Rationality and Ethics in Artificial Intelligence

Rationality and Ethics in Artificial Intelligence

Edited by

Boris D. Grozdanoff, Zdravko Popov
and Silviya Serafimova

Cambridge
Scholars
Publishing

Rationality and Ethics in Artificial Intelligence

Edited by Boris D. Grozdanoff, Zdravko Popov and Silviya Serafimova

This book first published 2023

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-9441-6
ISBN (13): 978-1-5275-9441-8

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This volume includes a selection of papers, written on the basis of talks delivered at three high-level conferences on Artificial Intelligence (AI), held in Sofia in 2018 and 2019. The topics cover a broad spectrum of AI related themes, among which approaches to artificial general intelligence (AGI) and human-like reasoning models, the problem of ethical AI, modern implementations of AI in the economy and defense sectors. The editors want to thank cordially to Prof. Anastas Gerdjikov, rector of the University of Sofia, Academician Julian Revalski, chairman of the Bulgarian Academy of Sciences, Gen Maj. Grudi Angelov, rector of the National Defense Academy "Georgi Sava Rakovski", the Japanese Ambassador to the Republic of Bulgaria H.E. Masato Watanabe, Dr. Ivo Trayanov, chairman of the Defense and International Security Institute (DISI), Prof. Zdravko Popov, in his quality of the chairman of the Public Policy Institute (PPI), Dr. Karina Angelieva, deputy minister of science and education, Dr. Hiroshi Yamakawa, the Director of the Dwango AI Laboratory and Chairperson of the Whole Brain Architecture Initiative (WBAI), and former Chief Editor of the Japanese Society for Artificial Intelligence (JSAI) and Dr. Momtchil Karpouzanov from the American University in Bulgaria (AUBG). The editors also wish to cordially thank Dr. Dimitar Popov for his invaluable help in proofreading the manuscript and editing the formalism.

CHAPTER I

A BRIEF HISTORY OF COMPUTER ETHICS AND HOW IT IS CONNECTED TO AI ETHICS

MARIANA TODOROVA

In his article "A very short history of computer ethics",[1] the author Terrell Bynum defines computer ethics as a scientific field that emerged in the first years of the outbreak of World War II, beginning with Massachusetts Institute of Technology professor Norbert Wiener. He defined the new field while participating in the development of an anti-aircraft weapon, the purpose of which was to intercept and track an enemy aircraft, then calculate its probable trajectory and inform the other parts of the weapon to activate the projectile. The emerging challenges for engineers, according to Bynum, led to the creation of a new branch of science by Wiener and his colleagues, which they called cybernetics, the science of information feedback systems. It was cybernetics combined with the digital computers created at the time that motivated the inventor to draw several important ethical conclusions.

In 1950, he (Wiener) published the first book of its kind on computer ethics (though he nowhere explicitly calls his reasoning that way): *The Human Use of Human Beings*,[2] where he spoke of the benefits and risks of automation and the use of computers. The text sounds like a come true (self-fulfilling prediction), as Wiener predicts that computers will enhance human capabilities, free people from repetitive manual labor, but also allow for processes of dehumanization and subordination of the human species. The

[1] Bynum,"A Very Short History of Computer Ethics", accessed July 8, 2022, https://web.archive.org/web/20080418122849/http:/www.southernct.edu/organizations/rccs/resources/research/introduction/bynum_shrt_hist.html.
[2] Norbert Wiener, *The Human Use of Human Beings. Cybernetics and Society* (New York: Doubleday Anchor Book, 1950).

author warns us not to accept computers as available entities, but to always keep in mind that they are trained (something that actually happens with machine learning) and can go beyond our control. Such a development is a prerequisite for complete dependence and even for control over humanity. The danger, he said, comes from the fact that computers cannot think abstractly and therefore cannot comprehend and evaluate human values.

Wiener adds that the invention of digital computers will lead to a second industrial revolution, which will have multi-layered dimensions, will unfold for decades and will lead to radical changes. For these reasons, he explicitly warns in the chapter, "Someone Communication Machines and Their Future," as well as throughout the book, that workers must adapt to changes in their jobs. Governments need to draft new laws and regulations. Industries and businesses need to create new policies and practices. Professional organizations need to prepare new codes for their members. Sociologists and psychologists need to study new phenomena, and philosophers need to rethink and redefine outdated social and ethical concepts.

Norbert Wiener made valuable recommendations nearly 70 years ago, which unfortunately do not find systematic application to this day. The changes he describes are under way, the exponential development of technology is accelerating, and although there are many projects of universities and research centers on the machine ethics of artificial intelligence, there is still a lack of serious discussion and consensus on key issues that touch.

Walter Manner was the researcher who formalized the term "computer ethics", defining it as part of the applied (ethics) in his work: *Starter Kit on Teaching Computer Ethics*[3] in 1976. He devoted his subsequent work to efforts to emancipate this title as a separate scientific field. The intention to strictly distinguish computer ethics from fundamental ethical issues is implicit.

[3] Walter Maner, *Starter Kit on Teaching Computer Ethics* (Self-published in 1978. Republished in 1980 by Helvetia Press in cooperation with the National Information and Resource Center for Teaching Philosophy).

James Moore, also known by his law of the same name, also dedicated an article[4] on this issue. He believes that the ambiguity surrounding computer ethics arises because there is a political vacuum over how to use computer technology. Through computers, we acquire new abilities that provide us with new opportunities and choices for action. Very often, according to him, there are no political measures for such situations, and if there are any, they are inadequate. For Moore, a central task in computer ethics is to determine what we need to do in specific computer-related situations, such as formulating policies and action guides. Computer ethics, according to him, must take into account both individual and social rights and policies. Therefore, it identifies four areas of computer ethics: 1) identifying the computer-generated policy vacuum; 2) clarification of conceptual ambiguities; 3) formulation of policies for the use of computer technologies 4) ethical justifications.

Moore correctly outlines the steps that should be taken to fill the ethical and then legal and regulatory gaps, but fails to point out that this task is too ambitious to be executed. Philosophers, anthropologists, psychologists and neuroscientists must take the lead in such a task, but they must work alongside representatives of labor and social systems, education, medicine, security and military technology. That is, with experts from all fields who will be influenced by computer technology and artificial intelligence.

Deborah Johnson also contributed to this issue. In her article, "Computer Ethics",[5] she defines it as a field that explores how computers provoke new varieties of classical moral problems and how they worsen and deepen when we apply common moral norms to emerging spheres. She does not share the opinion that this should be a new part of ethics, but simply that a new perspective is set for problems concerning property, power, personal space and responsibility, etc.

---

[4] Moor, James H. "What Is Computer Ethics?" In *Computers and Ethics*, ed. Terrell Ward Bynum (Basil Blackwell, 1985), 266 – 275.

[5] Deborah Johnson, "Computer Ethics", Prentice-Hall, reprinted by *Metaphilosophy,* Vol.16, No. 4 (October 1985): 319-322.

Back in 1995, Krystyna Górniak-Kocikowska predicted in her article[6] that computer ethics, then considered still part of applied ethics, would evolve into a system of global ethics applicable to every culture in the world.

She associates it with the times of the print media, arguing that Bentham and Kant developed their ethical theories in response to this discovery, and believes that the same will be repeated with computer ethics, which must respond to the computer-induced revolution. According to her, the nature of the expected phenomenon (computer revolution) is such that the ethics of the future will have a global character, in the sense that it will also address the integrity of human actions and relationships. Because computers know no boundaries and operate globally, they will set the stage for a new global universal ethic that must be universal to all human beings. The author adds that all local ethics (considering both individual areas and cultural and regional features (of Asia, Africa, Europe, America, etc.) may grow into a global ethic, inheriting computer ethics in information era.

Johnson (1985) also talks about global ethics, but uses a different meaning. For her, it will belong to new kinds of extensive moral problems. Inherited and contemporary ethical theories will continue to be fundamental, and this will not lead to a revolution in ethics. With the help of Terrell Bynum, two opposite concepts of computer ethics are revealed to us. On the one hand, there is the thesis of Wiener, Maner and Gorniyak-Kosikovska about a revolution in ethics and an obligation of humanity to rethink its very foundations, as well as of human life. On the other hand, Johnson's more conservative view is presented, which defends the position that ethics will remain “untouched” and that these are the same old ethical issues in a new reading, which in turn will make computer ethics meaningless as a separate part.

In the debate thus outlined, we can agree in part with statements from both theses. Ethics that address information and computer technology, as well as artificial intelligence, will be global and universal, as responses to the

[6] Górniak-Kocikowska, Krystyna. “The Computer Revolution and the Problem of Global Ethics.” In *Global Information Ethics*, eds. Terrell Ward Bynum and Simon Rogerson (Opragen Publications, 1996), 177–190.

consequences of interacting with artificial intelligence will not be a regional problem or only within nation states. We can assume nuances in the attitude and legal regulations towards the narrow, specialized artificial intelligence, which is not yet competitive, in terms of human brain capacity and awareness. For some cultures, nationalities, or global companies, it will probably be permissible for the personal assistant, artificial intelligence, to have a level of trust with which to perform delegated functions and decisions. For others, this will not be equally valid.

However, when addressing general or superartificial intelligence, it will be necessary for ethical aspects to be universally valid on a planetary level, addressed to all mankind.

Will the code of ethics for artificial intelligence be objective if the discussion is dominated by catastrophic or overly optimistic scenarios?

Mara Hvistendahl, a scientific correspondent for the Guardian, in her article "Can we stop AI outsmarting humanity"[7] considers (only) the emergence of artificial intelligence after about 30 years as the end of human evolution, began 50,000 years ago with the erection of the species Homo Sapiens and 5,000 years ago with the emergence of the phenomenon of civilization, citing the scientist Jean Tallinn. She is a co-founder of Skype and is strongly influenced by the scientist Elizer Yudkowski, according to whom artificial intelligence can hypothetically destroy humanity. On this occasion, Tallinn became a regular donor to the organization of Yudkowski, which dedicates its work to a project for the so-called "friendly artificial intelligence". The category of "friendly," according to Hvischendal, does not mean a reduction to skills such as dialogue with people or that it will be guided only by love and altruism. According to her, "friendly" can be defined as having human motivation, impulses and values. That is, it should not be useful for the machines of the future to arrive at the conclusion that they must erase us in order to achieve their goals. For these reasons, Tallinn founded the Center for the Study of Existential Risk in Cambridge.

[7] Mara Hvistendahl, "Can We Stop Outsmarting Humanity?", *The Guardian*, Match 28, 2019.

The concept behind Tallinn is that software does not need to be programmed to destroy humanity, but can "decide" so along the course of its existence. As we have already noted, this might be the product of a small error, software bug, etc. The scientist also refers to the example of Bostrom, who reveals that artificial intelligence could decide that the atoms in the human body are a good raw material and can be used in another way as a resource. Objections to such arguments come from the Technology Guild, which says it is too early to seek a solution against hostility. They recommend shifting the focus to current problems, such as the fact that most of the algorithms were created by white men and that this fact has given rise to the biases that accompany them.

Stuart Armstrong of the Future of humanity institute at the Oxford Institute also deals with these issues. Armstrong even goes further and suggests that purely physical intelligence be confined to a container and limited to answering questions. Its strategy is to protect people from possible manipulation. It also proposes to have a mechanism for disconnection from people or self-exclusion from the software itself under certain conditions. However, Armstrong fears that these conditions may be sufficient as a measure because artificial intelligence can learn to protect itself or at least develop "curiosity", as there is such an option. In this context, Hvistendahl reports on a programmer, Tom Murphy VII, who invented a program that could teach itself to play Nintendo computer games. In practice, this software is invincible and the only way to stop it is not to play (ibid.).

Tallinn, on the other hand, believes that even if the power button is masked and not of interest to artificial intelligence, there is still no solution to the problem of potential threat, as artificial intelligence may have secretly replicated itself hundreds of thousands of times on the web. For these reasons, researchers and practitioners are united around the idea of artificial intelligence being taught to recognize and study human values. That is, according to Tallinn, he must learn to evaluate people outside the canons of strict logic. For example, that we often say one thing but think another, that we enter conflicts or think differently when we are drunk, and so on. a state of irrationality (ibid.).

Tallinn takes as its best formula the statement of the Cambridge philosopher Hugh Price, who defines that artificial intelligence in ethical and cognitive aspects should be like a "superman". Other questions arise - if we do not want artificial intelligence to dominate us, then should we surpass it. And these questions again inevitably lead us to the presence of consciousness and free will in artificial intelligence.

Boris Grozdanoff in his article "Prospects for a Computational Approach to Savulescu's Artificial Moral Advisor"[8] proposes the creation of a universally valid homogeneous human ethic that can be codified so that artificial intelligence can study it and to create "Artificial moral agent" (AMA). According to him, ethics is also a product of human evolution, which was developed by science and, in particular by philosophy, to create a set of rules to bring order to society through which it can survive and prosper. Grozdanoff launched the thesis of circumstantial normativity in the form of an ethical system. It has to be formalized and axiomatized, which seems like a rather complicated and almost impossible task. In addition to being unified, encircled, and translatable into languages that AI can understand, it must also survive its incarnation in a general/general AI program. For Grozdanoff, the solution lies in the construction of a semantic engine (AMA) that can deliver and handle epistemological processes.

The formula that Hugh Price, Boris Grozdanoff and other scientists offer is correct, but much work is needed to precede it. Today, we are witnessing a resurgent wave of neoconservatism, which sharply criticizes liberal theories such as multiculturalism, globalism, etc. In parallel with these processes, we can predict a resurgence of nationalism, hardening of the concepts of "nation states" and "identities". This context would certainly prevent attempts to seek a universally valid formula for human ethics that could eventually be applied as a matrix for the training of general artificial intelligence. Cultural diversity and different human civilizational norms do not share the same views on the categories of "good" and "bad", human

[8] Boris D. Grozdanoff, "Prospects for a Computational Approach to Savulescu's Artificial Moral Advisor," *Етически изследвания,* бр.5, No. 3 (December 2020): 107-120.

rights, etc. their definition, which mostly fit world-class institutions such as the UN.

There is a call from the European Commission to companies, when creating new software involving artificial intelligence, to use the integration of ethical rules as a competitive advantage. In a document published on April 7, 2019, Europe stated that it would invent and propose global standards to guide all other players in this field. Such requests provoked sharp comments, including from Daniel Castro, vice president of the Information Technology and Innovation Foundation (ITIF).

Undoubtedly, it is crucial that there is a global strategic factor in the face of the European Commission in particular and the European Union in general, which is concerned and will try to impose and take precedence in ethical standards and frameworks for artificial intelligence. The problem is that this position will most likely be peripheral and non-binding.

UNESCO is also developing a solid and intergovernmental document - an ethical framework for artificial intelligence, which will also provide important and meaningful recommendations. From now on, it is important that all scientific papers and articles “meet” with political documents' positions to find the best solution for the ethical development of artificial intelligence.

References

Bynum, Terrell Ward. “A Very Short History of Computer Ethics”. Accessed July 8, 2022. https://web.archive.org/web/20080418122849/http:/www.southernct.edu/organizations/rccs/resources/research/introduction/bynum_shrt_hist.html.

Górniak-Kocikowska, Krystyna. “The Computer Revolution and the Problem of Global Ethics.” In *Global Information Ethics*, edited by Terrell Ward Bynum and Simon Rogerson, 177–190. (Guildford: Opragen Publications, 1996).

Grozdanoff, Boris D. “Prospects for a Computational Approach to Savulescu’s Artificial Moral Advisor.” *Етически изследвания*. бр. 5.,

No. 3 (2020): 107-120. https://jesbg.com/eticheski-izsledvania-br-5-kn3-2020/

Hvistendahl, Mara. “Can We Stop Outsmarting Humanity?” *The Guardian*, March 28, 2019. https://www.theguardian.com/technology/2019/mar/28/can-we-stop-robots-outsmarting-humanity-artificial-intelligence-singularity.

Johnson, Deborah. “Computer Ethics”, Prentice-Hall, reprinted by *Metaphilosophy*, Vol.16, No. 4 (1985): 319-322.

Maner, Walter. 1978. *Starter Kit on Teaching Computer Ethics* (self-published in 1978. Republished in 1980 by Helvetia Press in cooperation with the National Information and Resource Center on Teaching Philosophy).

Moor, James H. “What Is Computer Ethics?” In *Computers and Ethics*, edited by Terrell Ward Bynum, 266–275. (Basil Blackwell, 1985).

Wiener, Norbert. 1950. *The Human Use of Human Beings. Cybernetics and Society*. New York: Doubleday Anchor Book.

CHAPTER II

ETHICAL CLASHES IN THE PROSPECTS FOR AUTONOMOUS VEHICLES

SILVIYA SERAFIMOVA

Introduction

Clarifications

The moral issues regarding the use of autonomous vehicles (AVs) are not a new phenomenon in the ethical discourse.[1] Some of them date back to the concerns about the so-called trolley dilemma, introduced by Philippa Foot in 1967. The dilemma is a thought experiment according to which a fictitious onlooker can choose to save five people in danger of being hit by a trolley, by diverting the trolley to kill just one person.[2]

At first sight, the trolley dilemma looks like a utilitarian moral dilemma, which is based upon calculating the maximization of well-being for more representatives at the expense of the suffering of the few. If that were the case, there would be no dilemma whatsoever. The solution would be one to switch the trolley so that the five people can be saved. However, such a decision is an act utilitarian decision *par excellence*.

In turn, the trolley dilemma is modified within so-called moral design problem, which addresses the moral challenges in building AVs.[3] In this

[1] For the challenges in relating the trolley cases to the ethics of AVs, see Geoff Keeling,"The Ethics of Automated Vehicles," (PhD thesis, University of Bristol, 2020), 45-68.
[2] Philippa Foot, "The Problem of Abortion and the Doctrine of the Double Effect," *Oxford Review*, No. 5 (1967): 5-15.
[3] Geoff Keeling, "Commentary: Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios: Applicability of Value-of-Life-Based Models and

context, moral programming can be examined as displaying adaptations of the trolley dilemma, with the difference that the AVs are preprogrammed to make such decisions.[4]

Automated vehicle technologies are "the computer systems that assist human drivers by automating aspects of vehicle control" including a wide range of capabilities such as antilock brakes and forward collision warning, adaptive cruise control and lane keeping, as well as fully automated driving.[5] The moral concerns derive from the findings of theoretical research robotics, which show that crash-free environment is unrealistic.[6] This means that if a crash is unavoidable, "a computer can quickly calculate the best way to crash on the basis of a combination of safety, the likelihood of the outcome, and certainty in measurements much faster and with greater precision than a human can".[7]

Current projects for AVs aim at building partially autonomous vehicles, assuming that drivers can take back control of the vehicle under given circumstances. Such AVs are considered as artificial moral agents (AMAs) belonging to Levels 3 and 4 of NHTSA's classification.[8] According to the more "optimistic" projects of fully autonomous AVs, one should build AVs as artificial autonomous moral agents (AAMAs) belonging to Level 5 of the same classification. The moral challenges in building AVs concern the strive of the engineers for developing a "universally accepted moral code

---

Influences of Time Pressure," *Front. Behav. Neurosci*, No. 11 (December 2017): 247; Keeling, Geoff. "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles." In *Philosophy and Theory of Artificial Intelligence 2017* (*Studies in Applied Philosophy, Epistemology and Rational Ethics. 44*), ed. Vincent C. Müller (Springer 2018), 259-272. https://doi.org/10.1007/978-3-319-96448-5_29. I refer to the online version of this publication.

[4] Darius-Aurel Frank, Polymeros Chrysochou, Panagiotis Mitkidis and Dan Ariely, "Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles," *Sci Rep*, No 9, 13080 (September 2019): 2.

[5] Noah Goodall, "Ethical Decision Making during Automated Vehicle Crashes," *Transp. Res. Record J., Transp. Res. Board* 2424, No.1 (January 2014): 58.

[6] Goodall, "Ethical Decision Making during Automated Vehicle Crashes," 59.

[7] Goodall, "Ethical Decision Making during Automated Vehicle Crashes," 60.

[8] NHTSA's classification of AVs includes the following levels: no automation (Level 0), driver assistance (Level 1), partial automation (Level 2), conditional automation (Level 3), high automation (Level 4) and full automation (Level 5).

that could guide the machines' behavior".[9] The largest project of how one can "train" AI morality in such situations is so-called Moral Machine experiment. It is an online experimental platform that is designed to examine the moral dilemmas faced by AVs. It collects data of millions of moral decisions for the purposes of training machine-learning algorithms.[10]

Similar to the methodological pitfalls in objectifying the moral outcomes of the trolley dilemma, these regarding the Moral Machine experiment depend upon *who* decides for *whom* under *what* circumstances. Therefore, the moral outcomes can be evaluated by taking into account the plurality of the intersecting perspectives (passenger, pedestrian, observer), as well as the decision-making modes (deliberate, intuitive).[11]

The aforementioned specifications show that if one wants to find objective and morally justifiable solutions to the AV scenarios, one should constructively evaluate the role of different human predispositions in the process of moral decision-making. Recognizing the role of biases is of crucial importance for the AV manufacturers since "sourcing people's moral preferences on a large scale requires developing a standardized and reliable instrument that actually controls for participants' perspective and decision-making mode".[12]

Structure

The main objective of this paper is to demonstrate why finding some potential solutions to the moral design problem and Moral Machine experiment requires one to recognize the challenges in building AVs as moral rather than purely computational challenges. For the purposes of exemplifying the latter, I tackle the benefits and disadvantages of two types of AVs projects, viz. the Rawlsian collision algorithm, which displays a

[9] Frank, Chrysochou, Mitkidis and Ariely, "Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles," 1.
[10] Frank, Chrysochou, Mitkidis and Ariely, "Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles," 1.
[11] Frank, Chrysochou, Mitkidis and Ariely, "Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles," 2.
[12] Frank, Chrysochou, Mitkidis and Ariely, "Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles," 17.

contractualist project and some of Frank et al.'s thought experiments which represent utilitarian and deontological scenarios.[13]

The Section "A contractualist prospect for AVs" is devoted to the role of the Rawlsian collision algorithm. In addition to the investigation of Keeling's criticism of this algorithm, in Section "Rawlsian collision algorithm" I analyze why the way in which Leben elaborates upon Rawls' theory of the original position and the maximin rule necessitates one to rethink the role of value-of-life heuristics for the worst-off people. I also explore how such an analysis can contribute to revealing the diversity of core values: specifically, the values of survival probability and survival. These values are examined as related to the computation of life and death in AV accidents, as represented in Keeling's three scenarios. The scenarios are discussed in Section "Exemplifying Rawlsian collision algorithm". In Section "The role of biases", I aim to explore the impact of people's biased moral preferences upon the evaluation of survival probabilities.

In Section "Building "utilitarian" AVs", I analyze some Frank et al.'s thought experiments regarding the complicated use of AVs. In Section "The role of experimental ethics", special attention is paid to the challenges posed by evaluating the results through the methods of experimental ethics. Consequently, in Section "Some "hybrid" moral explanations", I investigate why the difficulties in justifying moral autonomy of AVs are driven by the way in which Green et al.'s exploration makes room for two triplets–these of *emotions–intuitive decision-making–deontological decision-making* and *cognitive knowledge–deliberate decision-making–utilitarian decision-making*. The objective is to demonstrate why the exaggerated trust in the triplets may trigger the misrecognition of a solution to one-versus-many case as a utilitarian solution, while it can be driven by purely deontological motivation. For the purposes of revealing the reasons behind the conflation of deontological and utilitarian decisions within the AV scenarios, I also examine the role of what I call utilitarian explanatory bias.

[13] The choice of projects demonstrates how both theoretical ethical approaches (these adopted in Rawlsian algorithm) and empirical ethical approaches (these incorporated into Frank et al.'s experiments) require further elaboration.

A Contractualist Prospect for AVs

Rawlsian Collision Algorithm

The moral design problem inherits some of the moral concerns about the trolley dilemma, when examined from a utilitarian perspective. That is why one should look to adopt another approach. An illuminative example of such an approach is found in Rawls' theory of justice. It is elaborated upon by Leben into so-called Rawlsian collision algorithm.

Leben's "contractualist" answer to the moral design problem is grounded into two main ideas borrowed from Rawls–these of the original position and the maximin rule.[14] The original position is "a hypothetical situation in which representative citizens decide on principles of justice to regulate the basic structure of society from a list of alternatives".[15] Each party represents the interests of a sub-group of citizens and all citizens have a representative.[16] The parties in the original position are supposed to decide from the perspective of so-called veil of ignorance. This is a situation where "no one knows his place in society, his class, position, or social status; nor does he know his fortune in the distribution of natural assets, his strength, intelligence, and the like".[17]

For the purposes of maintaining the objective foundations of justice, Rawls introduces so-called maximin decision procedure. The gist of the latter concerns the selection of principles, which provide "the greatest allocation of primary goods to the worst-off citizens".[18] The point is some minimal set

---

[14] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 4.
[15] John Rawls, *Theory of Justice* (Cambridge, Mass.: The Belknap Press of Harvard University Press, 1971), 122. Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 2.
[16] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles.", 2.
[17] Rawls, *Theory of Justice*, 137. Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 2.
[18] Rawls, *Theory of Justice*, 150-161. Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 3.

of rights, liberties and opportunities for the worst-off citizens to be guaranteed.[19]

Regarding the maximin rule, Leben modifies it due to the specificity of survival probabilities. As Keeling points out, the "iterated form of maximin described by Leben is called leximin".[20] The leximin rule resembles the maximin rule, while comparing the survival probabilities of the worst-off person on each alternative. However, the leximin can randomize between two or more alternatives having identical profiles of survival probabilities.[21] In other words, the leximin can compare the second-lowest survival probabilities with the remaining alternatives. By doing so, it can select the highest survival probability to the second worst-off person.

Such a conceptualization raises some significant challenges because the algorithms do not take into account the moral agents' value-of-life heuristics. These agents are recognized as belonging to the group of the worst-off people just because their life is at stake.

Furthermore, such algorithms rely upon one formalized in a moral sense presumption, namely, that survival is the highest good for all group members. Certainly, no one questions the fulfilled probability for survival as being high good in itself. However, the issue is that there might be representatives of the worst-off group who prefer to die rather than surviving with debilitating injuries.[22]

Leben himself is aware of this challenge, admitting that some non-fatal injuries might be evaluated as equivalent or worse than fatal injuries.[23] The problem is that when a lifelong debilitating injury is set versus fatal injury

---

[19] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 3.
[20] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles,", 4, Note 3.
[21] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 5.
[22] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 10.
[23] Derek Leben, "A Rawlsian Algorithm for Autonomous Vehicles," *Ethics Inf Technol* 19, No. 2 (March 2017): 111.

within the Rawlsian collision algorithm, the fatal injury is given priority in making the corresponding decision.

The strive to expand the scope of the maximin rule by introducing the leximin rule requires a reevaluation of Rawls' idea of life project. The leximin rule makes room for comparing the second-lowest survival probabilities on the remaining alternatives, but does not shed light upon whether it might be "more just" for the worst-off person or people to die than to suffer debilitating injuries. In turn, this specification necessitates one to reconsider Rawls' original position. One should also analyze how the veil of ignorance should be elaborated upon so that the graduation of justice can meet the requirements of *an internally differentiated* group of worst-off people.

As Keeling points out, Rawls does not assume the application of the maximin rule as universally valid.[24] He clearly denies its application as a general principle of rational decisions in the case of risk and uncertainty. This clarification puts in question the extrapolation of the maximin rule to that of the leximin, when survival probabilities are at stake. One of the reasons is that when such probabilities are tested, the parties reaching an agreement should not be indifferent to their own and others' life projects. A specification that contradicts the requirements set by the veil of ignorance.

Generally speaking, the moral challenge is to reveal why Rawlsian collision algorithm can only address the moral design problem, but cannot solve it. Keeling describes the gist of the problem by saying that Leben's answer concerns not a set of moral principles, but how one builds an algorithm based upon some principles.[25] Certainly, an algorithm grounded in contractualist principles resists some objections against potential utilitarian collision algorithms.[26] However, I would argue that in the strive for avoiding utilitarian relativism, one revives some crucial moral concerns.

---

[24] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 8.
[25] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 5.
[26] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 5.

Correspondingly to the pitfalls of utilitarian moral math, one may face a questionable contractualist moral math, which "threatens" moral evaluation with its formal regulations.

Exemplifying Rawlsian collision algorithm

Keeling describes three challenges which should be overcome if Leben's answer to the moral design problem is proved as satisfactory.[27]

Scenario 1

The AV can swerve left or right. If the AV swerves left, there is a 0% chance that its passenger will sustain a fatal injury and a 100% chance that its passenger will sustain a lifelong debilitating injury. If the AV swerves right, there is a 1% chance that its passenger will sustain a fatal injury and a 99% chance that its passenger will remain unharmed.

According to Keeling, Leben's algorithm chooses to swerve left because it gives the passenger the greatest survival probability.[28] Certainly, dividing rational preferences into strict and weak preferences necessitates the definition of the preferences to survival as strict preferences and those to non-fatal injuries–as weak preferences. However, regardless of the fact that the preference for survival is considered a strict preference in logical terms, it may turn out that it is a weak preference in moral terms. Extrapolating Keeling's concern about the selection of an alternative, which is not in the passenger's rational self-interest,[29] I would argue that the more serious problem is when the programming is not in the passenger's moral self-interest either.

[27] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 9.
[28] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 10.
[29] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 10-11.

Scenario 2

Keeling's main concern about the second scenario is that the maximin rule gives "undue weight to the moral claims of the worst-off".[30]

The AV can swerve left or right. If the AV swerves left, there is a 100% chance that its passenger will die, and twenty nearby pedestrians will be unharmed. If the driverless car swerves right, there is a 99% chance that its passenger will die, and a 100% chance that twenty nearby pedestrians will receive lifelong debilitating injuries.

Rawlsian algorithm selects the right swerve regardless of how many pedestrians will receive lifelong debilitating injuries.[31] Leben argues that he would always prefer to be one of the injured pedestrians claiming that such scenarios are unlikely to arise.[32] This argument is relevantly criticized by Keeling who claims that the low probability does not make the moral concerns of the pedestrians less important.[33]

Going back to Rawls' theory of the original position, it is apparent that Leben's assumption is grounded in the misinterpretation of the veil of ignorance. In the second scenario, the parties do not choose a principle of justice because it can provide an objective moral treatment to the group of the worst-off people, but because *they* want to be fairly treated if/when *they* occasionally fall into that group. Thus, the requirements of having objective knowledge in the original position and the maximin rule are not fulfilled.

In addition to Keeling's well-formulated concern that a personal preference to a non-fatal but debilitating injury is "not a good moral reason" to inflict a large number of injuries to prevent a single death,[34] one should take into

---

[30] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 11.
[31] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 11.
[32] Leben, "A Rawlsian Algorithm for Autonomous Vehicles," 114; Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 11.
[33] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 12.
[34] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 12.

consideration the problems of moral decisions' quantification. Swerving right, as Leben suggests, is not necessarily a morally acceptable option. This is possible only if both the decision-makers and those belonging to the worst-off group evaluate the survival as being the highest good.

That is why I would argue that the problems with the second scenario do not derive from "the undue weight" to the moral claims of the worst-off people, but rather from the fact that the due weight is not relevantly graduated in moral terms. The lack of such a graduation affects the way in which the group of the worst-off people is determined.

Scenario 3

Keeling points out that there is a scenario which includes an algorithm that assigns a higher survival probability to the worst-off people than Leben's algorithm. This is the greatest equal chance algorithm.[35]

The AV can swerve left or swerve right. If the AV swerves left, there is a 0% chance that Anne will survive, and a 70% chance that Bob will survive. If the AV swerves right, there is a 1% chance that Bob will survive, and a 60% chance that Anne will survive.

Leben's algorithm programs the AV to swerve right because it assigns a survival probability of 1% to the worst-off party.[36] The third scenario brings us back to the moral concerns about the one-versus-one case. When we have to decide who should die, taking into account that there are only two persons involved, the decision cannot be made on the basis of the people's number. One should know who these people are, as well as what their life projects look like. Otherwise, one cannot make an informed decision in moral terms.

This difficulty is not overcome by Keeling's algorithm of the greatest equal chances either. Even if the AV is programmed to construct a "weighted lottery between the alternatives, where the weightings are fixed to ensure

[35] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 12.
[36] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 12.

that the affected parties receive the greatest equal survival probabilities",[37] the following problem occurs. Precising the survival probability of 32.6%, which is certainly greater than 1%, is only the first step in resolving the moral dilemma. The precision of survival probabilities in programming the greatest equal chances does not trigger unquestionable moral consequences for the affected parties. For instance, computing the greatest equal chances for survival does not shed light upon the case when Anne, who could be a mother of three kids, or when Benn, who could be a researcher able to find a cure for cancer, should be sacrificed.[38]

Therefore, even if the most precise algorithm is elaborated upon, this algorithm does not make the moral concerns less significant. They derive not from the computation of life and death, but from life and death as such.

The Role of Biases

The analysis of the Rawlsian collision algorithm shows that the challenges in building AVs derive from people's moral decisions. The crucial role of biases is evident in the way in which the respondents give preference to saving the life of the pedestrian or that of the passenger(s) depending on whether their personal perspective is made salient or not.[39]

While analyzing and extrapolating the findings of the Moral Machine experiment, one should keep in mind that the Rawlsian collision algorithm explicitly avoids the recognition of the personal perspectives of the passenger, the pedestrian and the observer for the sake of achieving an

---

[37] Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 12.

[38] Having argued that there are some collisions in which the greatest equal chances algorithm is preferred to Leben's algorithm" (Keeling, "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles," 13), Keeling elaborates upon his view saying that the greatest equal chances algorithm is "not great" either (Keeling, "The Ethics of Automated Vehicles," 108). The reason is that it depends upon the ties between the affected parties, as well as assuming that each person has an equal moral claim to be saved (Keeling, "The Ethics of Automated Vehicles," 108). These conditions are not satisfied in AV collisions.

[39] Frank, Chrysochou, Mitkidis and Ariely, "Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles," 1.

optimal objectivity by applying the veil of ignorance. The idea is that the self-preserving intentions should be reduced to a minimum so that the group of the worst-off people can be determined in the most objective manner.

In this context, a detailed examination of the human decision-making biases within the Rawlsian collision algorithm can contribute to firstly, limiting the role of the parties' dominating self-preservation attitudes and secondly, demonstrating why neither the group of the decision-makers nor that of the addressees of the decisions are homogenous groups of moral agents. The second clarification, which concerns value-of-life heuristics, can make room for evaluating the heuristics in question in positive terms as well.

Regarding the future development of the Rawlsian collision algorithm within the framework of the Moral Machine experiment, one may investigate its effects if a graduated social norm violation is examined as a part of the decision-making process. Frank et al. provide a thought experiment (Study 6) assuming that the pedestrian's social norm violation results in significant likelihood of being sacrificed.[40] The graduation of the violated norm varying from a low norm violation (when the pedestrian walked in a street with no signs, traffic signals, or crosswalks), going through a control condition (when the pedestrian walked in the crosswalk) and ending up with a high norm violation (when the pedestrian jaywalked at a red light)[41] sets the question of coupling the issue of responsibility with that of guilt.[42]

If the pedestrians in the Rawlsian collision algorithm are placed in the worst-off group due to the objectivity of the high norm violation, does it mean that we can deprive them of the right to survive? Furthermore, do we have the

[40] Frank, Chrysochou, Mitkidis and Ariely, "Human Decision-Making Biases in the Moral Dilemmas of Autonomous
Vehicles," 11.
[41] Frank, Chrysochou, Mitkidis and Ariely, "Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles," 11.
[42] For the different evaluations of whether or not the jaywalker's awareness of undertaking a risk of death or a serious
harm assumes that they tacitly consent to being harmed or killed, see Keeling, "The Ethics of Automated Vehicles," 127-134.

right to deprive them of survival probability because we assume that it is their fault to fall into that group?

Elaborating upon this approach, I argue that Rawls' original position should be modified. If the parties agree to sacrifice themselves, in case they are norm violators, denying the survival probability by default is the only just decision. Certainly, such a line of thought reaches a dead end not only with respect to Rawls' theory of justice.

On the other hand, if the parties agree to sacrifice the pedestrians because they believe that the pedestrians are guilty, such an agreement hardly can be called moral at all. It questions the moral design problem by compromising the initial status of the pedestrians. The moral gist of the dilemma is who has the moral right to decide on behalf of others for their own sake so that one can avoid the implications of moral arbitrariness.

Building "Utilitarian" AVs

The Role of Experimental Ethics

Regarding the implications of moral agents' preferences, Bonnefon, Shariff and Rahwan[43] argue that the participants in the Moral Machine experiment favor "a utilitarian moral doctrine that minimizes the total casualties in potentially fatal accidents, but they simultaneously report preferring an autonomous vehicle that is preprogrammed to protect themselves and their families over the lives of others".[44] When the participants in the thought experiments think about the results of the dilemmas for the greater good of society, they seem to employ a utilitarian moral doctrine. Consequently, when they consider themselves and their loved ones, the participants show

[43] Jean-François Bonnefon, Azim Shariff and Iyad Rahwan, "The Social Dilemma of Autonomous Vehicles," *Science,* No. 352 (June 2016): 1573-1576.
[44] Azim Shariff, Jean-François Bonnefon and Iyad Rahwan, "Psychological Roadblocks to the Adoption of Self-driving Vehicles," *Nature Human Behavior*, No. 1 (September 2017): 694-696.. Frank, Chrysochou, Mitkidis and Ariely, "Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles," 1.