a program is SCIgen (SCIgen: An Automatic CS Paper Generator, 2015), which automatically generates texts. Nevertheless, these texts do not make any sense, and they only look like scientific papers. Text generation can be divided into two categories according to its method. These are combinatorial and automatic text generation. The first attempts at computer-generated texts were made in the 1950s when the so-called Combinatory Text Generation produced the first works.

Computational creativity is a discipline that falls under the study of artificial intelligence. One of the goals of this scientific field is to create software that will be able to work creatively according to parameters set by humans. However, exploring computational creativity raises many questions that are not easily answered. First and foremost is the question of what creativity is. Although the subject of much research, this ability is not clearly defined, and everyone has a different idea of the word. The general idea of creativity is that someone invents something new. This "new thing" is supposed to be something original and to achieve a goal (although this is sometimes unclear). The previous paragraph implies that its novelty and originality judge the outcome of the creative process. Historical creativity is judged by whether the final product of the creative process is novel in the context of all human history, while psychological creativity judges the novelty of the output of one particular person only in the context of his or her work. This approach to creativity implies that if computers were to achieve historical creativity, they would have to have access to historical data and interact with other creators. However, if this condition is not met, computers can only achieve psychological creativity.

In terms of originality, there is also the question of whether computers can create anything new, meaningful, surprising, and valuable at all (Bridy, 2012). Computer skeptics see the main problem precisely in the fact that software that mimics human creativity is made up of algorithms, which are defined as a finite and generalizable sequence of instructions, rules, or linear steps designed to achieve a specific, predefined goal (Gervas, 2009). In the context of this definition, then, the computer would function only as a tool for realizing the goals of the software author, not as the creator himself, so it would be impossible to speak of computer creativity. However, this definition only applies to deterministic algorithms. However, there are more types of algorithms that make up software capable of artistic production, such as in music and the visual arts. It is also possible to program algorithms to incorporate elements of randomness to achieve unexpected and surprising results. Another argument against the claim that using computers is incapable of originality due to their algorithmic nature is the claim that the brain itself is also just a machine. This implies that writers are just typewriters that process existing stories, create rules for their creation, and then write text according to those rules.

Computer-generated texts can, therefore, be considered a work protected by copyright law. However, the question concerning the authorship of these works remains (Barbosa, 2014). There are several possibilities. The author may be the creator of the program according to whose instructions the computer program generated the work. Alternatively, the person who used the program to generate the text can be described as the author. Another possibility is the program itself. The last option is to attribute the work to all of the above as a form of co-authorship. All of these approaches raise questions and issues that need to be addressed. Attributing authorship to the creator of a program that is designed to generate artwork follows the logic of determining authorship of computer games—whoever wrote the program that generates the artwork is the author of the artwork. In this case, however, the contribution of the computer program and its participation in the process of creation would be overlooked. In some works, this participation is only partial, but others (such as visual art or musical works) are the result of computer activity without human input.

Thus, it is possible to view the resulting work as a co-authorship between a human and a computer. In other countries, the law reflects these shifts in perceptions of authorship differently. In the US, the computer or the creator of a computer program is not considered the author of the work; the author is the user of the program who created the work with its help. In the UK and New Zealand, copyright law defines a computer-generated work as "a

work generated by a computer in circumstances such that there is no human author, and copyright is attributed to the person who has taken the necessary steps to produce the work (Bridy, 2012). In the electronic age, creating new and recycling old texts and manuscripts is deceptively easy.

Joshua Brown, head of New York-based investment advisory firm Ritholtz Wealth Management, wrote that there would be only three types of employees in the future. Those who will tell robots what to do. Those who will be told to do the same by robots. And finally, those who will fix the robots. In short, modern AI technology is already acting as a catalyst, multiplying opportunities but also multiplying problems and threats in equal measure and intensity.

## Liability for conduct and the problem of the distribution of unavoidable harm: Example of driverless cars

When Stanley Kubrick introduced his film 2001: A Space Odyssey to the world in 1968, the scene where an AI-controlled computer (HAL 9000) causes the death of a human being of its own volition might have seemed like a really big science fiction movie. However, in 2018, there was a case of an AI car running over a pedestrian crossing the road and causing her death for the first time (Levin and Wong, 2018). In the end, it turned out that the death of E. Herzberg was caused by a combination of the driver's inattention to the situation on the road at the time of the collision and the autonomous car's special settings preventing the AI from using the handbrake (manual braking was necessary as the AI only detected the person one second before impact). Still, it is possible that in the future, we will see cases where the car hits a pedestrian or causes an accident of its own "volition", i.e. in the exercise of full autonomy. The issue of liability for actions and the problem of the distribution of unavoidable harm in the context of autonomous intelligent systems and AI platforms are also crucial today (Hintze, 2016), as the world's largest car companies have regularly announced in recent years that the 21st century will bring the rapid rise of fully autonomous vehicles and that encountering them on the streets will be a common experience.

The AI algorithms driving autonomous cars need to be able to respond to a significant number of situations, and indeed, they can respond to a wide range of situations. But how will they cope in the event of an impending accident that cannot be averted (Walker, 2018; Condliffe, 2021). We are faced here with an ethical problem of distributing unavoidable harm that cannot be solved purely technically. Software engineers can undoubtedly create an algorithm that will cause an autonomous vehicle to change its path. But to write such a program and send the car out on the streets with it, they need to know whether the car is making "decisions" according to an algorithm that is ethically acceptable. And that is a problem. While being able to solve technical problems, can we solve ethical problems just as well (Dormehl, 2017)? In an effort to instil morality in their products, software engineers cannot help but turn to ethicists who are concerned with problems of the rightness or wrongness of human actions and choices, including the creation of algorithms that dictate to machines how they should behave. And ethicists can offer them some solutions, but none will be perfect, and none promises to win the universal favor of the entire human population. However, the ultimate choice rests on the shoulders of researchers and engineers. Simply put, our ethical positions are inconsistent; therefore, our recommendations on the ethics of autonomous intelligent systems and AI platforms will necessarily be inconsistent (Bonnefon et al., 2016).

The first task, which is far from easy, that the law will have to take on is the legal definition of artificial intelligence (European Commission, 2018). Such a definition is crucial to distinguish between situations in which the harm has arisen as a result of the actions of that particular system or, conversely, as a result of another computer program or product (Scherer, 2016). However, there is as yet no universally accepted definition of AI. The more an AI system is autonomous and, therefore, capable of making its own decisions without the possibility of prediction of such behavior by its programmers, manufacturers, or others, the more it is inconceivable that any of them could be held responsible for such behavior. In other words, it is unfair to hold any person responsible for harmful outcomes caused by the behavior of systems over which they had no control. However, such an argument cannot stand when it is realized that it is the principle of attributing responsibility to a person for a particular harmful result that is at the heart of the institution of strict liability, which has proved to be fundamental in law. At the same time, it is important to bear in mind the fundamental legal principle that the infliction of a harmful consequence as a result of an act (whether in this case the programming of an algorithm, launching an AI system, or, for example, a defective command to an AI by a particular person) always requires compensation. Accepting the idea that no person is responsible for the actions of an AI is, therefore, unthinkable (Hyatt and Paukert, 2018; Vladeck, 2014; Wein, 1992).

A possible solution to the question of responsibility for the AI's actions is to attribute responsibility to the AI itself. However, such a solution would require that AI systems be granted legal personality, making them the third type of person alongside natural and legal persons. The status of electronic personality then comes with recommendations and guidelines for possible future regulation of liability for damages caused by robots (Delcker, 2018; Hauser, 2017). In the long term, it proposes considering the introduction of a "special legal status for robots […] so that at least the most complex autonomous robots can have the status of an electronic person liable for damages caused by them and the possible use of an electronic person in cases where robots make autonomous decisions or are otherwise independently in contact with third parties". The proposal of electronic personhood has generated a lot of controversies because it contains a number of ambiguities and potential controversies. Nevertheless, the electronic person is a highly revolutionary proposal, which is accompanied by many as yet unanswered questions and which, if it is to be reflected in the current law, will thus be seen only in the more distant future. It will, therefore, have to be determined at the present time which of the persons involved in the "life" of the artificial intelligence, be it the manufacturer, the programmer, or the end-user can be regarded as the most appropriate person to impute the obligation to compensate for damage caused by the artificial intelligence system in situations where the artificial intelligence acts completely autonomously and, therefore, unpredictably for humans.

Statistics show that 9 out of 10 serious accidents on the road today are caused by human error. The World Health Organization adds that around 1.3 million people die each year as a result of road accidents (WHO, 2010). Autonomous cars represent an opportunity to improve these statistics and almost eradicate road accidents over time and reduce traffic collapses, improve the overall mobility of the population, and increase their productivity. However, even a computer or AI software or algorithm is not infallible, and accidents will continue to occur (especially in the early stages of development) even with autonomous cars. However, who will be liable for damages when a car is equipped with AI but is not yet fully autonomous and needs to be controlled by the driver? Moreover, who will be liable when no control is needed? Answers to these questions need to be found as soon as possible, as autonomous cars are likely to be among the first mass-marketed and used products fully controlled by AI. One of

the main issues that could hinder the development of autonomous vehicles is the complexity of determining liability for damage caused by an autonomous robot, platform, or AI algorithm (Garza, 2012). Indeed, at present, traditional rules for determining liability for damage caused by robots capable of making autonomous decisions cannot be applied to such robots because in this particular situation it would be "impossible to determine the party that should provide compensation and repairs the damage caused by the robot".

In this context, the car manufacturer would certainly equip autonomous cars with a system that would recognize whether the driver has taken the wheel or whether the car is being driven by artificial intelligence, and this would determine who is liable for the damage in each situation. In summary, there are three possible solutions for cars using fully or partially autonomous systems. In the case of cars that have recently been the subject of frequent headlines, i.e. cars capable of a certain degree of autonomous driving or cars that are only capable of maintaining a constant speed, for example, the legal system is content with the existing rules on liability for damage arising from the operation of vehicles (except in the rare case where the system does not give the driver time to react). The decisive fact here is the driver's obligation to exercise constant control over the system performing the driving. In the event that an accident occurs as a result of a programmer error, and, therefore, a faulty algorithm and the driver was not obliged to prevent the accident (was not obliged to monitor the steering or had no chance of averting the accident) or in any other way did not violate the rules for the operation of such a car (did not neglect safety precautions), the provision on liability for damage caused by a product defect can be used. In the remaining cases, where the maneuver of the car leading to the damage cannot be classified as a defect due to the full autonomy of the system, the introduction of strict civil liability for the car manufacturer or, jointly and severally, for the manufacturer of the AI system is the optimal solution (Paukert, 2018). However, it should be remembered that cars requiring this third modification are still not on our roads, and it is still a question of when this will actually happen.

However, the business model for autonomous cars remains largely unknown. Will you buy an autonomous car or share it? You will call an autonomous Uber and pay for the kilometers traveled. The question to be answered is as follows. So where do machines and platforms equipped with advanced AI algorithms get some approximation of the values that people would expect and want to have? I think one possible answer is a technique called "inverse reinforcement learning. However, the arguments that many scientists across disciplines claim that robots, computers, machines, and platforms equipped with advanced artificial intelligence (AI) must be able to make self-autonomous ethical decisions contain major methodological (Van Wynsberghe and Robbins, 2018) and conceptual contradictions associated with this approach.

One of the main benefits of the advent of self-driving cars is a reduction in accidents and, as a consequence, fewer fatalities and other injuries. However, it is likely that even the most advanced autonomous cars will not avoid accidents in the future, and it is these accidents that raise ethical questions about how the system should behave at critical moments (Ben-Shahar, 2016; Marchant and Lindor, 2012; Rejcek, 2017; Smith, 2018). In the case of accidents involving human-driven cars, situations in which the driver decides whether, for example, to steer the car off the road to avoid a collision, or, on the contrary, to brake hard, which may cause a collision with the following car, are of a sudden reactive and instinctive nature, in which the will to do harm cannot be observed. But how is the programmer supposed to set up an algorithm in advance to determine which harmful situation to prioritize? In the event of an unavoidable collision, should the car prioritize the lives of the occupants over those of persons crossing the street or over the lives of animals? According to the utilitarian concept, should it always prioritize the situation in which more lives are saved? When choosing between a collision with a helmeted motorcyclist and a non-helmeted motorcyclist, should it prefer to hit the compliant motorcyclist because he or she is more likely to survive, or instead punish the non-compliant motorcyclist even though he or she is likely to suffer a more serious injury? Is it fair to sacrifice the life of an old man to save a child?

In order to explore society's preferences, the Massachusetts Institute of Technology created the Moral Machine website [3], which offers visitors a series of model situations and explores how they would behave in those situations. According to one of the site's creators, it is also the largest international ethics study ever undertaken. The first results of this study, published recently, show that people prefer the life of a child to that of an adult, for example, and that the older the person, the more morally acceptable it is to trade their life for that of a younger person. Conversely, the study did not produce satisfactory results on the question of whether a car should prioritize the lives of its occupants or rather the lives of others. In this context, a Mercedes-Benz representative made a rather controversial statement in 2016 that the company's cars would primarily protect the life of the occupants (Morris, 2016). However insensitive this statement may seem in certain situations, its substance actually has a basis in research published in the same year. In fact, a group of researchers found that while people agree with the idea that an autonomous car should save as many lives as possible, even at the cost of sacrificing the car's occupants, this idea falls out of favor when the occupants are themselves, and they would not buy a car set up in this way for this reason. The majority of respondents to this study added that they do not agree that it should be governments that provide the way autonomous cars are programmed according to these utilitarian principles, and would rather have cars whose programming is not centrally regulated and which can thus behave in critical situations in any way preset by the manufacturer.

At the end of 2018, the German government came up with an action plan that responds to the conclusions of an ethics commission set up by the German Ministry of Transport and Digital Infrastructure, focusing on the issue of Level 4 and Level 5 autonomous cars. In its study, the Ethics Commission concluded, among other things, that (a) in the event of an unavoidable accident, an autonomous system must always prioritize human life over any other interest, be it the life of an animal or property, (b) any distinction based on age, gender, physical or mental state, etc. is unacceptable for the purposes of making decisions to resolve an unavoidable accident situation, and (c) a general algorithm setup that prioritizes injuries to fewer people is justifiable (Federal Ministry of Transport and Digital Infastructure, 2017). In its action plan, the German government has committed to the accelerated development of a legal framework for programming self-driving cars based on these principles (Luetge, 2017). Thus, in Germany, which is one of the pioneer countries of autonomous cars, it is likely to be the government (Burianski and Theissen, 2017), that will determine the answers to ethical questions in the end, although this may not be perceived by society as the most popular solution according to a 2016 study, if only because the life of the crew is not a priority in all situations according to this solution. Related to this is the frequently mentioned and as yet unclear question of whether the way an autonomous car is programmed should even be communicated to its occupants in advance. However, even the Action Plan acknowledges that "dilemma" situations will still need further analysis. Only in the coming months and years will we see how

governments, car manufacturers, programmers, and society, in general, will deal with all these ethical dilemmas.

## Conclusion

It turns out that the ability to understand the functioning of artificial intelligence will be much needed in the near future. Almost no one who comes into contact with its applications will be able to do without it. Artificial intelligence is taking over more and more human activities. The fact that, for the time being, all the development of technologies that use artificial intelligence with deep learning remains in the hands of humans and can, therefore, hopefully, be used to their advantage is crucial for future developments. We need to believe this and not lose optimism. Eric Schmidt put it well when he opened the AlphaGo battle with Lee Sedol: "Whatever the outcome, the winner will be humanity." However, caution and some doubt are in order. This is why we need to further develop the field of AI ethics in order to better prepare and model what possible effects of new phenomena such as the technological singularity and superintelligence, artificial consciousness, the interconnection of AI and nanotechnology, the exponential growth of various forms of AI, etc. will have on the current concept of moral responsibility and on the concept of Trustworthy AI.

Received: 14 February 2022; Accepted: 3 August 2022;
Published online: 17 August 2022

## Notes

1 https://ec.europa.eu/digital-single-market/en/artificial-intelligence#Coordinated-EU-Plan-on-Artificial-Intelligence
2 https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
3 http://moralmachine.mit.edu/

## References

Abbot E (1992) Flatland: a romance of many dimensions. Unabridged edition. Dover Publications
Allen C, Wallach W, Smit I (2006) Why machine ethics. IEEE Intell Syst 21(4):12–17. https://doi.org/10.1109/MIS.2006.83
Allen C, Smit I, Wallach W (2005) Artificial morality: top-down, bottom-up, and hybrid approaches. Eth Inf Technol 7(3):149–155. https://doi.org/10.1007/s10676-006-0004-4
Allen C, Wallach W (2011) Moral machines: contradiction in terms of abdication of human responsibility? In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp. 55–68.
Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28(4):15–26
Anderson M, Anderson SL (2010) Robot be good: a call for ethical autonomous machines. Sci Am 303(4):15–24
Barbosa P (2014) Towards a theory of computer generated texts. In: Torres R, Baldwin S. orgs. PO.EX: Essays from Portugal on Cyber literature and Intermedia by Barbosa P, Hatherly A, de Melo e Castro EM. West Virginia University Press, pp. 143–150
Bhaya A (2017) In a first, Sophia the humanoid robot gets Saudi citizenship, https://news.cgtn.com/news/3167444e32597a6333566d54/share_p.html
Ben-Shahar O (2016) Should carmakers be liable when a self-driving car crashes? https://www.forbes.com/sites/omribenshahar/2016/09/22/should-carmakers-be-liable-when-a-self-driving-car-crashes/?sh=5d7013a048fb. Accessed 22 Sept 2016
Boddington P (2017) Towards a code of ethics for artificial intelligence. In: Artificial intelligence: foundations, theory, and algorithms, 1st edn. Springer
Boden AM (2016) AI: its nature and future, 1st edn. Oxford University Press
Bonnefon JF, Sharif A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352(6293):1573–1576
Bostrom N (2016) Superintelligence: paths, dangers, strategies. Oxford University Press
Bridy A (2012) Coding creativity: copyright and the artificially intelligent author. Stanford Technol Law Rev. https://journals.law.stanford.edu/stanford-technology-law-review/online/coding-creativity-copyright-and-artificially-intelligent-author
Brynjolfsson E, McAfee A (2016) The second machine age: work, progress, and prosperity in a time of brilliant technologies, 1 edn. W. W. Norton & Company
Burianski M, Theissen CH (2017) An important milestone as Germany permits automated vehicles: market impact and outlook. https://www.deutscheranwaltspiegel.de/businesslaw/archiv/an-important-milestone-as-germany-permits-automated-vehicles-market-impact-and-outlook/. Accessed 7 Aug 2017
Callaway CH, Lester J (2002) Narrative prose generation. Artif Intell. http://linkinghub.elsevier.com/retrieve/pii/S0004370202002308
Condliffe J (2021) 2021 May be the year of the fully autonomous car. https://www.technologyreview.com/s/602196/2021-may-be-the-year-of-the-fully-autonomous-car/. Accessed 17 Aug 2016
Cotton B (2018) Calls to halt 'electronic personhood' for robots. https://www.busmessleader.co.uk/calls-to-halt-electronic-personhood-for-robots/45347/. Accessed 21 May 2018
Delcker J (2018) Europe divided over robot 'personhood'. https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/. Accessed 1 Apr 2018
Dignum V (2017) Responsible Artificial intelligence: Designing AI for human values, ITU Journal: ICT Discoveries, Vol 1 (Special Issue no. 1), pp. 1–8
Domingos P (2015) The master algorithm: how the quest for the ultimate learning machine will remake our world. Basic Books
Dormehl L (2017) Thinking machines: the quest for artificial intelligence and where it's taking us next. TarcherPerigee
European Commission (2018) Commission Staff Working Document: liability for emerging digital technologies. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018SC0137&from=en. Accessed 25 Apr 2018
E&T (2021) Uncanny humanoid robot 'Sophia' to enter mass production. https://eandt.theiet.org/content/articles/2021/01/uncanny-humanoid-robot-sophia-to-enter-mass-production/
European Parliament (2016) Draft report with recommendations to the Commission on Civil Law Rules on Robotics. http://www.europarl.europa.eu/sides/getDocJo?pubRef=//EP//NONSGML%2BCOMPARL%2BPE-582J443%2B01%2BDOC%2BPDF%2BV0//EN
Federal Ministry of Transport and Digital Infastructure (2017) Ethics Commission: automated and connected driving. Report. https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile
Frey C, Osborne B (2013) The future of employment: how susceptible are jobs to computerisation? Oxford Martin Programme on Technology and Employment, Oxford. http://www.oxfordmartin.ox.ac.uk/downloads/academic/future-of-employment.pdf
Garza A (2012) „Look Ma, No Hands!": wrinkles and wrecks in the age of autonomous vehicles. New Engl Law Rev 46(3), 581–616
Gervas P (2009) Computational approaches to storytelling and creativity. AI Mag. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2250/2101
Gunkel DJ (2014) A vindication of the rights of machines. Philos Technol 27(1):113–132. https://link.springer.com/article/10.1007/s13347-013-0121-z
Hauser M (2017) Do robots have rights? The European Parliament addresses artificial intelligence and robotics. https://www.cms-lawnow.com/ealerts/2017/04/do-robots-have-rights-the-european-parliament-addresses-artificial-intelligence-and-robotics. Accessed 6 Apr 2017
Hintze A (2016) Understanding the four types of AI, from reactive robots to self-aware beings. https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616. Accessed 14 Nov 2016
Hyatt K, Paukert Ch (2018) Self-driving cars: a level-by-level explainer of autonomous vehicles. https://www.cnet.com/roadshow/news/self-driving-car-guide-autonomous-explanation/. Accessed 29 Mar 2018
Kirchberger T (2017) European Union policy-making on robotics and artificial intelligence: selected issues. Croat Yearbook Eur Law Policy 13:191–214
Labbé C, Labbé D (2013) Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? Scientometrics 94(1):379–396. https://doi.org/10.1007/s11192-012-0781-y.
Levenson E (2014) A. Times journalist explains how a Bot wrote his earthquake story for him. The Wire http://www.thewire.com/technology/2014/03/earthquake-bot-los-angeles-times/359261/
Levin S, Wong J (2018) Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian. https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe. Accessed 19 Mar 2018
Lin P, Abney K, Bekey G (2012) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, MA
Luetge Ch (2017) The German Ethics Code for automated and connected driving. Philos Technol 30:547–558. https://link.springer.com/article/10.1007/s13347-017-0284-0
Marchant G, Lindor R(2012) The coming collision between autonomous vehicles and the liability system Santa Clara Law Rev 52(4):1321–1340
Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21. https://doi.org/10.1109/MIS.2006.80

Morris D (2016) Mercedes-Benz's self-driving cars would choose passenger lives over bystanders. [cit. 4.7.2018], dostupné na: http://fortune.com/2016/10/15/mercedes-self-driving-car-ethics/. Accessed 15 Oct 2016
Oliveira A (2017) The digital mind: how science is redefining humanity. MIT Press
Oremus V (2014) The First News Report on the L.A. Earthquake was written by a robot, slate. https://slate.com/technology/2014/03/quakebot-los-angeles-times-robot-journalist-writes-article-on-la-earthquake.html
Parker PM (2007) Method and apparatus for automated authoring and marketing. https://patents.google.com/patent/US7266767?oq=philip+m+parker
Paukert CH (2018) Why the 2019 Audi A8 won't get Level 3 partial automation in the US. https://www.cnet.com/roadshow/news/2019-audi-a8-level-3-traffic-jam-pilot-self-driving-automation-not-for-us/. Accessed 14 May 2018
Ramasubramanian S (2021) Makers of humanoid robot Sophia are building its sibling. The Hindu. https://www.thehindu.com/sci-tech/technology/makers-of-humanoid-robot-sophia-are-building-its-sibling/article34243237.ece
Rejcek P (2017) When intelligent machines cause accidents, who is legally responsible? https://singularityhub.com/2017/01/30/when-intelligent-machines-cause-accidents-who-is-legally-responsible/. Accessed 30 Jan 2017
SCIgen: An Automatic CS Paper Generator (2015) PDOS: Parallel & Distributed Operating Systems Group. https://pdos.csail.mit.edu/archive/scigen/
Scherer M (2016) Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. Harvard J Law Technol 29(2):353–400
Shalvey K (2021) Humanoid robot Sophia has moved into the art world with a music project and an NFT sale, which reached almost $700,000. https://www.businessinsider.com/sophia-the-robot-launches-music-career-hanson-robotics-nft-sale-2021-4
Smith O (2018) A huge global study on driverless car ethics found the elderly are expendable. https://www.forbes.com/sites/oliversmith/2018/03/21/the-results-of-the-biggest-global-study-on-driverless-car-ethics-are-in/?sh=3b279ca64a9f. Accessed 21 Mar 2018
Stone Z (2017) Everything you need to know about Sophia, The World's First Robot Citizen. Forbes. https://www.forbes.com/sites/zarastone/2017/11/07/everything-you-need-to-know-about-sophia-the-worlds-first-robot-citizen/?sh=712f3fe46fa1
United Nations (2018) At UN, robot Sophia joins meeting on artificial intelligence and sustainable development. https://www.un.org/en/desa/un-robot-sophia-joins-meeting-artificial-intelligence-and-sustainable-development
Vladeck D (2014) Machines without principles: liability rules and artificial intelligence. Wash Law Rev 89(117):117–150
Van Wynsberghe A, Robbins S (2018) Critiquing the reasons for making artificial moral agents, science and engineering ethics. https://doi.org/10.1007/s11948-018-0030-8
Wallach W (2007) Implementing moral decision making faculties in computers and robots. AI Soc 22(4), 463–475. https://doi.org/10.1007/s00146-007-0093-6
Walker J (2018) The self-driving car timeline—predictions from the Top 11 Global Automakers. https://www.techemergence.com/self-driving-car-timeline-themselves-top-11-automakers/. Accessed 29 May 2018
Weng Y (2016) A European perspective on robot law: Interview with Mady Delvaux–Stehres. http://robohub.org/a-european-perspective-on-robot-law-interview-with-mady-delvaux-stehres/. Accessed 15 Jul 2016
Wein L (1992) The responsibility of intelligent artifacts: toward an automation jurisprudence. Harvard J Law Technol 6:103–154
WHO (2010) Global plan for the decade of action for road safety 2011–2020. WHO. https://www.who.int/roadsafety/decade_of_action/plan/plan_english.pdf

## Acknowledgements

This work was produced at the Department of Philosophy, Faculty of Philosophy and Arts Trnava University in Trnava and supported by the project Revised Anthropology: the Concept of the Human in the 21st Century, VEGA 1/0174/21.

## Competing interests

The author declares no competing interests.

## Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

## Informed consent

This article does not contain any studies with human participants performed by any of the authors.

## Additional information

**Correspondence** and requests for materials should be addressed to Tomas Hauer.