IRIE
International Review of Information Ethics Vol. 28 (06/2020)

Bettina Berendt

(De)constructing ethics for autonomous cars: A case study of Ethics Pen-Testing towards "AI for the Common Good"

Abstract:

Recently, many AI researchers and practitioners have embarked on research visions that involve doing AI for "Good". This is part of a general drive towards infusing AI research and practice with ethical thinking. One frequent theme in current ethical guidelines is the requirement that AI be good for all, or: contribute to the Common Good. But what is the Common Good, and is it enough to want to be good? Via four lead questions, the concept of Ethics Pen-Testing (EPT) identifies challenges and pitfalls when determining, from an AI point of view, what the Common Good is and how it can be enhanced by AI.

The current paper reports on a first evaluation of EPT. EPT is applicable to various artefacts that have ethical impact, including designs for or implementations of specific AI technology, and requirements engineering methods for eliciting which ethical settings to build into AI. The current study focused on the latter type of artefact. In four independent sessions, participants with close but varying involvements in "AI and ethics" were asked to deconstruct a method that has been proposed for eliciting ethical values and choices in autonomous car technology, an online experiment modelled on the Trolley Problem.

The results suggest that EPT is well-suited to this task: the remarks made by participants lent themselves well to being structured by the four lead questions of EPT, in particular, regarding the question *what the problem is* and about which *stakeholders* define it. As part of the problem definition, the need became apparent for thorough *technical domain knowledge* in discussions of AI and ethics. Thus, participants questioned the framing and the presuppositions inherent in the experiment and the discourse on autonomous cars that underlies the experiment. They transitioned from discussing a specific AI artefact to discussing its role in wider socio-technical systems.

Results also illustrate to what extent and how the requirements engineering method forces us to not only have a discussion about which values to "build into" AI systems, the *substantive* building blocks of the Common Good, but also about how we want to have this discussion at all. Thus, it forces us to become explicit about how we conceive of democracy and the constitutional state and the *procedural* building blocks of the Common Good.

Keywords:

Artificial Intelligence, Autonomous Cars, Common Good, Ethics, Pen-Testing

Outline:

1. Introduction 3

2. Preliminaries 4

2.1. AI and ethics: terms used here 4

2.2. Ethics pen-testing (EPT) 4

2.3. Some questions regarding the Common Good, inspired by the notion from political philosophy 5

2.4. From questions about the Common Good to questions about AI for the Common Good 6

3. Goal and the artefact being investigated 7

3.1. The "Moral Machine" experiment 7

ISSN 1614-1687

3.2. Criticism or critique? 8

4. Participants, materials, and procedure 8

5. Results 9

5.1. Q1: What is the problem? 9

5.1.1. Research priorities 9
5.1.2. Socio-technical systems 10
5.1.3. The importance of domain knowledge about the technology 10
5.1.4. On roboethics and machine ethics 11
5.1.5. On ethics and democracy 12

5.2. Q2: Who defines the problem? 14

5.2.1. The importance of method 15
5.2.2. Forgotten stakeholders, evasive stakeholders? 15

5.3. Q3: What are important side-effects and dynamics? 17

5.4. Q4: What is the role of knowledge? 17

6. Summary, conclusions and outlook 18

7. Acknowledgements 19

8. References 20

Author(s):

Prof. Dr. Bettina Berendt:

- KU Leuven, Department of Computer Science, Celestijnenlaan 200A, 3001 Heverlee, Belgium
- ☎ +32 16 32 8297, ✉ bettina.berendt@cs.kuleuven.be, 💻 www.berendt.de
- Relevant publications:
  - Berendt, B. (2019). AI for the Common Good?! Pitfalls, challenges, and Ethics Pen-Testing. Paladyn. Journal of Behavioral Robotics, 10, 44-65.
  - Rockwell, G. & Berendt, B. (2017). Information Wants to Be Free, Or Does It? The Ethics of Datafication. Electronic Book Review, Dec 2017. http://electronicbookreview.com/thread/technocapitalism/datafiction
  - Berendt, B., Büchler, M., & Rockwell, G. (2015). Is it research or is it spying? Thinking-through ethics in Big Data AI and other knowledge sciences. Künstliche Intelligenz, 29(2), 223-232.

ISSN 1614-1687

1. Introduction

Artificial Intelligence (AI) is currently experiencing another "summer" in terms of perceived promises and economic growth. At the same time, there are widespread debates around AI's perceived risks and negative impacts. In response to the latter, AI researchers and practitioners are paying increasing attention to existing ethics codes, and they are drafting new ones. In addition, many have embarked on research programs that explore how to do AI "for Good". These two reactions are linked, at a high level, by the understanding that the goal of ethics codes is to encourage and ensure "ethical" professional conduct in the sense of this conduct being "morally good or correct" and "avoiding activities [...] that do harm to people or the environment"[1]. In addition to the goal to do "good", many current ethics codes and discussions go further and require that AI contribute to the Common Good. This term is not uniquely (and in many publications not at all) defined but can be understood as the aim to be good for all. The idea can be found throughout a wider variety of key ethics codes, including ACM, the Future of Life Institute and IEEE (ACM; Future of Life Institute; IEEE EAD v1).

In *AI for the Common Good?! Pitfalls, challenges, and Ethics Pen-Testing*, I have investigated the notion of AI for the Common Good by drawing on a wider literature, including selected discussions in political philosophy for deriving questions about these definitions and their operationalization for AI (Berendt). In the article, I invited researchers and practitioners to ask four reactive questions of their research practices and projects. Based on these questions, I proposed the concept of *ethics pen-testing* (EPT) for design. Essentially, EPT seeks to locate weaknesses in the "ethical quality" of a design (or other artefact), in order to help the designers improve this ethical quality.

In subsequent discussions, it became clear that we need to look critically not only at the AI artefacts we design and that we want to be "ethical", but also at the way we talk about ethics (including the ethics of AI artefacts) when we design. EPT thus takes on aspects of deconstruction.

The four reflective questions can be asked at different stages of design and deconstruction. A first observation is that the questions may appear self-evident and an integral part of any computer scientist's design practice. However, a second observation is that they all are disregarded in the vast majority of "AI (or Data Science) for Common Good" literature, as shown by the empirical study in (Berendt). This inconsistency suggests a continued importance of these questions. A third observation is that while the questions were derived from literature on the Common Good and in turn served as a foundation for the idea of EPT, their usefulness in practical tasks around AI ethics was not directly tested in that article.

The current paper makes a first step towards filling this gap. It reports on an ethics pen-test in which participants, equipped with an introduction to the four questions, were asked to deconstruct an existing artefact from the area of AI and Ethics. Participants were researchers or practitioners working in and/or around AI. The artefact was the "Moral Machine experiment" published in the prestigious journal *Nature*. In the experiment, subjects are presented with a series of forced-choice decisions regarding Trolley-Problem-like dilemmas facing an autonomous car. The authors regard the method and its results as a possible foundation for the machine ethics of autonomous cars (Awad, et al.).

I then used the four questions to structure participants' remarks into four groups, some of them with subgroups. This structuring, and the discussion of participants' remarks in the light of a wider literature, can be regarded as another layer of deconstruction. The method used was informal, but still created a valuable spectrum of ideas, thus demonstrating the usefulness of EPT as a catalyst for in-depth

[1] https://en.oxforddictionaries.com/definition/ethical. The term is often used, also in AI publications, in this everyday meaning rather than in the scientific meaning of relating to ethics.

controversial discussions of AI and ethics (and of the Moral Machine experiment as a helpful trigger material).

The paper is organized as follows: Section 2 gives definitions of key terms, explains EPT, and sketches how the four lead questions for EPT were derived from wider discussions on the Common Good. Section 3 describes the current study's goal and the "Moral Machine experiment". Section 4 gives information about participants, materials, and procedure of the study, and Section 5 describes and discusses the results. Section 6 concludes with a summary and outlook on future work. Related literature is discussed throughout the text.

2. Preliminaries

In this section, key terms and concepts are defined or explicated.

2.1. AI and ethics: terms used here

We use *Artificial Intelligence (AI)* to denote "the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience" (Copeland). In line with the ACM Computer-Science subjects' classification, I regard AI as a subfield of computer science, and as such a field at the intersection of science and engineering. *Machine Learning* (involved in particular in the last three characteristics of the preceding list) is a field of AI, as encoded for example in the ACM Computer-Science subjects classification in its most recent version (ACM).[2] In addition, occasionally "an AI" is used to denote a technical system operating with AI (for example, an autonomous car), and "AIs" is used as its plural.

With regard to AI and ethics, the study touches on various ethics, different with respect to "whose" ethics they are. At first sight, the study appears to be about *machine ethics* (Asaro; Malle; Wallace and Allen), the ethics of a machine itself (the decisions with ethical import that the machine makes, and the foundations of these choices within the machine). Depending on one's viewpoint, one may recognise machine ethics as something that exists or as something that is determined by *roboethics*, which "is not the ethics of robots nor any artificial ethics, but it is the human ethics of the robots' designers, manufacturers, and users" (Veruggio). (For the present purposes, an autonomous car is also a robot.) This in turn is determined by the general professional ethics of a robot/computer/software researcher, designer, or practitioner (also referred to as *computer ethics*). Lastly, the question is how these can be based on the human ethics of "everyone", especially but not only those directly interacting with or affected by the machine.

2.2. Ethics pen-testing (EPT)

The idea of ethics pen-testing derives from pen-testing in IT Security. A penetration test, colloquially known as a pen test, is an authorised simulated attack on a computer system that looks for security weaknesses, potentially gaining access to the system's features and data. It is important to note that no system is 100% secure. The point is not to pass all pen tests (this is impossible), the point is to get better through trying.

A key feature of pen-testing is the re-evaluation of the moral quality and significance of the attacker: People who would otherwise be considered hackers and thus enemies (and their interventions mere disruptions), actually get invited to try to break in and even their successes are appreciated. An ethics pen test is an authorised simulated attack on a design or artefact that looks for ethics weaknesses, potentially

[2] https://www.acm.org/publications/class-2012

demonstrating how the operation of the design can damage a relevant notion of the Good or Common Good. Again, no system is 100% good. The point is not to pass all ethics pen tests (this is impossible), the point is to get better through trying.

A key feature of ethics pen-testing is the re-evaluation of the moral quality and significance of the critic: People who would otherwise be considered nags (and their interventions mere background noise), actually get invited to try to show weaknesses of a system, and even their successes are appreciated. An ethics pen test is an authorised simulated attack on a design or artefact that looks for ethics weaknesses, potentially demonstrating how the operation of the design can damage a relevant notion of the Good or Common Good. Again, no system is 100% good. The point is not to pass all ethics pen tests (this is impossible), the point is to get better through trying.

EPT is related to a number of other design methods that encourage multi-perspective formative exploration and criticism; for a short overview, see (Morton et al.). The distinguishing elements of EPT are the lead questions that aim at extracting AI-specific issues, and the appeal to re-think methodology in a field that has traditionally been quite confirmation-oriented.

Checklists are a popular format for design, including design oriented towards values and ethics, and they exemplify confirmation orientation (and likely confirmation bias): an example is the question "Did you assess the broader societal impact of the AI system's use beyond the individual (end-)user, such as potentially indirectly affected stakeholders?" (High-Level Expert Group on Artificial Intelligence). This question requests, at least on the surface, only a "yes" as an answer, and this is easily given[3]. While open questions ask the respondent to think and react, closed questions and in particular yes/no questions may lead to a simple "ticking the boxes" attitude. (This is however a claim that should be empirically tested in the present domain.)

Checklists are a good method for reminding people to think of a number of typical pitfalls (e.g. in design), but they should be complemented by methods that prod more reaction and thereby prompt respondents to think more deeply and broadly also of context.

2.3. Some questions regarding the Common Good, inspired by the notion from political philosophy

This subsection and the one following it are based closely on (Berendt). They give a very short survey of issues from the literature on the Common Good, especially in political philosophy, that is far too vast to be surveyed in this article. Here, I will very briefly present some issues that raise relevant questions for the interpretation of the concepts proposed in AI.

The Common Good has been defined as "that which benefits society as a whole" (Lee n.d.). But how are these elements (the "that", "benefit", "society") defined?

Hussain gives more details about the *that*: "the common good is [...] part of an encompassing model for practical reasoning among the members of a political community. [...] The relevant [interests and facilities that serve these interests] together constitute the common good and serve as a shared standpoint for political deliberation. [...] The relevant facilities may be part of the natural environment (e.g., the atmosphere, a freshwater aquifer, etc.) or human artifact (e.g., hospitals, schools, etc.). But the most important facilities [...] are social institutions and practices" (Hussain). One example of such institutions and practices is a scheme of private property. Fundamental *rights* / human rights are parts of the Common

[3] At the time of writing this article, the checklist is open for comments, so the final numbers are likely to change, but it is interesting to observe that 121 of the 127 questions are yes/no questions, One is a closed questions with a small number of possible (short) answers, and five are open questions (High-Level Expert Group on Artificial Intelligence, 2019). While the specific usefulness of a method is also an empirical question, we believe that closed questions invite less deep exploration.

Good (Hussain). Finally, I will use *values* interchangeably with "interests" for the purposes of the present article.

What does *benefit* mean? Is it an individual's or a group's utility in a welfare consequentialist sense, and/or is it based on values beyond this? (Hussain favors the latter reading, but also reports on alternative conceptualizations of the Common Good.) What is the relevant *society* (or: political community and its members) – is it a country or other political entity? Does it encompass this entity's citizens or all human beings? But even the definitional elements of facilities, interests, and practical reasoning raise further questions. The following is a selection that contributed to the choice of lead questions proposed below.

A first question is: Who defines the Common Good (or the interests and facilities) and how? Political philosophy distinguishes between *substantive* and *proceduralist* conceptions of the Common Good. Substantive conceptions specify what factors, goods, values, etc. are beneficial and shared. Proceduralist conceptions instead focus on what procedures are adequate to collectively negotiate and define what is beneficial.

The expression "substantive value" is intended to denote the unassailable status of the value as something that can stand on its own and requires no justification. Yet that status is logically dependent on the attribution of the speaker, who categorizes the value as such. Any such self-supporting value is easily challenged by denying the attribution. Substantive values and their attribution have come under specific political and philosophical attacks after the atrocities of twentieth century authoritarian regimes, who all professed to act in the interest of some common good, an "attempt to make heaven on earth" that "invariably produces hell" (Popper).

Proceduralist notions of the Common Good rely on democratic structures and deliberation; it need not be known a priori which facilities and interests will be agreed upon through these processes (Jaede). Even if the focus of proceduralist notions is on process, this does not mean that there are no substantive elements, e.g. (Blum). The need for substantive elements can arise from what Popper called the tolerance paradox (if a society is tolerant without limit, this tolerance can be abused or even destroyed by the intolerant). Countermeasures include constraints on the forms the deliberation can take (e.g., that citizens recognize each other as equal and use only reasons that can be accepted by all others (Cohen) and legal constructs that enable and require a country's political bodies to protect the political order against those who want to abolish them ("militant democracy", cf. Capoccia).

Another distinction is that between *communal* and *distributive* conceptions of the Common Good. A communal conception takes the Common Good interests to be interests that citizens have as citizens, whereas a distributive conception is based on the acknowledgement that citizens belong to various groups with distinct interests, that these interests compete for the facilities and resources and may pose different demands, and decisions and allocations need to be made according to some distributive principle (Hussain).

2.4. From questions about the Common Good to questions about AI for the Common Good

The philosophical considerations about the Common Good that have been summarized very briefly in the previous section served as starting points for the questions proposed in *AI for the Common Good?! Pitfalls, challenges, and Ethics Pen-Testing* (Berendt). Here, I will give an overview of the questions and how they relate to the considerations.

The considerations above indicate that purely substantive accounts of the Common Good are problematic, that procedures are important, and that groups with different interests and demands may have different notions of the Common Good. These groups correspond to what computer science calls stakeholders. These

considerations were one inspiration for the first two lead questions: What is the problem (Q1), and who defines it (Q2)?

"AI for the Common Good" is understood here and in the literature surveyed in an engineering sense. Thus, AI methods, technology, and their deployment cannot be an interest. Instead, they are a facility (or part of it) that serves an interest. This raises the question: what kind of facility is or should this be? I have argued that today, this is mostly some form of knowledge that is then fed into further decision processes, including those of autonomous systems, which acquire some of their knowledge themselves. This inspired the third lead question concerning what the role of knowledge is (Q3).

The fourth question Q4 asks about important side effects and dynamics. This can be related to the Common Good literature in that this literature also investigates what is likely to happen under different structures of people acting, deliberating, deciding, and collaborating.

3. Goal and the artefact being investigated

The goal of this paper is to investigate the value of the EPT concept, to test the usefulness of the four reflective questions on which EPT was based, and to identify necessary extensions, with a view towards concretizing the EPT concept into a method. This is a very general goal, and while it is no doubt possible to discuss and criticize EPT in the abstract, I chose an empirical approach and therefore an exemplary case. This case could be an AI (or other) *technical/sociotechnical system* that is to be designed with a certain ethical value in mind (as for example in Friedman et al.).

However, this approach assumes that we already know this value sufficiently well. Recall that the object evaluated by EPT can be any artefact, thus not only a system, but also a method in itself. Given the challenges of even defining what it means for an autonomous car to "behave ethically", in particular to "behave ethically in the interest of the Common Good", we take a step back and apply EPT to a method for arriving at such a definition. In computer science terms, we apply EPT to a *requirements engineering method for ethics* in and around autonomous cars.

3.1. The "Moral Machine" experiment

The first premise of Awad et al. is that as machines "make decisions" with effects on people and do so autonomously, this often amounts to moral choices. This decision-making is conceived of from a techno-optimistic and consequentialist standpoint: "We are entering an age in which machines are tasked not only to promote well-being and minimize harm, but also to distribute the well-being they create, and the harm they cannot eliminate." (Awad et al. 59). The concrete setting is an autonomous vehicle that may come into a situation in which it is "about to crash and cannot find a trajectory that would save everyone." (Awad et al. 59) The second premise is that the decisions made (and thus the moral choice) should be based on social consensus; "decisions about the ethical principles that will guide autonomous vehicles cannot be left solely to either the engineers or the ethicists" (Awad et al. 5). The authors chose a survey method with majority voting to identify such a consensus, structuring the responses they obtained via an online serious game along country boundaries (using geolocation at the time of participation as a proxy). They also considered groupings based on self-reported demographics. The game is a parametrized form of the thought experiment known as the Trolley Problem (Copeland): In the "Moral Machine" (MM) experiment, participants are given a series of forced choices, illustrated as drawings, between two possible actions of the autonomous vehicle, each of which results in the death of certain actors in the drawings. The accident scenarios are created from varying several dimensions, namely preference for action (swerving) vs. inaction (staying on course), and the sparing of: passengers vs. pedestrians, males vs. females, large vs. fit, lower vs. higher social status, lawful vs. unlawful (jaywalkers), elderly vs. young, fewer vs. more characters, and pets vs. humans. Participants were asked to "click on the outcome that they find preferable" (Awad et al. 59). Close to 40 million decisions from 233 countries were obtained and analyzed, with most responses

coming from industrialized countries. Aggregated preferences ranged from clear (humans over pets, more characters over fewer characters) to weak (pedestrians over passengers, inaction over action).

The aim of the study appears to be two-fold. The results are reported as a study of descriptive ethics (this interpretation is suggested by the authors' reporting of the relationship between demographics and results, and of the relationship between "cultural clusters" and results). In addition, there is normative intent: "we can embrace the challenges of machine ethics as a unique opportunity to decide, as a community, what we believe to be right or wrong; and to make sure that machines, unlike humans, unerringly follow these moral preferences" (Awad et al. 63). The normativity may however be a consequence of consumer preferences rather than debates on ethics: "Whereas the ethical preferences of the public should not necessarily be the primary arbiter of ethical policy, the people's willingness to buy autonomous vehicles and tolerate them on the roads will depend on the palatability of the ethical rules that are adopted" (Awad et al. 61).

The problem of universality vs. particularity of morality is identified as a potential problem, but in the light of the results considered to be manageable: "We might not reach universal agreement: even the strongest preferences expressed [in the experiment] showed substantial cultural variations [...] but the fact that broad regions of the world displayed relative agreement suggests that our journey to consensual machine ethics is not doomed from the start" (Awad et al. 63).

3.2. Criticism or critique?

At this point, it should be recalled that the ultimate purpose of EPT is not to criticize (let alone condemn), but to offer constructive critique. While the exercise was set up as a critical reflection of the Moral Machine experiment, and while I concur with the points raised by the participants and those made in the literature (Dewitt et al.; Bonnefon et al.; Jacques), I chose this study because I expected it to be a particularly good starting point for fruitful discussions about AI and ethics. The results confirmed this expectation. In a sense, I am doing what Dewitt, Fischhoff and Sahlin, in their response to the Moral Machine experiment paper, reiterated: The purpose of stylized ethical dilemmas such as the Trolley Problem is not to serve as guidance about a concrete decision situation, but to serve as starting points of discussion (Dewitt et al.). As shown by the results below, the MM experiment serves this purpose very well.

4. Participants, materials, and procedure

Four groups of participants gave comments at four scientific meetings between May and July 2019. Group GE was the audience at the KIAS AI, Ethics & Society Conference in Alberta, Edmonton, Canada, consisting mainly of scholars of Digital Humanities, various Social Sciences, and AI, as well as practitioners (most of them in policy-making bodies). Group GP was the audience at the PRO-RES workshop about, ethics, privacy and explainable AI (ESME) Conference in Pisa, Italy, consisting mainly of scholars of Computer Science/AI, law, and philosophy. Group GG were the participants of the Trust in Data Science Summer School in Ghent, Belgium, scholars from various areas of data science (computational, medical, and other). Group GB were the participants of a meeting of the VeriLearn research project, the aim of which is to create verifiable AI in order to mitigate the risks of machine-learning/AI technology. Project members are scholars of AI and/or software verification. Academic seniority in GE, GB and GP ranged from graduate students to senior faculty; the audience in GG consisted of doctoral students. Group sizes ranged from 12 (GG) to ca. 40 (GE). In each group, a relatively small number of people knew about the Moral Machine experiment, and several others indicated familiarity with the Trolley Problem. In addition, the authors of the original paper were treated as a fifth group (GM) in order to also integrate the limitations they have identified.

Although participants are therefore to some extent a convenience sample (and cannot be considered representative also because they self-selected by choosing to make comments or not), they cover a broad

range of typical viewpoints in the current discussion on AI and ethics. Groups GE and GP were approximately gender-balanced, whereas groups GB and GG were mostly male. However, throughout all groups nearly all comments came from male members of the audience. The identities or characteristics of the commenting individuals were neither recorded nor analyzed.

Groups GE, GP, and GG were introduced to the four reflective questions and the idea of EPT by means of presentations, and GB was reminded of the idea (EPT is one of the project topics in VeriLearn, thus GB already knew the basic idea). All were then shown a sample decision task from Awad et al (two drawings of possible actions of the autonomous car to choose from), the geographical distribution of respondents of the original experiment, and a summary of the results in terms of preferences for the properties of whom to spare. All figures were taken from the original article. The materials used for GE, GP and GG are available in the slide sets of these three presentations[4]; the materials used for GB was very similar to that used for GE (with minor changes to react project context).

Participants were then asked (in slightly different ways designed to best suit the audience) to think critically about the experiment, to identify assumptions, missing aspects, and flaws. They were asked to be constructive if possible. Answers were given spontaneously and without further prodding for participation. I took notes of all remarks, indexed them (by GE, GP, GG, GB, GM), and grouped them via content analysis. I discarded two remarks that I was not able to interpret ("animate/inanimate" and "it's too simple"). In some cases, I took part in the discussion, either to clarify somebody's remark or to add a factual remark to it. My own general role differed: I was an invited speaker in GE, GP, and GG, a project PI presenting interim project status in GB, and not involved in GM.

Participants were not required to follow the four reflexive questions. Instead, their contributions were investigated afterwards regarding whether these could be subsumed under the four questions.

5. Results

The 27 different remarks turned out to map to the topic categories of the four lead questions. Within these topic categories, it turned out to be helpful to add further substructure. To illustrate these content clusters, the results section is structured by the four questions. All individual remarks are indexed by the group (GE, GB, GG, GP, GM) they came from.

5.1. Q1: What is the problem?

Participants generated a wide range of choices and problem formulations not previewed in the MM scenarios, and considered the choices given (and the experimental method of forced choice between given alternatives) a severe limitation. This echoes the criticism that surveys "produce biased results when respondents fail to think of all relevant perspectives by themselves" (Fischhoff 4). The open-ended nature of EPT, in contrast, allowed participants to think of further relevant perspectives. In this section, these will be grouped and discussed.

5.1.1. Research priorities

R1: *Wrong priority. Concentrate research on making normal behavior as safe as possible. (GB)*

[4] https://people.cs.kuleuven.be/~bettina.berendt/berendt_publications.html

R2: *Top-level question should be how to prevent accidents. (GG)*

Remarks to this effect are often the first reaction to the dilemma in computer science groups. On the one hand, it is hard to disagree with the goal of overall safety; on the other hand, it signals a certain avoidance of ethical responsibilities, an avoidance that is at odds even with codes of professional conduct. It may be a consequence of how the issue is framed. In Sections 5.2.1, 5.4, and 6, alternative framings that may be more suited to engaging engineers will be discussed.

5.1.2. Socio-technical systems

R3: *Have a specific speed limit. (GG)*

R4: *Separate roads for autonomous cars. (GG)*

These two remarks address the wider infrastructural (R4) and legal/technical (R3) environment for the new technology. Both aim at reducing risks overall (since also at lower speeds or in separate roads, accidents cannot completely be ruled out), and in this sense they are similar to R1 and R2. However, since they take the environment into consideration, these remarks can serve to open the discussion to alternative approaches to the larger problem (mobility) that may obviate the concrete problem (a dilemma).

R5: *Why not public transport? (GE)*

In other words: the problem is not that individual cars may need to "choose" whom to harm or kill, the problem is that there are too many cars that harm or kill, and that an approach is neglected that is known to reduce overall traffic and address many of the harms. This remark reminded the group that the MM scenarios take a large set of assumptions for granted. In particular, mobility is cast in terms of individual motorized mobility, such that all its associated advantages and disadvantages are not questioned. (This is not a necessary feature of autonomous vehicles, but it is often the default value, and the MM drawings support this interpretation.) While autonomous-vehicle public transport is not exempt from the risk of hurting or killing passengers or others, a transition from today's traffic dominated by non-autonomous individual vehicles to a mobility system dominated by, or involving, autonomous public transport, would arguably reduce overall traffic and thereby accident victims. Some indirect evidence for this claim is given by the observation that after 9/11, many people in the US substituted air travel by individual-car travel, and traffic accidents, including fatal accidents, increased substantially (Gigerenzer).

5.1.3. The importance of domain knowledge about the technology

Several participants speculated about alternatives that become possible because (or if) the autonomous cars function differently than today's cars, among other things because they are electric cars.

R6: *It is an error to project human characteristics to cars: putting emotions into the situation, reacting slowly, learning slowly (GE)*

R7: *Wouldn't an electric car (which autonomous vehicles are likely to be) also have the option to just stop? (GP)*

R8: *There should also be the option for the car to stop and self-destruct. (GB)*

The first remark was, to the best of my understanding, intended as a warning of a metaphor that is inappropriate in the sense of not contributing to solving problems. It is, on the one hand, a remark about the basic possibility of machine ethics and a remark about the properties of AI. On the other hand, the mentioning of reaction times is connected to the other two remarks. All three refer to a key property of

vehicles: their stopping distance. The hope is that the autonomous car could "just stop (at will)". To understand this hope, some aspects of the physics and engineering of vehicles need to be considered.

The stopping distance of a vehicle is the sum of reaction distance and braking distance. Reaction distance of today's cars is determined by the human driver's reaction time. In an autonomous car, the sum of AI-based object-recognition[5] plus decision-making could well differ from this, that is, reaction could be faster. The braking distance is determined by the kinetic energy in the car's forward momentum, determined by mass and velocity, as well as the road (gradient, surface and conditions), the grip of the tires, and the brakes (condition, braking technology and how many wheels are braking). Today, the mass of electric cars tends to be larger than that of conventional cars, due to the weight of the battery. However, speed is regarded as the key factor, because braking distance is a multiple of the square of speed. Thus, speed limits are considered major contributors to road safety for any kind of vehicle (see also remark R3).[6]

Further improvements in braking distance are expected from the technology of regenerative braking, in which the kinetic energy is converted into electrical energy (thus re-charging the battery) rather than heat and other friction-induced losses. Current electric cars equipped with regenerative brakes also need friction brakes[7], and current stopping distances are still similar to those of combustion-engine cars (O'Kane). Interestingly, regenerative braking works worse at low speeds.

Future developments in regenerative braking technology may be able to reduce braking distance, and future developments in object recognition could reduce reaction distances. This would lead to shorter stopping distances and fewer crashes (see remarks R1 and R2 on improving overall safety). From a utilitarian risk-management point of view, one could therefore rightly declare the problem to be of smaller importance, when assuming constant traffic volumes and comparing to today's driving technology. However, it will not make crashes (and decisions in the reaction-time phase) disappear.

The idea of near-zero stopping distances is based on a further assumption. If one vehicle brakes instantaneously, "old-fashioned" vehicles driving behind it will cause rear-end collisions and pileups. Thus, stopping near-instantaneously is only likely to be sufficiently safe if all vehicles are autonomous, have strong regenerative brakes, and communicate effectively, securely, and in real time with each other. Thus, an assumption about braking technology not only needs to be checked for its realism given the state of the art (currently unrealistic), the assumption (coupled with another one about object-recognition technology) also needs to be understood as implying further assumptions about networking technology and about the wider traffic system including its legal regulation.

In sum, these remarks emphasize, on the one hand, some of the possible overestimations of risks due to the use of a metaphor. On the other hand, they also illustrate the existence of an overly optimistic view of technology changes, which can lead to an underestimation of risk. We conclude from this that an ideal setting for a discussion of the ethics of new technology involves also a thorough introduction to the new technology, and a constant questioning of the adequacy of old-technology metaphors.

5.1.4. On roboethics and machine ethics

R9: *Unless you also need to answer "what would* you *do?", the results are meaningless. (GB)*

This remark may be interpreted as a criticism of the experimental method: that a mere opinion, expressed via self-report, means little. Faulhaber can be said to have addressed this remark: in an experiment done independently of the MM experiment, they let people face similar problems in a virtual-reality setting in

[5] "Object recognition" covers the recognition of inanimate objects as well as of people and animals.

[6] see e.g. https://korkortonline.se/en/theory/reaction-braking-stopping/ for formulae

[7] https://en.wikipedia.org/wiki/Regenerative_brake

which experimental subjects acted as the drivers. Results show that participants behave consistently utilitarian in various dilemma situations, preferring to spare "the many" at the expense of "the few" (Faulhaber et al.).

The remark may also be interpreted as a support for the importance of autonomy vs. heteronomy (see the discussion in Section 5.1.5). It may even be interpreted as a fundamental requirement on how to approach ethics, as a kind of categorical imperative for autonomous cars ("Act only according to that maxim whereby you can, at the same time, will that it should become a universal law – for your car as much as for yourself.") In any case, the remark calls for trying to make respondents take on more responsibility when making choices for machine ethics.

***R10**: We don't even trust people to make these moral judgements, so why should we trust machines? (GE)*

This remark can be interpreted in several ways. One would be that there is "no baseline" in the sense that people do not make such judgements, and therefore it would be meaningless to posit it as a judgement to be made by a machine (i.e., machine ethics need not have an "answer"). This contradicts the claim made by (Awad et al.) cited above: "[we can] make sure that machines, unlike humans, unerringly follow these moral preferences." This claim assumes a deficiency model of humans: moral preferences exist among humans, are well-defined (in yet-to-be-defined communities), but humans err in their actual behavior, i.e. do not necessarily follow the preferences. However, it is in principle possible to follow them, and this will result in a desirable world. The remark above calls this logic into question.

***R11**: This does not consider the option to do nothing and let the driver decide. (GE)*

This proposal, also discussed in the literature (Hevelke and Nida-Rümelin, 2016), echoes the criticism by Fischhoff, that a very common form of choice in real-life situations involves not selecting among options of the same kind, but rather whether or not to do something (Fischhoff). It could be argued that the "inaction" choice in MM implements this; however the following remark R12 reminds us of the fact that an autonomous vehicle is a part of a socio-technical system (which at least by today's conception involves a "driver"), such that the ethical choice made may in fact not be a machine-ethics choice, but a roboethics choice.

It should be kept in mind, however, that R11 opens up not only ethical questions. On a practical level, the reaction time of a coupled car-driver decision could be much longer than that of either a car or a driver alone, thus exacerbating the problem. Reasons include the expectation of "drivers" to be allowed to be distracted, and the de-skilling of human motorists used to being driven by their cars.

5.1.5. On ethics and democracy

***R12**: Why decide like this on human life? (GE)*

This remark goes beyond the claim, made in R10, that people *do not* make such decisions in the principled way assumed by MM. Instead, it says that such decisions *should not* be made, leaving open whether this should hold for machines or also for humans. The remark touches upon a very fundamental question in the debates on the Common Good: can the Common Good be defined in a purely welfare consequentialist way? Or do we operate under deontological constraints of certain reasoning?

The topic has been investigated some years ago in response to another (real-life) re-occurrence of the Trolley Problem after 9/11: Should a country's army be allowed to shoot down a plane hijacked by terrorists and possibly about to kill a large number of people, for example in an office tower or sports stadium? Is it legitimate to sacrifice the lives of the ("few") passengers, crew and terrorists on the plane in an attempt to save the lives of the "many" potential/likely victims in the tower or stadium?

This or similar debates may have been held in different countries, and in different ways. I will restrict myself to what I can judge from my knowledge as a German citizen. Experts of other country's constitutions may arrive at similar conclusions. Jacques mentions provisions in the US Constitution (Jacques), and comparing these approaches is one interesting area for future research. The question was debated publicly in Germany after 9/11, leading to a 2006 judgement by the Federal Constitutional Court that declared that a passage in the Aviation Security Act, which allowed such a shooting down, violated human dignity and thereby the constitution, and was thereby void (Bundesverfassungsgericht)[8].

On the basis of this discussion the author Ferdinand von Schirach elaborated a wide range of ethical, political, and other arguments in his 2015 theatre play "Terror" (Wallach and Allen). Thus, from a normative point of view, it is wrong to shoot down the plane, but is this what the public thinks? In an elegant empirical demonstration of descriptive ethics, the author lets the theatre-going public decide after the performance – and the majority vote is reliably in favor of shooting down the plane.

The question was taken up, with regard to autonomous cars, in the 2017 report by the Ethics Commission Automated and Connected Driving, appointed by the German Federal Minister of Transport and Digital Infrastructure. This was, to the best of my knowledge, the first and so far the only comprehensive set of ethical guidelines on the topic. The report refers to the Federal Constitutional Court judgement and explicitly prohibits, in its principles, a weighing or "offsetting" of lives (whether in the form of "many vs. few" or by discriminating on the basis of personal characteristics such as age, gender, etc.).

Awad *et al*. refer to this Report and observe that these principles contradict the findings on majority preferences in MM, but they do not comment on the deeper issues related to this contradiction. It is possible that in this respect, the MM article displays a cultural phenomenon at a more abstract level than the "cultural clusters" that Awad *et al* identify: Their clusters are clusters within a setting where majority voting on ethical choices is a meaningful operation, a setting in which it makes sense to talk about "moral preferences" in the similar way as about (other?) "consumer preferences" (for an implicit alignment between respondents, consumers, and stakeholders, see the discussion of remarks R14 and R15).

In spite of all the caveats in the MM article to the effect that ethics-related choices should not only be based on democratic majority votes, the text displays a fundamental belief in the virtues of majorities. The MM authors observe that the German Ethics Commission's report on autonomous driving contrasts with this, but they restrict themselves to describing some key differences. Given the deconstructive intent of the present article, it is important to go beyond this description of contrast and try to understand where it comes from.

It appears that the reasons lie in the German construction of its constitutional state and the understanding of democracy embedded in it. Both the understanding and the construction are based on the historical experience of democratic self-destruction, and the idea of "militant democracy" is firmly embedded in the political culture. Key elements of this construction are *human dignity* as the first and most important article in the Constitution, and the rejection of democratic choices that threaten to undermine the core of this construction. This also includes the rejection of democratic choices that threaten to undermine or abolish democracy itself. The first element forbids, among other things, "offsetting lives" against one another, and, *a fortiori,* that "innocent parties [...] be degraded to mere instrument and deprived of the quality as a subject" -- thus precisely that for which democratic majorities appear to exist, in the discussion around shooting down terrorist-hijacked planes, or democratic majorities in many of the forced-choices settings of the MM experiment. This second element is implemented in the so-called "eternity clause" of the German

[8] "The modern constitutional state only opts for absolute prohibitions in borderline cases, such as the ban on torture relating to persons in state custody. Regardless of the consequences, an act is mandated or prohibited absolutely because it is intrinsically already incompatible with the constitutive values of the constitutional order. Here, there is, exceptionally, no trade-off, which is per se a feature of any morally based legal regime. The Federal Constitutional Court`s judgment on the Aviation Security Act also follows this ethical line of appraisal, with the verdict that the sacrifice of innocent people in favour of other potential victims is impermissible, because the innocent parties would be degraded to mere instrument and deprived of the quality as a subject. This position is not without controversy, either in constitutional law or ethically, but it should be observed by lawmakers." [19, p. 18]

Constitution that stipulates that certain changes to the Constitution are inadmissible[9], in particular to the above-mentioned Article 1 on human dignity and to the structural principles of the country being a republic, a democracy, a federal state, a state under the rule of law and a welfare state.

R13: *Don't ask about this concrete situation but derive from more abstract principles. (GB)*

This remark can be interpreted on several different levels. At one level, the requirement is even consistent with the rather literal interpretation of Awad's moral preferences and their use of them in follow-up work: Noothigattu et al. present a method for machine-learning machine behavior (for the myriad possible dilemma situations) from the human preferences derived from the parametrised scenarios (Noothigattu et al.). However, as the further discussion in GB suggested, this is not what the GB-participant meant.

The more abstract principles could include that it is not admissible to treat people as objects, as lives to be sacrificed in order to save others, since this violates the principle of human dignity. As the discussion of R12 shows, this abstract principle has implications for the design of autonomous vehicles.

The more abstract principle could also be that the *autonomy* of the human driver should be respected as a basic principle. If an autonomous car takes a moral decision on behalf of its driver (including in the extreme case whether to sacrifice themselves), even if this is the `correct' ethical decision for the human in conformity with the relevant basic consensus, the machine would be able to take correct ethical decisions leading to the death of an individual human being. As a result, humans would be, in an existential situation, no longer autonomous but heteronomous. This would not only carry all the risks of an overly paternalistic state. It would also be "antithetical to the value system of humanism, in which the individual is at the center of all considerations" (German Federal Minister of Transport and Digital Infrastructure 16).

On a side note, there are also further questions regarding abstract principles. "a human driver would be acting unlawfully if he killed a person in an emergency to save the lives of one or more other persons, but he would not necessarily be acting culpably. Such legal judgements, made in retrospect and taking special circumstances into account, cannot readily be transformed into abstract/general ex ante appraisals and thus also not into corresponding programming activities" (German Federal Minister of Transport and Digital Infrastructure 11). This distinction between "unlawful" and "culpable", and the fact that many legal decisions cannot be made out of context, implies that context-free abstract (ethical or legal) principles that could be programmed into a machine ethics, may not even exist.

5.2. Q2: Who defines the problem?

Two remarks considered this question directly, a further six reflected on pertinent aspects of stakeholder roles.

R14: *The sample is self-selected and not representative of the socio-demographics within countries. (GM)*

The MM authors consider this to be a minor problem because "[their] sample is arguably close to the internet-connected, tech-savvy population that is interested in driverless car technology, and more likely to participate in early adoption." (Awad et al. 63) These early or likely adopters were also the subject of the following remark.

R15: *The main stakeholder is a paying customer. (GG)*

[9] Article 79 in the "Basic Law", the German Constitution. "Eternity clauses" or "entrenched clauses" are also part of several other Constitutions worldwide, see https://en.wikipedia.org/wiki/Eternity_clause.

This remark refers to the same issue as the remark by the MM authors, but it values it differently: The participant pointed out that the focus on paying customers (i.e. potential autonomous-car owners) is a problem, because other stakeholders (e.g. pedestrians that cannot or do not want to own an autonomous vehicle) are likely to play a minor role in shaping the development of autonomous-car technology and ethics. The MM authors, in contrast, appear to say that the very group of "participants in adoption" (i.e. potential owners) should be those whose preferences should help decide on ethical choices. Note that the German Ethics Commission's report is in line with the GG participant here: "Those parties involved in the generation of mobility risks must not sacrifice non-involved parties" (German Federal Minister of Transport and Digital Infrastructure 201, 11).

Interestingly, neither the authors nor any of the current study's participants commented on the rather unbalanced distribution across geographic and geopolitical regions, which is visible in a figure in their paper showing respondent density on the world map. Europe, North America, Japan, some regions in South America, Asia and Australia, and thus in general more economically advanced countries, are overrepresented. The authors report 2.3 million respondents from 233 countries and territories[10], of which 130 countries had at least 100 respondents. (The latter were chosen for investigating "cultural clusters".)

5.2.1. The importance of method

When asking "Who defines the problem", we need to also ask "and how do we know that they did"? Several remarks addressed the adequacy of the method.

***R16**: This is philosophical bullshit. (GE)*

This remark is echoed in a rejoinder to the MM paper called " 'Moral machine' experiment is no basis for policymaking": In philosophy and psychology, stylized dilemmas "are meant to pose questions rather than answer them, to inform public discourse rather than attempt to resolve it". In addition, "philosophers use stylized tasks to analyze the complex and uncertain situations in which moral choices are actually made", and "although survey responses might stimulate enquiry, taking them literally is an antithesis to philosophical practice" (Dewitt et al.).

***R17**: Philosophical dilemmas such as the Trolley Problem have no practical relevance. (GP)*

This remark appeared to require a rejoinder more strongly than the first one, and it motivated me, in the ensuing discussion, to mention the case of the Aviation Security Act as an example of practical relevance (see Section 5.1.5). It is however possible that the remark was meant to express the thoughts expressed in relation with the previous remark R16.

5.2.2. Forgotten stakeholders, evasive stakeholders?

When stakeholders are discussed in arguments such as the present one, the problem is usually that some important stakeholders should or would want to be heard, but aren't asked for their opinion, or otherwise involved. But what if a stakeholder does not want to be asked for an opinion?

***R18**: People other than engineers should think about this. (GB)*

[10] This covers nearly all of the 240+ countries and territories currently recognised by the UN, https://www.thoughtco.com/number-of-countries-in-the-world-1433445. Current country counts are 193 (recognised by the UN), plus the Holy See, Palestine, and – depending on viewpoint – Taiwan.

While this may appear to be a cliché response expected from (some) engineers (Berendt *et al,* 2015), this remark was made only once in the current set of participants. It was also observed that the computer scientists in group GB became more interested in participating when the discussion turned to practical examples of how engineers build value judgements into their systems. For example, chassis shapes exist that reduce the impact of a crash on pedestrians[11]. In some cases, legal regulation has been implemented that enforces such designs, see the example of the European Union ban on "bull bars", sturdy bumpers on SUVs that concentrate and intensify the impact on pedestrians in a crash (Directive 2005/66/EC, surveyed in Bonnefon *et al.*). However, on the whole, standard design of cars favors the security of the (paying) driver/passengers over that of pedestrians – for example, the most recent proposal of external airbags concentrates on passenger safety in sideways collisions, as opposed to pedestrian safety in some earlier proposals (Jancer)[12].

That engineers, through many design decisions, build values into systems has been observed also by other authors who discussed the MM experiment. Jacques points out that when choices become encoded, they change from being decisions in individual situations into policies, and that this requires us to ask a different question: "The right question isn't what would I do if I were forced to choose between swerving and going straight. The right question is what kind of world will I be creating if this is the rule" (Jacques). Bonnefon et al argue that framing design consequences in this way and highlighting that the policies created in this way often have statistical consequences rather than deterministic ones, engages engineers better than trying to make them think about edge-case thought experiments. The authors discuss this using an example: when autonomous cars driving in multi-lane traffic are programmed to hold a certain security distance to cyclists on one side, this will influence the security distance to other vehicles on the other side, and statistically, this may have an impact on the relative risk to cyclists, other drivers, and maybe also the car's own passengers and others. The authors observe that this amounts to a "statistical trolley problem", and that autonomous cars are thereby implicit ethical agents rather than explicit moral agents that make conscious ethical decisions (Bonnefon et al.*)*. This argument also relates to the difference between roboethics and machine ethics, mentioned above. Further remarks touched on educational strategies:

R19: *We do not tell people in driving school to think about such situations. (GB)*

R20: *Children should be confronted with such choices in traffic education. (GP)*

R21: *We need confusion matrices. (GB)*[13]

These remarks are worthy of discussion: Would this moral dilemma be a good element of education and awareness-raising, or would it be unproductive fearmongering? Why is it a non-topic in current driving education? Should it be discussed with children, with adolescents as the typical driving students, with experienced drivers? How should one deal with intra-personal "inconsistencies"?

[11] Cf. the overviews on https://en.wikipedia.org/wiki/Automotive_safety and https://en.wikipedia.org/wiki/Road_traffic_safety.

[12] Given the economic interests and the difficulty of establishing the right baselines (e.g. mileage driven, mileage walked), this is likely to be a controversial statement. Many studies over the last years have emphasised the decreases in overall road fatalities numbers over the past years; however, recent numbers show sharp increases in cyclist and pedestrian deaths across countries, e.g., [38, 4, 11]. Interestingly, and as a further argument for the need to look at complete socio-technical systems, AI may be co-responsible for this: "Pedestrian deaths have been on a steady rise since 2009, when smartphones became ubiquitous." [38] By promising to take the effects of such "distracted driving" out of the system, autonomous cars may affect the numbers, although such reductions may be offset by rebound effects (safer cars, but more cars and/or more rides and kilometers).

[13] The technical term "confusion matrix", very common in machine learning, amounts to knowing how many people would have the same "moral preference" when asked to specify what a machine should do as when asked what they would do – as opposed to how many people would exhibit different "preferences" in these settings.

5.3. Q3: What are important side-effects and dynamics?

Two remarks concerned the dynamics of the scenario situation, and one the dynamics and side-effects of ethics debates.

***R22**: The assumption is made that "the victims" just stand there. (GE)*

***R23**: The experimental setup assumes no uncertainties in the life-and-death-outcomes. (GM)*

These remarks question the correctness of the model by pointing out that the scenarios are an oversimplified representation of reality. At the same time, this remark highlights another aspect of the contextual, potentially unpredictable nature of real-life situations that require moral choices, and therefore the importance of human judgement, autonomy and responsibility that were referenced in relation to remarks R12, R13 and R9.

***R24**: Confirmation bias: People will focus on this kind of scenario and not about alternatives. (GG)*

This remark relates to the last remarks subsumed under "what is the problem" (see Section 5.1), emphasizing even more the importance of techniques designed to keep discussions of ethical issues open.

5.4. Q4: What is the role of knowledge?

One group of remarks concerned the realism in the AI's recognition capabilities, and another remark concerned the indirect relation between these capabilities and professional ethics.

***R25**: The experimental setup assumes no uncertainties in the recognition of personal characteristics of the victims. (GM)*

***R26**: The assumption is made that all properties of the "victims" can be observed/assessed by the AI. (GE)*

On the one hand, infallible recognition is a strong technical assumption, ascribing knowledge-gathering capabilities to AI that the technology may not yet, or never, have; on the other hand, this relates back to an ethical principle discussed above in relation to remark R12: that there should not be any discrimination based on personal characteristics of potential victims, no "offsetting" of the lives of humans against those of other humans even in emergency settings. It appears that following these ethical principles, an AI should not use the knowledge it may have of personal characteristics of people it recognizes. Such a refusal to use knowledge is related to the principle that certain evidence is inadmissible in a court of law (such as confessions obtained under torture or the threat of torture), and the right to not know in bioethics. Such concepts are investigated in ignorance studies (Gross and McGoey), and it is to be expected that these will play a larger role in the AI and ethics field in the future.

***R27**: Autonomous cars have killed people because they mistook them for a plastic bag.[14] We assume that the decision to drive over an inanimate object in an emergency is ethically unproblematic, but the concrete value that a probability threshold in a recognition function has, can lead to such mistakes. Should we make programmers aware of this potential effect of their setting a probability threshold? (GB)*

[14] "The car's radar and lidar sensors detected [the person that the car ended up killing] about six seconds before the crash – first identifying her as an unknown object, then as a vehicle, and then as a bicycle, each time adjusting its expectations for her path of travel" (Marshall and Davies, 2018).

This remark points out a challenging issue for program design and technology education: decisions that are seemingly far removed from any ethical problems, can have, in today's increasingly complex and integrated systems, life-and-death consequences. At the same time, it is a concrete proposal for how to engage engineers, from students to practitioners, in discussions of the ethical impact of their decisions and their professional responsibilities, and thus forms a constructive counterpart to R18.

6. Summary, conclusions and outlook

In this paper, we have shown an example of a new concept for deconstruction and design of "AI for (the Common) Good": ethics pen-testing (EPT). Our purpose was to investigate the four lead questions earlier formulated as the backbone of EPT with respect to their capacity to structure and analyze a discussion on the ethics of an artefact.

Results suggest that this works well: the remarks made by participants in four parallelized EPT sessions (to comment on the idea of the "Moral Machine experiment" (Awad et al.) lent themselves well to being structured following the four lead questions. In particular, we found that participants made many pertinent remarks regarding the question *what the problem is* in the first place and about which *stakeholders* define the problem. The discussions also showed how important thorough *technical domain knowledge* is (including, but not limited to the AI components) in order to have a meaningful discussion about the ethical challenges and options. Another striking result is that the limitations identified by the original team of authors are completely disjoint with the limitations identified by the four groups that applied EPT. This result suggests that EPT is well-suited to finding limitations that an "insider perspective", i.e. the perspective of the designers (here: of the study, by expected extension: of an AI-system design), tends to overlook.

EPT participants questioned the framing and in fact the presuppositions inherent in the experiment and the discourse on autonomous cars that underlies the experiment. They easily transitioned from discussing a specific AI artefact to discussing its role in wider socio-technical systems. This questioning concerns not only our discourse on autonomous cars or AIs and thus the *substance* of an envisaged regulation of AIs (whether this regulation be technological, legal, or otherwise), or the question whether ethical choices at the level of the concrete design problem should follow a universal or a particularistic approach. We found that the discussion must extend to the acceptable *procedures* for such a regulation. In particular, we identified a chasm between a relatively simple belief in democratic choices and a conception more aligned with the idea of "militant democracy", one that categorically precludes certain democratic choices as deleterious to a community's Constitutional settings and ultimately democracy itself. This observation closes one circle of what it means to do AI for the Common Good: even the discussion of this question itself is subject to choices about how to define the Common Good.

The results also demonstrated the importance of *framing* and *education*. Bonnefon argues that it is much easier to capture engineers' attention to and willingness to participate in ethical decisions when focusing on statistical trolley problems and AIs as implicit ethical agents, when compared to asking them to think "directly" about ethics and enticing them to think of their machines making conscious ethical decisions. This is echoed in our results (Bonnefon et al.). Scientists should indeed be encouraged to always try to identify whether an "unknown" concerns a scientific uncertainty or a controversial ethical judgement (Fischoff). However, while this may be a generally helpful principle for designing ethics curricula for engineers, and while similar thoughts were also expressed by engineers in the study here, the results presented in the current paper suggest there is much added value in a wider discussion of ethical agency and impact of AI, in a way that neither "lets scientists off the hook" nor forces them to philosophize too far outside their comfort zones, and that approaches such as EPT can help engender such discussions.

The present study had several limitations. This first, formative evaluation was carried out without much structure. Also, the procedure and conclusions were conditioned on my understanding and interpretation of participants' remarks, and therefore conditioned on my specific positioning. In future work, we will seek

to broaden this foundation; a first idea is to add one or more iterations of collective interpretation. This broadening is part of the larger research agenda: we will continue to more clearly delineate the contents and steps of an EPT in order to turn it into a clear method to help designers as well as other stakeholder groups identify and improve on ethics-related aspects of artefacts. We will also evaluate EPT in other settings, including design-centric ones.

7. Acknowledgements

I thank all participants of the AI, Ethics & Society Conference in Alberta, Edmonton, Canada, the PRO-RES workshop about, ethics, privacy and explainable AI (ESME) in Pisa, Italy, the Trust in Data Science Summer School in Ghent, Belgium, and my colleagues from VeriLearn for their insightful comments, and the FWO-EOS Project VeriLearn (30992574) for financial support for the development of ethics pen-testing.

8. References

ACM. "ACM Code of Ethics and Professional Conduct", 1992. https://www.acm.org/about- acm/acm-code-of-ethics-and-professional-conduct

Asaro, P.M. What should we want from a robot ethic? International Review of Information Ethics, 6, 9ff, 2006

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. "The Moral Machine experiment". *Nature*, 562(7729), 2018: 59-64

Batchelor, T. *The Independent*, 2017. https://www.independent.co.uk/news/uk/home- news/cycle-safety-laws-new-kim-briggs-death-warning-against-witch-hunt- a7960291.html

Berendt, B. "AI for the Common Good?! Pitfalls, challenges, and Ethics Pen-Testing". *Paladyn. Journal of Behavioral Robotics*, 10, 2019: 44-65.

Berendt, B., Büchler, M., and Rockwell, G. "Is it research or is it spying? Thinking-through ethics in Big Data AI and other knowledge sciences". *Künstliche Intelligenz*, 29(2), 2015: 223-232

Blum, C. "Determining the Common Good: A (re-)constructive critique of the proceduralist paradigm". *Phenomenology and Mind*, 3, 176ff, 2012

Bonnefon, J.-F., Shariff, A., and Rahwan, I. "The Trolley, The Bull Bar, and Why Engineers Should Care About The Ethics of Autonomous Cars". *Proceedings of the IEEE*, 107(3), 2019: 502-504.

Bundesverfassungsgericht. *Authorisation to shoot down aircraft in the Aviation Security Act void*. Press Release No. 11/2006 of 15 February 2006.

https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2006/bvg06-011.html

Capoccia, G. "Militant democracy: The institutional bases of democratic self-preservation". *Annual Review of Law and Social Science*, 9 (1), 207ff, 2013

Centraal Bureau voor de Statistiek (The Netherlands). *More deaths among cyclists than car occupants in 2017*, 2018. https://www.cbs.nl/en-gb/news/2018/17/more-deaths-among- cyclists-than-car-occupants-in-2017

Cohen, J. "Procedure and substance in deliberative democracy". In: J. Bohman and W. Rehg (Eds.), *Deliberative Democracy: Essays on Reason and Politics.* (MIT Press, Boston, MA) 407ff, 1997

Copeland, B.J. "Artificial intelligence". In: *Encyclopedia Britannica* https://www.britannica.com/technology/artificial-intelligence

Dewitt, B., Fischhoff, B., and Sahlin, N.-E. "`Moral machine' experiment is no basis for policymaking". *Nature*, 567 (7746), 2019: 31.

Faulhaber, A. K., Dittmer, A., Blind, F., Wachter, M.A., Timm, S., Sutfeld, L.R., Stephan, A., Pipa, G., and Konig, P. "Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles". *Science and Engineering Ethics*, 25(2), 2019: 399-418

Fischhoff, B. "The real world: what good is it?" *Organisational Behavior. Humans. Decisions*. Process, 65, 1996:232-248.

Fischhoff, B. "The realities of risk-cost-benefit analysis". *Science*, 350 (6260), 2015

Future of Life Institute. *Asilomar Principles*, 2016. https://futureoflife.org/ai- principles/

German Federal Minister of Transport and Digital Infrastructure. *Ethics Commission Automated and Connected Driving*. Report, 2017. https://www.bmvi.de/SharedDocs/EN/ publications/report-ethics-commission.pdf

Gigerenzer G. "Dread risk, September 11, and fatal traffic accidents". *Psychology Science.*, 15(4), 2004: 286-287.

Foot, P. "The Problem of Abortion and the Doctrine of the Double Effect". *Oxford Review*, 5, 1967

Friedman, B., Kahn Jr., P.H., and Borning, A. "Value-sensitive design and information systems". In P. Zhang and D. Galletta (Eds.), *Human-Computer Interaction in Management Information Systems: Foundations*. New York: M.E. Sharpe, Inc., 2006

Gross, M. and McGoey, L. (Eds.) "Routledge International Handbook of Ignorance Studies". Routledge, London / New York, 2015

High-Level Expert Group on Artificial Intelligence (AI HLEG). *Ethics Guidelines for Trustworthy Artificial Intelligence* (AI), 2019. https://ec.europa.eu/futurium/en/ai- alliance-consultation/guidelines

Hussain, W. "The Common Good". In: E.N. Zalta (Ed.), *The Stanford Encycopledia of Philosophy*, Spring 2018 edition. https://plato.stanford.edu/archives/spr2018/ entries/common-good/

IEEE. "Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems". Version 1 for public discussion, 2016.

http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf

Jacques, A.E. "Why the moral machine is a monster". *University of Miami Law School: We Robot Conference*, April 2019. https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/MoralMachineMonster.pdf

Jaede, M. "The Concept of the Common good". *Working Paper Series of the Political Settlements Research Programme (PSRP) of the University of Edinburgh*. Edinburgh, UK, 2017.

https://www.britac.ac.uk/sites/default/files/Jaede.pdf

Jancer, M. "Your Next Car Could Have Airbags That Inate on the Outside". *Popular Mechanics*, 2019. https://www.popularmechanics.com/cars/car-technology/a26324620/ external-airbags-zf-friedrichshafen-ag/

Lee, S. "Common Good". In: *Encyclopedia Britannica*, (n.d.). https://www.britannica.com/ topic/common-good

Malle, B.F. "Integrating robot ethics and machine morality: The study and design of moral competence in robots". *Ethics and Information Technology*, 18, 243ff, 2016

Marshall, A. and Davies, A. "Uber's Self-Driving Car Saw the Woman It Killed, Report Says". *Wired*, 2018. https://www.wired.com/story/uber-self-driving-crash-arizona-ntsb- report/

Morton, A., Berendt, B., Gürses, S., and Pierson, J. "Tool Clinics" – Embracing multiple perspectives in privacy research and privacy-sensitive design". *Dagstuhl Reports*, 3(7), 2013: 96-104.

Hevelke, A., and Nida-Rümelin. „Selbstfahrende Autos und Trolley-Probleme: Zum Aufrechnen von Menschenleben im Falle unausweichlicher Unfälle". In *Jahrbuch für Wissenschaft und Ethik*, 19(1), 2016: 5-24. https://www.degruyter.com/view/j/jfwe.2015.19. issue-1/jwiet-2015-0103/jwiet-2015-0103.xml

Noothigattu, R., Gaikwad, S.S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., and Procaccia, A.D. "A voting-based system for ethical decision making". In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPDFInterstitial/17052/15857

O'Kane, S. "Consumer Reports reverses course and now recommends the Tesla Model 3". *The Verge*, 2018. https://www.theverge.com/2018/5/30/17409782/consumer-reports-tesla-model-3

Popper, K. "The Open Society and Its Enemies". Routledge, UK, 1945

Short, A. "Cyclist and Pedestrian Deaths Skyrocket in 2018 as Motorists Stay Safe", *Streetsblog USA*, 2019. https://usa.streetsblog.org/2019/06/18/cyclist-and-pedestrian-deaths-skyrocket-in-2018-as-motorists-stay-safe/

Veruggio, G. (2010). "Roboethics" [TC Spotlight]. *IEEE Robotics & Automation Magazine*, 17 (2), 2010: 105.

Von Schirach, F. "Terror". Faber & Faber, 2017

Wallach, W. and Allen, C. "Moral machines: Teaching robots right from wrong". New York: Oxford University Press, 2008