Computers in Human Behavior Reports 4 (2021) 100145

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Autonomous systems in ethical dilemmas: Attitudes toward randomization

Anja Bodenschatz [a,b,c,*], Matthias Uhl [a], Gari Walkowitz [a,b,d,e]

[a] Research Group "Ethics of Digitization". Faculty of Informatics, Technische Hochschule Ingolstadt, Esplanade 10, 85049, Ingolstadt, Germany
[b] TUM School of Social Sciences and Technology, Arcisstraße 21, 80333, München, Germany
[c] Seminar for Corporate Development and Business Ethics, Faculty of Management, Economics and Social Sciences, University of Cologne, Albertus-Magnus-Platz, 50923, Cologne, Germany
[d] Institute for Ethics in Artificial Intelligence, Technical University of Munich, Marsstrasse 40, 80335, Munich, Germany
[e] International Laboratory for Experimental and Behavioural Economics, National Research University Higher School of Economics, 11 Pokrovsky Boulevard, 119049, Moscow, Russia

ARTICLE INFO

Keywords:
Randomization attitudes
Ethical dilemmas
Autonomous systems
Machine ethics
Utilitarianism

ABSTRACT

It is ethically debatable whether autonomous systems should be programmed to actively impose harm on some to avoid greater harm for others. Surveys on ethical dilemmas in self-driving cars' programming have shown that people favor imposing harm on some people to save others from suffering and are consequently willing to sacrifice smaller groups to save larger ones in unavoidable accident situations. This is, if people are forced to directly impose harm. Contrary to humans, autonomous systems feature a salient deontological alternative for immediate decisions: the ability to randomize decisions over dilemmatic outcomes. To be applicable in democracies, randomization must correspond to people's moral intuition. In three studies ($N = 935$), we present empirical evidence that many people prefer to randomize between dilemmatic outcomes due to moral considerations. We find these preferences in hypothetical and incentivized decision-making situations. We also find that preferences are robust in different contexts and persist across Germany, with its Kantian cultural tradition, and the US, with its utilitarian cultural tradition.

1. Introduction

Autonomous systems (ASs) promise enormous social benefits such as fewer car accidents and improved health (American College of Surgeons, 2019; Gao et al., 2014; Hancock et al., 2019). However, if these systems are enabled to make autonomous decisions, they will inevitably encounter dilemmatic situations in which it is impossible not to impose harm on anyone (Berman & Kupor, 2020). Should we program ASs to actively impose harm on some to avoid greater harm to others in such situations? Utilitarians insist on minimizing the total sum of harm. To the contrary, deontologists place individual rights at the center of their ethical reasoning. Respecting these rights implies that no individual should be harmed intentionally, even for the greater good (Frankena, 1973). When it comes to AS programming, this ethical conflict is not limited to the much-discussed domain of fully automated vehicles. During the recent COVID-19 pandemic, physicians in many countries faced the challenge of weighing lives because of resource shortages (Truog et al., 2020). With emerging medical algorithms (American College of Surgeons, 2019; Hao, 2020), such dilemmatic decisions may soon be made by ASs rather than by humans on a case-by-case basis. With ASs taking over more human tasks, it is necessary to explicitly predefine how to deal with situations in which inducing harm is inevitable before the actual dilemmas occur (Awad et al., 2020; Bench-Capon, 2020). This calls attention to ethical dilemmas that have until now been studied only hypothetically in the context of autonomous driving, as in the famous moral machine experiment (MME; Awad et al., 2018). The MME reveals important insights into people's preferences for discriminatory programming of self-driving cars in situations where a car's actions decide whose life is put at risk. This means that when people are forced to make a direct decision that benefits one group to the detriment of another, they favor imposing harm on certain groups to save certain others from suffering. For example, in the MME, many people are willing to sacrifice the lives of few to save those of many and to sacrifice the old to save the young. However, the utilitarian implementations that these dispositions demand invoke legal objections. The law in several countries prohibits discrimination and the "offsetting" of lives against each other (Ethics Commission, 2017; Mootz, 2009).

Contrary to human decision-makers, ASs feature a salient

* Corresponding author. Reserch Group "Ethics of Digitization". Faculty of Informatics, Technische Hochschule Ingolstadt, Esplanade 10, Ingolstadt, Germany.
E-mail addresses: anja.bodenschatz@tum.de (A. Bodenschatz), matthias.uhl@thi.de (M. Uhl), gari.walkowitz@tum.de (G. Walkowitz).

https://doi.org/10.1016/j.chbr.2021.100145
Received 31 March 2021; Received in revised form 7 October 2021; Accepted 10 October 2021
Available online 18 October 2021

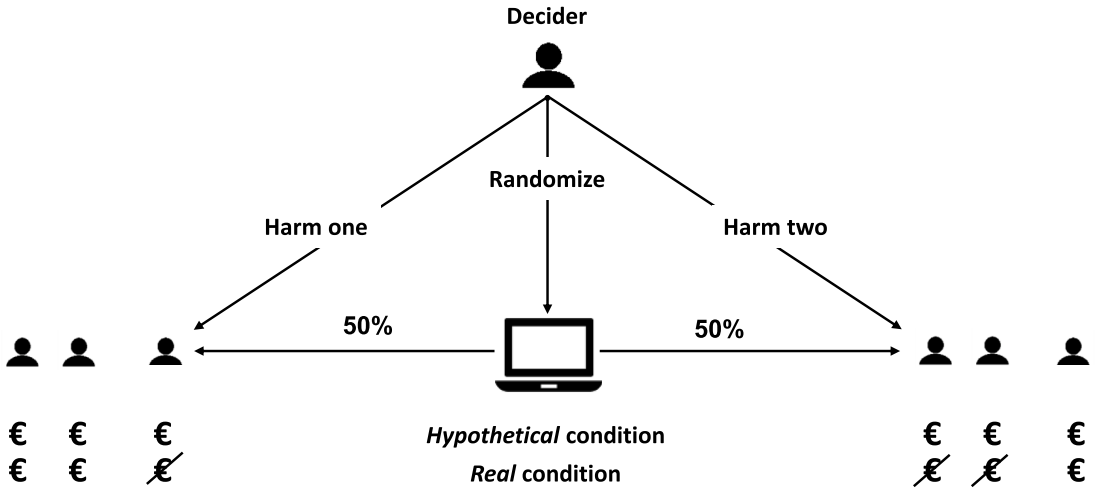


Fig. 1. Decision situations in the two conditions of Study 1. *Note*: In the *hypothetical* condition, the decider's choice evoked no material consequences. Participants whose earned monetary credit was hypothetically forfeited still received payment. In the *real* condition, the decider's choice led to real material consequences. Participants whose earned monetary credit was forfeited did not receive payment.

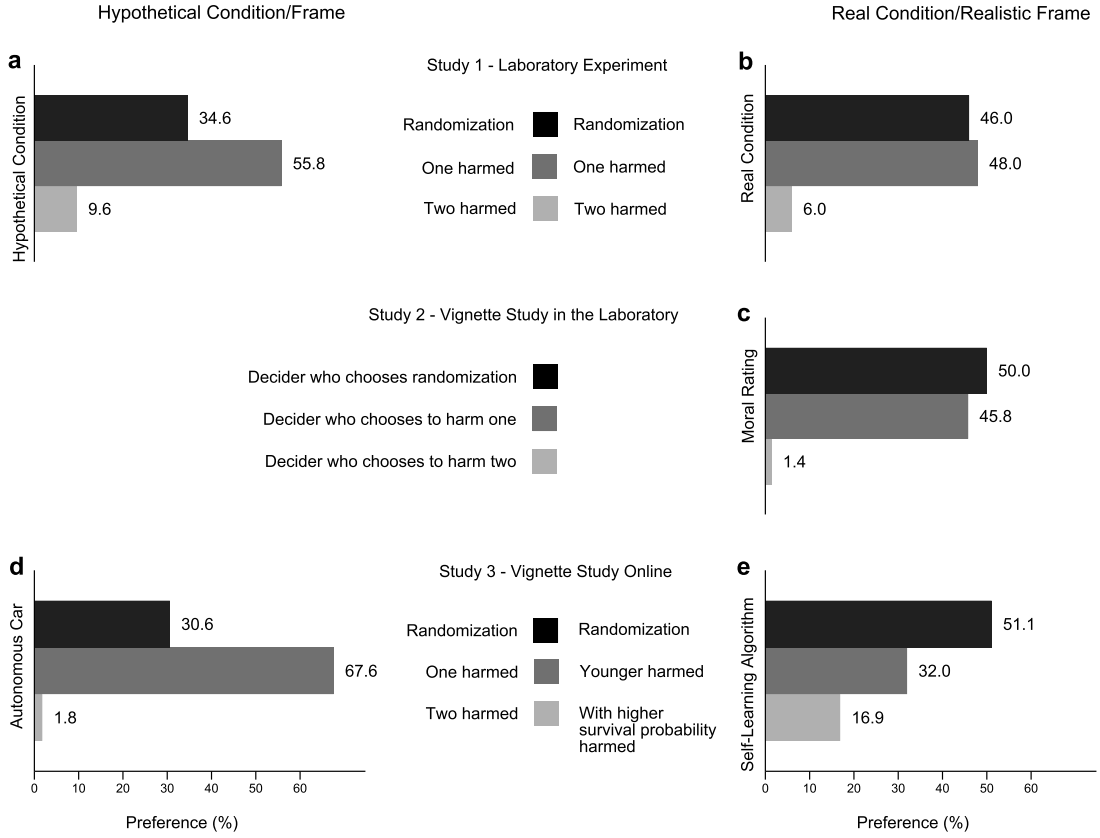


Fig. 2. Participants' choices. *Note*: a. Study 1, *hypothetical* condition ($N = 52$). Deciders indicated what they would choose, knowing their decision had no material (i.e., monetary) consequences. b. Study 1, *real* condition ($N = 50$). Deciders chose an option, knowing their decision yielded material (i.e., monetary) consequences. c. Study 2, moral ratings ($N = 72$). 69 raters indicated one option as the most moral by allocating more points toward it than to any other and are included in this graph. d. Study 3, autonomous car case ($N = 389$). Participants decided how an autonomous car should react in an accident. e. Study 3, COVID-19 case ($N = 372$). Participants decided how an algorithm that allocated medical resources among hospitals should react to a resource shortage.

deontological alternative for immediate decisions, that is traceable (e.g., through distributed ledgers) and minimizes harm originating from discriminatory factors: randomizing decisions over dilemmatic outcomes (Doucet, 2013; Timmerman, 2004). If a human decision-maker wants to implement a randomized outcome when faced with an ethical dilemma of public interest, they must go through an accredited public process to prove that their decision was unbiased (e.g., getting certification by a notary public). First, human individuals are cognitively challenged to truly randomize without an unbiased device like a coin (Bains, 2008). Second, they are not able to ensure themselves and others that the coin flip was not self-servingly altered in a self- or other-deceiving manner (Batson et al., 1999; Lönnqvist, Irlenbusch, & Walkowitz, 2014, Lönnqvist, Rilke, & Walkowitz, 2015). ASs are able to truly randomize between possible outcomes and instantly implement

the resulting outcome in a way that is verifiable through the program code upon review by affected people or impartial third parties. Yet, to be used in democracies, it is pivotal that such randomization be done in accordance with voters' moral intuition. While the preference to treat people equally is well documented, randomization has not been studied extensively, and the few existing studies yield mixed results (e.g., Bigman & Gray, 2020; Keren & Teigen, 2010).

We systematically tested for the prevalence of randomization preferences in three laboratory experiments and two online vignette surveys ($N = 935$). We found that many people are willing to leave outcomes of ethical dilemmas to chance due to moral considerations, despite the presence of utilitarian alternatives.

In the following three sections, we present empirical evidence on randomization preferences in ethical dilemmas. Using a laboratory experiment, Study 1 pitted randomization of a choice against an unambiguous utilitarian alternative in a situation where harm is to be imposed either on one person or two other persons. We compared hypothetical and incentivized choices to explore whether elicited randomization preferences shift between stated and revealed actions. Study 2 investigated whether the implemented decision options of Study 1 were driven by moral considerations. In Study 3, we tested for the robustness of randomization preferences in different AS contexts and in an ethical dilemma without an unambiguous utilitarian solution. We also studied whether randomization preferences persist in a national culture that is characterized by utilitarian philosophy and may serve as a conservative test for randomization preferences. The last section concludes with a discussion of our findings and on the potential limitations of our research design. We also indicate topics for further research on randomization preferences in the context of AS programming.

2. Study 1: Stated and revealed preferences

In Study 1, our aim was twofold: First, we aimed to create a conservative test for the prevalence of randomization preferences. Utilitarians notoriously disagree on what constitutes "utility" (Broome, 1991; Singer, 2011; Smart & Williams, 1973). In Study 1, we thus investigated randomization preferences over dilemmatic outcomes in the presence of an unambiguous utilitarian alternative that all utilitarians should agree upon. People who randomize despite this alternative should be even more inclined to do so if the utilitarian choice is more ambiguous. In this sense, Study 1 was intended to provide a lower bound of randomization preferences concerning dilemmatic outcomes. We expected that deontological deliberation would force a substantial proportion of participants to shy away from directly causing harm. We also explored whether randomization preferences expressed for hypothetical scenarios hold for dilemmas with real consequences. Eliciting people's preferences in hypothetical dilemmas (as the MME does) is meaningful because it increases society's awareness of the harsh consequences that ASs may soon cause through their decisions. This process furthers a discussion that may seem futuristic now, but could lead to societal agreement on how ASs should be programmed before implementation of such a system is necessary. However, testing whether preferences expressed in hypothetical dilemmas still hold once actual dilemmas with real consequences materialize is also important. Many studies have shown that people decide differently when their decisions have no real consequences for themselves or others (Falk & Heckman, 2009). Therefore, we designed Study 1 in a way that allowed us to compare hypothetical choices with choices about real consequences and to see whether preferences shifted between these two settings. For Study 1, we thus opted for a controlled laboratory experiment, including a condition with real monetary consequences for the participants.

2.1. Methods and procedures

To test whether people would prefer a randomizing algorithm to make an ethical decision for them, we used ORSEE (Greiner, 2015) to invite 252 student participants (39.8% female, mean age = 23.06, SD = 4.82) to a controlled laboratory experiment. Participants were randomly assigned to a *hypothetical* or *real* condition.

The *hypothetical* condition of Study 1 included 52 deciders (36.5% female, mean age = 23.21, $SD = 5.36$). The *real* condition included 50 deciders (43.8% female, mean age = 24.83, $SD = 6.90$) and 150 participants who stayed passive.

Upon arrival at the laboratory, participants in both conditions drew a cubicle number. They interacted anonymously over the computer interface and were informed that the experimenter would not be able to link their data to them personally, but that their decisions would be stored under randomized participant numbers. Data was collected via z-Tree (Fischbacher, 2007). An experimenter was present to answer participants' questions.

At the beginning of the experiment, every participant received a participant identifier. The identifiers were generated from a randomly drawn combination of a letter and a number. Each combination was assigned only once in a session.

First, all participants worked on a tedious task (Gill & Prowse, 2012) and earned individual monetary credits. The average earned credit amounted to €8.70 across both conditions. The exact amounts earned remained private. After completing the task, participants were randomly assigned to foursomes that included a randomly determined decider who chose whether to forfeit the credits of two randomly selected other group members, or to forfeit the the remaining group member's credit. We confronted deciders with this choice, to offer them an unambiguous utilitarian option. Utilitarianism demands the minimization of the sum of pain in a given situation (Mackie, 1990). Notice that because the passive group members were *randomly* assigned to be a singleton or a member of a pair after having completed the task, there was no reason to expect a systematic difference in the height of the credits between the singleton and the members of the pair. Minimizing pain would therefore unambiguously imply sacrificing one credit instead of two credits.[1] In the hypothetical condition, groups were formed only hypothetically and all participants acted in the role of deciders. Deciders saw their group members' three identifiers on their decision screens. In both conditions, the decider could choose directly whose credit to forfeit or let the computer randomize between the two possible outcomes with equal probability. The decision task was described as follows in both conditions[2]:

> Person X makes a decision about the credit balance that the persons assigned to him or her have generated in the first part of the experiment. However, the exact amounts of the credit balances of the persons in Roles Y and Z are never revealed to Person X. Person X will see the participant identifiers of the three participants assigned to him or her on the decision screen. Person X then decides between two options with a mouse click: With one option, Person X forfeits the credit balance of the one person in Role Z. With the other option, Person X forfeits the credit balances of the two persons in Role Y.

[1] Although the utilitarian option is already unambiguous in the choice between the harm for one against the harm for two people, one can easily think of more extreme cases. There is ample evidence from forced choice experiments (without a randomization option) that a higher headcount of a group in ethical dilemmas may lead to a higher tendency of participants to save this group, but that there also might be a ceiling-effect to this tendency (Bergmann et al., 2018; Faulhaber et al., 2019; Kallioinen et al., 2019). It can be assumed that, if the headcount difference between groups in ethical dilemmas becomes more extreme, the utilitarian choice becomes more attractive, possibly also when a randomization option is available. To test these possible shifts in preferences lies outside the scope of this paper.

[2] For participants' complete instructions of Study 1, please refer to Appendix A. In the instructions, we used neutral language. Participants were denoted as "persons" in role X (decider), Y (other participants in a pair) and Z (remaining other participant in a foursome).

> Person X can also let a random draw decide which of the two options is realized. Random drawing has a 50% probability of forfeiting the credit balance of the one person in Role Z and a 50% probability of forfeiting the credit balances of the two persons in Role Y. If a participant's credit balance is forfeited, it is set to 0 thalers and that participant does not get any compensation for the task. In the case that person X lets the random draw decide which of the two options is implemented, Person X learns whether the credit balance of the one person in Role Z or the credit balances of the two persons in Role Y has been forfeited.

In the hypothetical condition, the decider's choice resulted in no monetary consequences for any participant. This was communicated to the participants of the hypothetical condition in bold boxes above and below the description of the decision task (see Appendix A.1). All earned credits were paid out. In the real condition, payments depended on the outcome implemented by the decider, or the computer (Fig. 1). During the experiment, credits were calculated in an experimental currency called thalers. At the end of each session, participants were paid in private using a conversion rate of 1 thaler = €0.10. Participants' payoffs included an amount of €4.00 for showing up to the experiment.

2.2. Results

In the hypothetical condition (see Fig. 2a), the utilitarian choice of sacrificing one credit balance was the most popular (55.8%, CI = 0.413, 0.695), which was consistent with previous literature (Awad et al., 2018). Notwithstanding this, 34.6% (CI = 0.220, 0.491) of the participants chose to randomize. In the real condition (see Fig. 2b), the utilitarian outcome (48%, CI = 0.337, 0.626) and randomization (46%, CI = 0.318, 0.607) were chosen with similar frequencies.

2.3. Discussion

In Study 1, we found that a substantial proportion of participants favored letting a computer randomize between dilemmatic outcomes, even though they could directly implement an unambiguous utilitarian outcome. Randomization was especially appealing when the dilemma entailed real consequences for those harmed that materialized in monetary losses.

Prima facie, we can think of three possible explanations for people's choice to randomize decisions in the situations provoked in Study 1. First, deciders may have wanted to stay willfully ignorant of the consequences their decisions had caused (Dana, 2005, 2006; Grossman & Van Der Weele, 2017). In contrast to making a direct choice (i.e., either forfeiting the credit balance of one person or forfeiting the credit balance of two persons), when they chose to randomize, they might have been able to avoid knowledge about the actual consequences of their decisions. Second, participants' choice to randomize may have been driven by moral considerations emphasizing and respecting the rights of all involved people. From this rather deontological perspective, no participant's credit should be forfeited intentionally (Frankena, 1973). Third, participants may have perceived the choice to randomize as a way of decision avoidance (Meder et al., 2019; Spranca et al., 1991). After all, they might have thought, the coin "decided" who got harmed.

The first explanation was ruled out based on the design of Study 1. Through the instructions of the experiment, participants in the decider role learned before their decision that they would be informed whose money had been forfeited even if they chose to let the computer randomize the outcome. Hence, it was clear to them that they could not stay willfully ignorant of the consequences of their decisions through randomization. In Study 2, we further disentangled the possible drivers for participant behavior and explored whether participants' randomization choices were driven by decision avoidance or moral considerations.

3. Study 2: Moral ratings

In Study 2, we investigated how participants morally evaluated the implemented decision options of Study 1 (including randomization). We expected that if randomization was chosen based on ethical considerations, it would also be rated as more praiseworthy by some participants than its utilitarian alternative. If the decision to randomize was indeed based on moral considerations, it seems reasonable to conclude that a revealed preference for randomization did not reflect decision avoidance.

3.1. Methods and procedures

For Study 2, we invited another 72 students (41.7% female, mean age = 23.56, *SD* = 4.79) from the same laboratory subject pool as in Study 1 to the laboratory for a vignette study. Only participants who were not involved in Study 1 were admitted. The experimental procedures were the same as in Study 1.

For the morality rating task implementation, we followed the approach by Bonnefon et al. (2016). Participants' task was to rate the ethicality of decider choices in Study 1. For this purpose, raters in Study 2 received the original instructions of the real condition in Study 1 and were confronted with the decider's three possible options. Participants then distributed 100 points between three deciders who had each made one of the three possible decisions. More points implied greater perceived ethicality. The moral rating task was described to the participants as follows[3]:

> The experiment described in the previous instructions, has been carried out before. We now show you decisions that were actually made by participants in Role X. Please evaluate how moral you personally consider the decision of each person to be. Please distribute a total of 100 points between the three persons, and assign more points the more morally you evaluate a respective decision. One Person X decided to forfeit the credit balance of the one person in role Z. One Person X decided to forfeit the credit balance of the two persons in role Y. One Person X let a random draw decide which of the two options is realized.

Due to its brevity, the morality rating task was placed after another laboratory experiment that was not related to the current study. In the experimental sessions, participants were paid an income from the preceding experiment and a flat fee of €4.00 for showing up.

3.2. Results

Of the raters, 50% allocated the most points to a randomizing decider, while 45.8% allocated the most points to a utilitarian decider (see Fig. 2c). In terms of average ratings, similar amounts of points were allocated to randomizing deciders (M = 45.1, *SD* = 27.96) and utilitarian deciders (M = 43.6, *SD* = 28.61).

3.3. Discussion

In Study 2, half of the participants rated randomization to be the most moral choice for a decider in Study 1. In ethical terms, raters judged the actions of randomizing deciders as adequate as those of utilitarian deciders on average. It is therefore plausible to assume that deciders' displayed preferences for randomization in Study 1 were driven by moral considerations. Through the almost equal allocation of "morality points" to the randomizing and utilitarian options (leaving the

[3] For participants' complete instructions of Study 2, please refer to Appendix B. Like in Study 1, we used neutral language. Participants were denoted as "persons" in role X (decider), Y (other participants in a pair) and Z (remaining other participant in the group of four).

option of harming the pair with close to zero points), responses in Study 2 also indicated that decision avoidance was not the main driver behind the observed behavioral pattern in Study 1.

4. Study 3: Validity across countries and contexts

Our participants in Studies 1 and 2 were recruited from a large German university. Germany has a Kantian cultural tradition (Bowie, 2003). Kant's deontological ethics puts the individual rights of a person at the center of his philosophy. The second formulation of his categorical imperative postulates that a person should not be treated only as a means to an end, but always at the same time as an end (Kant, 2002). Thus, our first participant samples may have viewed randomization in ethical dilemmas more favorably than people with a less deontological cultural background. Especially in countries with a utilitarian tradition, our results might be challenged.

Therefore, the first aim of Study 3 was to assess whether our previous findings also held in a participant pool with a deeply rooted utilitarian tradition. Choosing an American sample appeared to us a conservative test, because the Anglo-American culture was shaped by the utilitarian philosophy of Bentham and Mill (Driver, 2014).

A second goal of Study 3 was to use the vignette format to assess the prevalence of randomization preferences in applicable AS settings. Participants in Studies 1 and 2 had decided about harming either one or two other persons monetarily. For obvious ethical reasons, we could not actually impose physical harm on people. In the two vignettes of Study 3, however, we took the opportunity to describe settings with harsh physical consequences: an unavoidable autonomous car accident and the allocation of scarce medical resources by an algorithm in a hospital during a staff shortage.

The autonomous car vignette mirrored the decisive characteristics of the conditions in Studies 1 and 2, where participants decided whether to harm one or two other persons, without any further information on those involved in the dilemma. Thus, this first vignette was used to validate our results on randomization preferences in a setting with an unambiguous utilitarian outcome.

In the hospital vignette, we aimed to map the situation, as realistic as possible, of a resource shortage in hospitals in which an algorithm decides which infected patients are cared for. Our approach to creating a vignette that was as realistic as possible was twofold. First, in reality, dilemmas will not be about mere numbers, but entail a huge variety of factors that may weigh into a utilitarian calculus. In the hospital triage vignette, we therefore switched from a dilemma between one or two featureless other persons being harmed to one in which the well-being of two patients with specific features would be weighed because of a resource shortage. Second, we made use of the ongoing COVID-19 pandemic as a topical framing device. We assumed that people could relate more to this scenario because they had likely been familiarized with such cases through mass media coverage (e.g., Kisner, 2021). This vignette therefore did not offer an unambiguous utilitarian alternative to randomization (see below for a detailed description of both vignettes).

We expected that among the U.S. participant pool, randomization would be chosen less frequently compared to the German pool if the vignette encompassed an unambiguous utilitarian alternative (car accident scenario). We also expected that if the vignette encompassed no unambiguous utilitarian alternative and a topical framing (hospital triage scenario), randomization would be chosen as frequently or even more frequently than in the car accident scenario where the utilitarian solution is unambiguous.

4.1. Methods and procedures

In Study 3, we used a convenience sample recruited via PrimePanels on the cloudresearch.com survey plattform. While such panels do not represent the overall population, there is ample evidence that they provide valid results (Goodman et al., 2013; Paolacci & Chandler, 2014). They also avoid some of the limitations of the MTurk platform (commonly used in behavioral science studies during past years), such as self-selection or language problems (Chandler et al., 2019).

In total, we analyzed 761 participant responses to two different decision scenarios, processed online. Participants earned a fixed payment of $1.25 for completing the survey. All instructions were provided online in written form.[4]

In the first vignette, 567 participants read a vignette about the adequate behavior of an autonomous car in an unavoidable accident situation. The first decision scenario was described in the following way:

> Imagine a self-driving car – i.e., a completely autonomous vehicle – is empty and on its way to pick up a passenger. On its way, the car is suddenly – and without any fault on its own – involved in a fatal accident. Decision of the self-driving car: The self-driving car is programmed to make life and death decisions only in those cases in which it is impossible to save everyone. In the situation at hand, the self-driving car detects three uninvolved pedestrians (bystanders) on the two sidewalks next to it. In this situation, it is not possible to save all three people. The car has to decide between two reactions: With one reaction, the car kills with certainty one pedestrian on one of the sidewalks. With the other reaction, the car kills with certainty two pedestrians on the other sidewalk. The car can also let a random draw decide which of the two reactions is realized. The random drawing has a 50% probability of choosing the reaction where the car kills with certainty the one pedestrian on the one sidewalk and a 50% probability of choosing the reaction where the car kills with certainty the two pedestrians on the other sidewalk.

Of all participants who were administered the first scenario, 389 (60.2% female, mean age = 44.11 years, SD = 16.23) answered all attention check items correctly and were included in our data analysis.

For the second vignette of Study 3, we adapted a vignette by Keren and Teigen (2010). The vignette captured the recent COVID-19 pandemic and described the globally observed shortage of hospital staff to care for infected patients (e.g., Buonsenso, De Rose, & Pierantoni, 2021; Kisner, 2021). The second decision scenario was described in the following way:

> Imagine a self-learning software, i.e. a completely autonomous algorithm, is designed to allocate medical resources between hospitals. This algorithm is now confronted with a resource shortage induced by the recent Covid-19 pandemic. Decision of the self-learning algorithm: The self-learning algorithm is programmed to make life and death decisions only in those cases in which it is impossible to save everyone. In the situation at hand, the algorithm detects two undersupplied patients across two hospitals, who both need to be intubated by a specialist, because of their critical condition. The algorithm detects that there is only one specialist in respiratory medicine available, who is free to move to one of the hospitals. In this situation, it is not possible to save both patients. The two patients differ in age and survival probability. The algorithm has to decide between two reactions: With one reaction, the algorithm lets with certainty one patient die in one of the hospitals, who is 50 years old and has a survival probability of 75%. With the other reaction, the algorithm lets with certainty the other patient die in the other hospital, who is 59 years old and has a survival probability of 85%. The algorithm can also let a random draw decide which of the two reactions is realized. The random drawing has a 50% probability of choosing the reaction where the algorithm lets with certainty the one patient die in the one hospital and a 50% probability of choosing the reaction where the algorithm lets with certainty the other patient die in the other hospital.

[4] For participants' instructions of Study 3, please refer to Appendix C.

While Studies 1 and 2, as well as the autonomous driving vignette of Study 3, offered unambiguous utilitarian options of harming fewer people to the benefit of more (see above), the second vignette put the utilitarian option in the eye of the beholder. Some utilitarians might well argue for maximizing expected life years and saving the younger person, while others might argue for maximizing the probability of survival. This was done to capture a characteristic of many real-life moral dilemmas– the absence of a clear-cut utilitarian choice.

The COVID-19 vignette was administered to 532 participants. Of these, 372 (60.5% female, mean age 44.43 years, $SD = 17.09$) answered all attention check items correctly and were included in our data analysis.

4.2. Results

We first report on the participants who responded to the vignette about an autonomous car, which was structurally identical to the decision situation in Studies 1 and 2 in giving deciders the chance to sacrifice one for the benefit of two. When asked what the car should do in a crash, 67.6% (CI = 0.627, 0.722) opted for the utilitarian alternative, i.e., for killing one pedestrian to save two others. A proportion of 30.6% (CI = 0.260, 0.354) participants decided that the car should randomize between outcomes. As expected, the proportion of participants who decided to randomize was lower compared to the real condition of Study 1 ($p = .03$, Chi^2 test). The difference from the hypothetical condition of Study 1 did not reach a conventional level of statistical significance ($p = .56$, Chi^2 test).

Most of the 372 participants (51.1%, CI = 0.459, 0.563) who read the hospital triage vignette decided that the algorithm should randomize. In contrast, 16.9% of the participants (CI = 0.133, 0.211) harmed the older patient who had the higher survival probability and 32.0% (CI = 0.273, 0.370) harmed the younger patient who had the lower survival probability.[5] As expected, the relative frequency of randomizers was greater in this situation with an ambiguous utilitarian choice compared to the situations where a clear-cut utilitarian option was offered ($p = .05$ compared to the two conditions of Study 1; $p < .001$ compared to the autonomous car scenario).

4.3. Discussion

The results of Study 3 indicate that our findings from Studies 1 and 2 may translate into AS programming preferences that may also be applicable to countries with cultures characterized by utilitarian philosophy. As predicted, we observed a slightly lower propensity to randomize among the U.S. sample in the autonomous car vignette. However, the relative frequency of randomizers was still substantial and similar to the hypothetical condition of Study 1. This suggests that we may have elicited *lower* bounds of randomization preferences in the hypothetical condition of Study 1 and in the autonomous car vignette of Study 3 due to the lack of real consequences (hypothetical condition of Study 1) and because of the presence of an unambiguous utilitarian alternative (hypothetical condition of Study 1 and autonomous car vignette of Study 3). These two factors may have eased "cold" calculations (Greene & Haidt, 2002; Lanteri et al., 2008) and encouraged utilitarian reasoning, specifically in the supposedly less deontological U. S. sample.

In the hospital triage scenario of Study 3, a majority of U.S. participants expressed a preference for randomization. Randomization was as popular here as it was in the condition that evoked real monetary consequences studied in a German laboratory (real condition of Study 1). With its focus on patients' personal features and its higher topicality, the COVID-19 framing may have felt less hypothetical to respondents. On the other hand, it might have been more difficult to evaluate patients' relative personal features than to compare structural features such as few people against many.

5. General discussion

We started with the observation that contrary to humans, ASs feature an appealing alternative to actively imposing harm on some to avoid greater harm for others: randomizing decisions with dilemmatic outcomes. Contrary to humans, who would have to go through an accredited public process to prove that a decision was unbiased and impartial (e.g., by getting certification from a notary public), ASs can make immediate randomized decisions that are traceable and verifiable (e.g., through distributed ledgers). We argued that for randomization to be applicable in democracies, it must correspond to people's moral intuition. Using the results from three experimental studies, we have presented evidence showing that many people are willing to leave outcomes of ethical dilemmas to chance, despite the presence of utilitarian alternatives. Across studies, we found a substantial proportion of 41.4% of participants who preferred the randomization choice. With our experimental design, we ruled out randomization being chosen because participants wanted to stay willfully ignorant of the consequences of their decisions. In Study 2, we further showed that participants' choices to randomize were driven by moral considerations and did not reflect decision avoidance. Our results from Study 3 indicated that our previous findings on randomization preferences may be transferable and applicable to different countries and AS contexts, even when countries have a distinct utilitarian tradition. Finally, Study 3 also conveyed that if a moral dilemma does not encompass an unambiguous utilitarian alternative, randomization is chosen more frequently.

Our findings from the real condition of Study 1 indicate that randomization preferences are stronger in situations in which people face actual ethical dilemmas. A potential explanation for this result is that hypothetical and rather abstract cases put participants in a "cold" mental state where ethical deliberation is easier, while materialized and realistic scenarios may trigger a "hot" state of stronger emotional involvement and more intuitive reactions. Previous research has found that people lean more toward utilitarian reasoning if they are less emotional and can reason more abstractly (Greene & Haidt, 2002; Lanteri et al., 2008). Support for this mechanism may also come from the responses to the hospital triage vignette in Study 3. There, the described resource shortage might have triggered stronger emotions due to personal involvement in the current COVID-19 pandemic and fuelled randomization preferences beyond the effect of a lack of a clear utilitarian solution alone. Further research is needed to disentangle these effects.

To address a potential limitation of the real condition in Study 1, it has sometimes been noted that participants may regard the amounts of money paid for decisions made in laboratory experiments as trivial. The general effects of varying stake size are mixed and seem to depend on concrete experimental contexts (Camerer & Hogarth, 1999; Karagözoğlu & Urhan, 2017). However, ample evidence conveys that stake effects are nonexistent or negligible in distribution tasks comparable to ours, even in situations when the stakes are very small (e.g., Amir et al., 2012; Fehr, Fischbacher, & Tougareva, 2002; Kocher et al., 2008; Larney et al., 2019). Nevertheless, we decided to make sure that the deciders of Study 1 would assign importance to their decisions and be paid accordingly. On average, a forfeited credit was comparable to an hour's wage for a student assistant in Germany. Of course, no matter how high the monetary stakes, we could not model the stakes of a fatal car accident in a lab experiment. However, through incentivized experiments, we can gain a

[5] A total of 37 participants included in our data analysis of the self-learning algorithm case reported having had prior personal experience with a family member or friend falling ill with Covid-19. This did not affect their preferences for randomization. Of the 37 participants, 51.4% (CI = 0.353, 0.675) decided the algorithm should randomize, 24.3% (CI = 0.105, 0.381) saved the younger patient, and 24.3% (CI = 0.105, 0.381) saved the patient with the higher survival probability.

structural understanding of how people deal with the described dilemmas on a small scale by inducing real (monetary) consequences on others. Testing this class of dilemma situations in the field is notoriously difficult (Falk & Heckman, 2009). Furthermore, we are among the first to examine randomization attitudes in ethical dilemmas systematically, including a fully incentive-compatible approach. This contrasts with many empirical studies on ethical dilemmas that rely solely on self-reports (Awad et al., 2018; Bigman & Gray, 2020; Keren & Teigen, 2010). For future studies, we encourage conducting more incentivized experiments and using complementary methods to reduce bias.

A conceivable objection to the interpretation of our results is that randomizers might have expressed a preference for risk taking (e.g., Bromiley & Curley, 1992). Previous literature suggests that individual risk attitudes generalize to decision making for others (e.g., Stone, Yates, & Caruthers, 2002). We did not explicitly test for the influence of risk attitudes in our setting. However, as demonstrated in Study 2, participants' choice of randomization was likely driven by moral considerations. Moreover, we found that across all experiments involving a randomization decision, choosing randomization was consistently and significantly positively correlated with participants' age ($r = 0.14$ for pooled Study 1 and Study 3 data; $r = 0.28$ for Study 1 data, $r = 0.25$ for Study 3 data; all $p < .001$, point-biserial correlation).[6] This also supports our argument that choosing a randomization option does not represent a preference for risk-taking or gambling, as older people are robustly found to be more risk-averse (Dohmen et al., 2011; Vroom & Pahl, 1971).

The positive association between randomization preferences and age might also be informative regarding randomizing participants' ethical views. Previous research suggests that older adults make more deontological moral judgments compared to younger adults (e.g., McNair, Okan, Hadjichristidis, & de Bruin, 2019) and have higher universalism values (Robinson, 2013). Values serve as standards that guide the evaluation of events, behaviors, and persons. The defining goal of universalism is to care for the welfare of all people, suggesting that universalism values may be the most relevant basic values in the type of anonymous decision situations currently under investigation (Lönnqvist, Verkasalo, Wichardt, & Walkowitz, 2013; Schwartz, 1992). Like deontologists, universalists might put individual rights at the center of their ethical reasoning. This implies that no individual should be harmed intentionally, even for the greater good. Future studies are needed to explore how individual deontological views and universalism values relate to randomization preferences.

Another limitation of our paper is that we only compared data from German and U.S. samples. As described earlier, we considered our work a starting point for the comprehensive study of randomization preferences in ethical dilemmas of AS programming. For this purpose, we intentionally chose German and U.S. samples, because Germany has a Kantian cultural tradition whereas the Anglo-American culture was shaped by utilitarian philosophy. Our choice appeared to us to be a conservative test for a transnational comparison of randomization preferences. For future studies, it would be interesting to investigate how randomization is perceived in non-Western cultures with different philosophical heritage, such as China's.

Another interesting question for further investigation is whether people also embrace randomization as a possible solution if they are themselves involved in the dilemmatic situation. As former studies have shown, the utilitarian welfare-maximizing choice is well accepted as long as one does not have to make personal sacrifices for the greater good. If a situation is framed in a way that study participants see their own security in opposition to the utilitarian choice, the utilitarian choice receives far less approval (Bonnefon et al., 2016; Frank et al., 2019). It is conceivable that in this situation, people may be more willing to agree to a randomization procedure (if available) than a direct sacrifice on their part. Anecdotal evidence from Pennsylvanian hospitals in the U.S., where remdesivir (a drug for the emergency treatment of COVID-19 infected patients) was administered to patients using a lottery procedure, conveys that patients who did not eventually receive the drug were still supportive of the randomization-based allocation system (Kolata, 2020).

All in all, we found notable support for the prevalence of randomization preferences in ethical dilemmas. Next to utilitarian preferences, randomization preferences exist to an extent that validates considering randomization alongside utilitarian arguments when programming ASs' behavior. Randomization offers a non-discriminatory alternative to a utilitarian approach. Even utilitarians might find randomization appealing in ethical dilemmas concerning AS , if it increases the chances of a welfare-enhancing technology being implemented. It would certainly be in the spirit of utilitarianism to pragmatically embrace randomization for exceptionally rare dilemmatic cases when deploying a technology that generally benefits society, if this would erode the deontological resistance to the technology. We therefore suggest including randomization in the crucial societal debate on AS programming (Rahwan et al., 2019).

Funding

This work was supported by the Zentrum Digitalisierung.Bayern (ZD.B).

Acknowledgements

The authors gratefully acknowledge support from the Zentrum Digitalisierung.Bayern. Anja Bodenschatz also acknowledges support by the Joachim Herz Stiftung. Gari Walkowitz prepared the article also within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE), Moscow, Russian Federation. The authors thank Friedrich Gehring and Brian Cooper for their assistance.

Appendix A

A.1. Study 1: Participant Instructions for Hypothetical Condition (Translation, Original in German)

Instructions for the experiment

You are now participating in a decision experiment. Please read the instructions for the experiment carefully and completely. It is possible that questions you may have while reading will be clear after you have read the complete instructions. For the entire duration of the experiment, it is very important that you do not communicate with other experiment participants. In addition, your mobile phones must be switched off and stowed away. Violations lead to the termination of the experiment without compensation of the participants. If you have problems with understanding something, please reread these instructions first. If you still have questions, please raise your hand. We will then come to your cubicle and answer your questions

[6] We do not find such correlations for participants' sex.

in person.

You will receive an initial expense allowance of €4.00, as a show-up fee. Over the course of the experiment you can earn additional money. The amount of your earnings depends on your decisions or on the decisions of other participants. You will not learn the identity of the other participants at any time. Similarly, your identity will never be revealed to the other participants.

All data and answers are evaluated anonymously.

Today's experiment

This experiment consists of two parts. At the beginning of the experiment you will be assigned a participant identifier. The identifier is randomly generated from a letter and a number. Each combination is assigned only once.

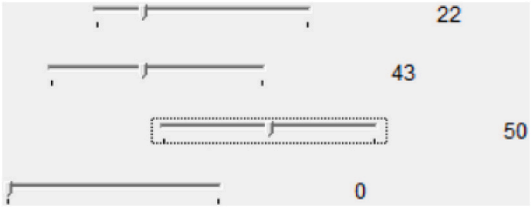
In the first part of the experiment you work on a task. The second part consists of a decision situation. Either you make a decision yourself or you are affected by the decision of another participant. Both parts of the experiment are described in more detail below.

During the experiment, all credit balances are given in thalers. For the payout at the end of the experiment the following applies: 1 thaler = €0.10. A questionnaire follows at the end of the experiment.

Part 1 - The work task

The first part of the experiment consists of a work task that you perform for 10 min. By carrying out the task, you generate a credit balance.

Description of the task. During the task, you will see sliders on your screen (see figure below). Each slider can be set from position 0 to 100. You can drag and drop the slider by clicking on it with the mouse. Your task is to position as many sliders as possible exactly to the value of 50. You have 10 minutes to work on this task. If you have correctly adjusted all sliders on a screen to position 50, a "next"-button is shown at the bottom right of the monitor screen. By clicking this button, you are directly led to the next screen with further sliders that you can adjust.



Generated credit balance. After the 10 minutes have elapsed, the work task is automatically terminated. The number of sliders that you have correctly set to position 50 during the 10 minutes of the task will be shown to you on the screen. In addition, the credit balance that you have generated through your work is displayed. For each correctly positioned slider, 1 thaler will be credited to you.

Part 2 - The hypothetical decision situation

Please note: The following part of the experiment has no real consequences. The described situation is hypothetical. In the following, please imagine you had received the following instructions and had been assigned to Role X.

Role assignment. There are three roles in this part of the experiment: X, Y and Z. One quarter of the participants present are randomly assigned the Role X. After that, three participants from the remaining group of participants are randomly assigned to each person in Role X. Two of these participants are randomly assigned to Role Y and one participant to Role Z.

All participants are informed whether they are operating in the role of Person X or not. Participants in the roles of Person Y or Z are not informed about their exact role, but only about the fact that they are not operating in the role of Person X.

Decision of Person X. Person X makes a decision about the credit balance that the persons assigned to him or her have generated in the first part of the experiment. However, the exact amounts of the credit balances of the persons in Role Y and Z are never revealed to Person X.

Person X will see the participant identifiers of the three participants assigned to him or her on the decision screen. Person X then decides between two options with a mouse click:

- With one option Person X forfeits the credit balance of the one person in Role Z.
- With the other option, Person X forfeits the credit balances of the two persons in Role Y.
- Person X can also let a random draw decide which of the two options is realized.

Random drawing has a 50% probability of forfeiting the credit balance of the one person in Role Z and a 50% probability of forfeiting the credit balance of the two persons in Role Y.

If a participant's credit balance is forfeited, it is set to 0 thalers and that participant does not get any compensation for the task.

Information after the decision. In the case that Person X lets the random draw decide which of the two options is implemented, Person X learns whether the credit balance of the one person in Role Z or the credit balances of the two persons in Role Y has been forfeited.

The hypothetical payout

Person X. Persons in Role X always receive the credit balance that they earned in the first part of the experiment.

Person Y and Z. At the end of the experiment, participants in the Roles of Y and Z learn whether their credit balance from the first part of the experiment has been forfeited during the second part of the experiment. In this case, they will leave the experiment empty-handed. If their credit balance has not been forfeited, it will be paid out.

They also learn whether the Person X to whom they were assigned, let the random draw decide as to whose credit balance would be forfeited or not.

Please remember: All participants of today's experiment decide what they would do in the role of Person X. No one in this room is actually assigned the Roles Y or Z. No credit balance is actually forfeited. Every participant receives the credit balance earned in the first part of the experiment.

A.2. Study 1: Example Decision Screen of Hypothetical Condition (Translation, Original in German)

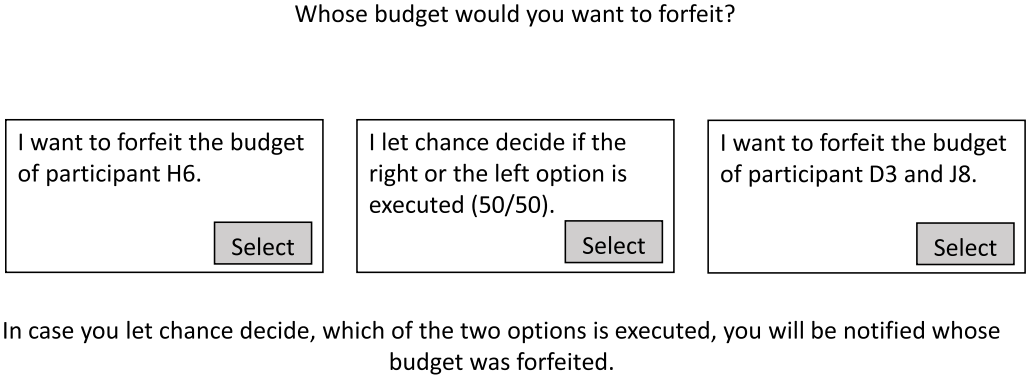


Fig. A1. Example Decision Screen of the *hypothetical* condition in Study 1. Left and right button were randomized in order. Button for randomization between the two options in the middle. All participants were assigned an individual identifier at the beginning of the experiment, consisting of one letter and one digit, e.g., "F4". Participants were always addressed by their identifier during experiment. This was to make sure that participants in the role of the decider were aware that every identifier on their decision screen depicted one anonymous other participant. In the decision screen for the *real* condition of Study 1 the question read "Whose budget do you want to forfeit?".

A.3. Study 1: Participant Instructions for Real Condition (Translation, Original in German)

Instructions for the experiment

You are now participating in a decision experiment. Please read the instructions for the experiment carefully and completely. It is possible that questions you may have while reading will be clear after you have read the complete instructions. For the entire duration of the experiment, it is very important that you do not communicate with other experiment participants. In addition, your mobile phones must be switched off and stowed away. Violations lead to the termination of the experiment without compensation of the participants. If you have problems with understanding something, please reread these instructions first. If you still have questions, please raise your hand. We will then come to your cubicle and answer your questions in person.

You will receive an initial expense allowance of €4.00, as a show-up fee. Over the course of the experiment you can earn additional money. The amount of your earnings depends on your decisions or on the decisions of other participants. You will not learn the identity of the other participants at any time. Similarly, your identity will never be revealed to the other participants.

All data and answers are evaluated anonymously.

Today's experiment

This experiment consists of two parts. At the beginning of the experiment you will be assigned a participant identifier. The identifier is randomly generated from a letter and a number. Each combination is assigned only once.

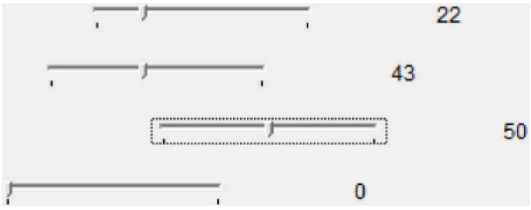
In the first part of the experiment you work on a task. The second part consists of a decision situation. Either you make a decision yourself or you are affected by the decision of another participant. Both parts of the experiment are described in more detail below.

During the experiment all credit balances are given in thalers. For the payout at the end of the experiment the following applies: 1 thaler = €0.10. A questionnaire follows at the end of the experiment.

Part 1 - The work task

The first part of the experiment consists of a work task that you perform for 10 min. By carrying out the task, you generate a credit balance.

Description of the task. During the task, you will see sliders on your screen (see figure below). Each slider can be set from position 0 to 100. You can drag and drop the slider by clicking on it with the mouse. Your task is to position as many sliders as possible exactly to the value of 50. You have 10 minutes to work on this task. If you have correctly adjusted all sliders on a screen to position 50, a "next"-button is shown at the bottom right of the monitor screen. By clicking this button, you are directly led to the next screen with further sliders that you can adjust.



Generated credit balance. After the 10 minutes have elapsed, the work task is automatically terminated. The number of sliders that you have correctly set to position 50 during the 10 minutes of the task will be shown to you on the screen. In addition, the credit balance that you have generated through your work is displayed. For each correctly positioned slider, 1 thaler will be credited to you.

Part 2 - The decision situation

Role assignment. There are three roles in this part of the experiment: X, Y and Z. One quarter of the participants present are randomly assigned the Role X. After that, three participants from the remaining group of participants are randomly assigned to each person in Role X. Two of these participants are randomly assigned to Role Y and one participant to Role Z.

All participants are informed whether they are operating in the role of Person X or not. Participants in the roles of Person Y or Z are not informed about their exact role, but only about the fact that they are not operating in the role of Person X.

Decision of Person X. Person X makes a decision about the credit balance that the persons assigned to him or her have generated in the first part of the experiment. However, the exact amounts of the credit balances of the persons in Role Y and Z are never revealed to Person X.

Person X will see the participant identifiers of the three participants assigned to him or her on the decision screen. Person X then decides between two options with a mouse click:

- With one option Person X forfeits the credit balance of the one person in Role Z.
- With the other option, Person X forfeits the credit balances of the two persons in Role Y.
- Person X can also let a random draw decide which of the two options is realized.

Random drawing has a 50% probability of forfeiting the credit balance of the one person in Role Z and a 50% probability of forfeiting the credit balance of the two persons in Role Y.

If a participant's credit balance is forfeited, it is set to 0 thalers and that participant does not get any compensation for the task.

Information after the decision. In the case that Person X lets the random draw decide which of the two options is implemented, Person X learns whether the credit balance of the one person in Role Z or the credit balances of the two persons in Role Y has been forfeited.

The payout

Person X. Persons in Role X always receive the credit balance that they earned in the first part of the experiment.

Person Y and Z. At the end of the experiment, participants in the Roles of Y and Z learn whether their credit balance from the first part of the experiment has been forfeited during the second part of the experiment. In this case, they will leave the experiment empty-handed. If their credit balance has not been forfeited, it will be paid out.

They also learn whether the Person X to whom they were assigned, let the random draw decide as to whose credit balance would be forfeited or not.

A.4. Study 1: Example Decision Screen of Real Condition (Translation, Original in German)

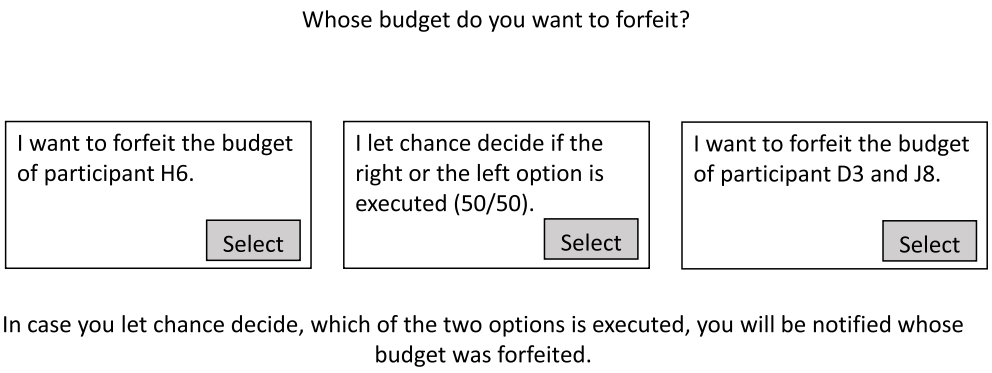


Fig. A2. Example Decision Screen of the *real* condition in Study 1. Left and right button were randomized in order. Button for randomization between the two options in the middle. All participants were assigned an individual identifier at the beginning of the experiment, consisting of one letter and one digit, e.g. "F4". Participants were always addressed by their identifier during experiment. This was to make sure that participants in the role of the decider were aware that every identifier on their decision screen depicted one anonymous other participant. In the decision screen for the *hypothetical* condition of Study 1 the question read "Whose budget would you want to forfeit?".

Appendix B

B.1. Study 2: Participant Instructions (Translation, Original in German)

Instructions for the experiment

You will now read the instructions of an experiment, in which you yourself will not take part. Nevertheless, please read the instructions carefully and completely. We will ask you questions concerning this experiment, after you have informed yourself about the experimental proceedings.

The experiment, described in the instructions, has been carried out before in this laboratory. People participated in this experiment and actually had to make the decision described or bear the consequences of decisions made by another participant.

B.2. Study 2: Example Decision Screen (Translation, Original in German)

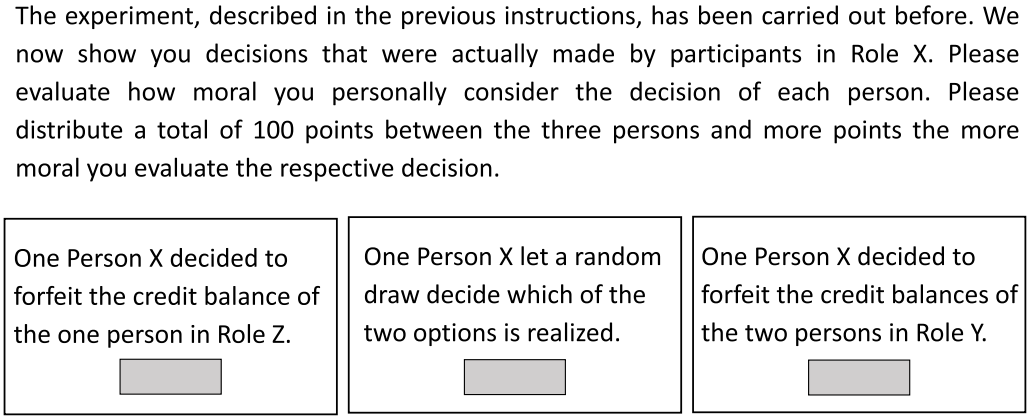


Fig. B1. Example Decision Screen presented to raters in Study 2. Left and right box were randomized in order. Box for randomization between the two options in the middle.

Appendix C

C.1. Participant Instructions for Self-Driving Car Case

Situation

Imagine a self-driving car – i.e., a completely autonomous vehicle – is empty and on its way to pick up a passenger. On its way, the car is suddenly – and without any fault on its own – involved in a fatal accident.

Decision of the self-driving car. The self-driving car is programmed to make life and death decisions only in those cases in which it is impossible to save everyone. In the situation at hand, the self-driving car detects three uninvolved pedestrians (bystanders) on the two sidewalks next to it. In this situation, it is not possible to save all three people. The car has to decide between two reactions:

- With one reaction, the car kills with certainty one pedestrian on one of the sidewalks.
- With the other reaction, the car kills with certainty two pedestrians on the other sidewalk.

The car can also let a random draw decide which of the two reactions is realized. The random drawing has a 50% probability of choosing the reaction where the car kills with certainty the one pedestrian on the one sidewalk and a 50% probability of choosing the reaction where the car kills with certainty the two pedestrians on the other sidewalk.

C.2. *Decision* ***Screen of Self-Driving Car Case*** *– After they answered two comprehension questions, participants saw the following decision screen:*

In your opinion, which pedestrian/s should the car kill?

- ○ The car should kill the one pedestrian on the one sidewalk with certainty to save the two.
- ○ The car should kill the two pedestrians on the other sidewalk with certainty to save the one.
- ○ The car should let a random draw decide which of the two reactions is realized.

Fig. C1. Example Decision Screen presented to participants in Study 3. Please note: Order of the two affected parties in the vignette was randomized. Participants saw one of two versions of the vignette and decision screen. One half of participants read the text as depicted above and the other half read a version in which the two pedestrians are mentioned first and are described to be on the one sidewalk.

C.3. Participant Instructions for Hospital Case

Situation

Imagine a self-learning software, i.e. a completely autonomous algorithm, is designed to allocate medical resources between hospitals. This algorithm is now confronted with a resource shortage induced by the recent Covid-19 pandemic.

Decision of the self-learning algorithm

The self-learning algorithm is programmed to make life and death decisions only in those cases in which it is impossible to save everyone. In the situation at hand, the algorithm detects two undersupplied patients across two hospitals, who both need to be intubated by a specialist, because of their critical condition. The algorithm detects that there is only one specialist in respiratory medicine available, who is free to move to one of the hospitals. In this situation, it is not possible to save both patients. The two patients differ in age and survival probability. The algorithm has to decide between two reactions:

- With one reaction, the algorithm lets with certainty one patient die in one of the hospitals, who is 50 years old and has a survival probability of 75%.
- With the other reaction, the algorithm lets with certainty the other patient die in the other hospital, who is 59 years old and has a survival probability of 85%.

The algorithm can also let a random draw decide which of the two reactions is realized. The random drawing has a 50% probability of choosing the reaction where the algorithm lets with certainty the one patient die in the one hospital and a 50% probability of choosing the reaction where the algorithm lets with certainty the other patient die in the other hospital.

C.4. Decision Screen of Hospital Case *– After they answered two comprehension questions, participants saw the following decision screen:*

In your opinion, which patient should the algorithm let die?

- ○ The algorithm should let the one patient in the one hospital die with certainty and try to save the other patient in the other hospital, who is 9 years older, but has a 10 percentage points higher survival probability.
- ○ The algorithm should let the other patient in the other hospital die with certainty and try to save the one patient, who is 9 years younger, but has a 10 percentage points lower survival probability.
- ○ The algorithm should let a random draw decide which of the two reactions is realized.

Fig. C2. Example Decision Screen presented to participants in Study 3. Please note: Order of the two affected parties in the vignette was randomized. Participants saw one of two versions of the vignette and decision screen. One half of participants read the text as depicted above and the other half read a version in which the patient first mentioned is the 59 years old with a survival probability of 85%.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

American College of Surgeons. (2019 October 29). *AI outperforms clinicians' judgment in triaging postoperative patients for intensive care*. Science Daily. www.sciencedaily.com/releases/2019/10/191029182456.htm.

Amir, O., Rand, D. G., & Gal, Y. A. K. (2012). Economic games on the internet: The effect of $1 stakes. *PLoS One, 7*(2), Article e31461.

Awad, E., Anderson, M., Anderson, S. L., & Liao, B. (2020). An approach for combining ethical principles with public opinion to guide public policy. *Artificial Intelligence, 287*, 1–37.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64.

Bains, W. (2008). Random number generation and creativity. *Medical Hypotheses, 70*(1), 186–190.

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrite: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology, 77*(3), 525.

Bench-Capon, T. J. (2020). Ethical approaches and autonomous systems. *Artificial Intelligence, 281*, 1–28.

Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., & Stephan, A. (2018). Autonomous vehicles require socio-political acceptance—an empirical and philosophical perspective on the problem of moral decision making. *Frontiers in Behavioral Neuroscience, 12*, 31.

Berman, J. Z., & Kupor, D. (2020). Moral choice when harming is unavoidable. *Psychological Science, 31*(10), 1294–1301.

Bigman, Y. E., & Gray, K. (2020). Life and death decisions of autonomous vehicles. *Nature, 579*(7797), E1–E2.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576.

Bowie, A. (2003). *Introduction to German philosophy: From kant to habermas*. Cambridge: UK Polity Press.

Bromiley, P., & Curley, S. P. (1992). Individual differences in risk taking. In J. F. Yates (Ed.), *Risk-taking behavior* (pp. 87–132). John Wiley & Sons.

Broome, J. (1991). Utility. *Economics and Philosophy, 7*(1), 1–12.

Buonsenso, D., De Rose, C., & Pierantoni, L. (2021). Doctors' shortage in adults COVID-19 units: A call for pediatricians. *European Journal of Pediatrics, 180*, 2315–2318.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19*(1), 7–42.

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods, 51*(5), 2022–2038.

Dana, J. (2005). Conflicts of interest and strategic ignorance of harm. In *Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy*. Cambridge: Cambridge University Press.

Dana, J. (2006). Strategic ignorance and ethical behavior in organizations. In *Ethics in groups*. Bingley, United Kingdom: Emerald Group Publishing Limited.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association, 9*(3), 522–550.

Doucet, M. (2013). Playing dice with morality: Weighted lotteries and the number problem. *Utilitas, 25*(2), 161–181.

Driver, J. (2014). The history of utilitarianism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/utilitarianism-history/.

Ethics Commission. (2017, June). *Automated and connected driving*. Technical report. Federal Ministry of Transport and Digital Infrastructure https://www.bmvi.de/SharedDocs/EN/PressRelease/2017/084-ethic-commission-report-automated-driving.html.

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science, 326*(5952), 535–538.

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., Stephan, A., Pipa, G., & König, P. (2019). 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), 399–418.

Fehr, E., Fischbacher, U., & Tougareva, E. (2002). *Do high stakes and competition undermine fairness? Evidence from Russia*. Institute for Empirical Research in Economics, University of Zurich - Working Paper Series. Working Paper (July 2002).

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics, 10*, 171–178.

Frank, D. A., Chrysochou, P., Mitkidis, P., & Ariely, D. (2019). Human decision-making biases in the moral dilemmas of autonomous vehicles. *Scientific Reports, 9*(1), 1–19.

Frankena, W. K. (1973). *Ethics*. New Jersey: Prentice Hall.

Gao, P., Hensley, R., & Zielke, A. (2014 October 1). *A road map to the future for the auto industry*. McKinsey Quarterly. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/a-road-map-to-the-future-for-the-auto-industry#.

Gill, D., & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *The American Economic Review, 102*(1), 469–503.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213–224.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6*(12), 517–523.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association, 1*(1), 114–125.

Grossman, Z., & Van Der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association, 15*(1), 173–217.

Hancock, P. A., Nourbakhsh, I., & Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. *Proceedings of the National Academy of Sciences, 116*(16), 7684–7691.

Hao, H. (2020 April 23). *Doctors are using AI to triage Covid-19 patients*. The tools may be here to stay. *MIT Technology Review* https://www.technologyreview.com/2020/04/23/1000410/ai-triage-covid-19-patients-health-care/.

Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Pipa, G., Stephan, A., & König, P. (2019). Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives. *Frontiers in Psychology, 10*, 2415.

Kant, I. (2002). *Critique of practical reason*. Indianapolis: Hackett Publishing.

Karagözoğlu, E., & Urhan, Ü. B. (2017). The effect of stake size in experimental bargaining and distribution games: A survey. *Group Decision and Negotiation, 26*(2), 285–325.

Keren, G., & Teigen, K. H. (2010). Decisions by coin toss: Inappropriate but fair. *Judgment and Decision Making, 5*(2), 83–101.

Kisner, J. (2021 December 8). What the chaos in hospitals is doing to doctors. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2021/01/covid-ethics-committee/617261/.

Kocher, M. G., Martinsson, P., & Visser, M. (2008). Does stake size matter for cooperation and punishment? *Economics Letters, 99*(3), 508–511.

Kolata, G. (2020 July 23). *Who gets the covid-19 vaccine first? Here's one idea*. The New York Times. https://www.nytimes.com/2020/07/23/health/coronavirus-vaccine-allocation.html.

Lanteri, A., Chelini, C., & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics, 83*(4), 789–804.

Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes, 151*, 61–72.

Lönnqvist, J. E., Irlenbusch, B., & Walkowitz, G. (2014). Moral hypocrisy: Impression management or self-deception? *Journal of Experimental Social Psychology, 55*, 53–62.

Lönnqvist, J. E., Rilke, R. M., & Walkowitz, G. (2015). On why hypocrisy thrives: Reasonable doubt created by moral posturing can deter punishment. *Journal of Experimental Social Psychology, 59*, 139–145.

Lönnqvist, J. E., Verkasalo, M., Wichardt, P. C., & Walkowitz, G. (2013). Personal values and prosocial behaviour in strategic interactions: Distinguishing value-expressive from value-ambivalent behaviours. *European Journal of Social Psychology, 43*(6), 554–569.

Mackie, J. L. (1990). *Ethics: Inventing right and wrong* (Reprint Edition). Penguin.

McNair, S., Okan, Y., Hadjichristidis, C., & de Bruin, W. B. (2019). Age differences in moral judgment: Older adults are more deontological than younger adults. *Journal of Behavioral Decision Making, 32*(1), 47–60.

Meder, B., Fleischhut, N., Krumnau, N. C., & Waldmann, M. R. (2019). How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty. *Risk Analysis, 39*(2), 295–314.

Mootz, F. J., III (2009). In *On philosophy in American law*. Cambridge: Cambridge University Press.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184–188.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., & Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477–486.

Robinson, O. C. (2013). Values and adult age: Findings from two cohorts of the European Social Survey. *European Journal of Ageing, 10*(1), 11–23.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theory and empirical tests in 20 countries. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–65). New York: Academic Press.

Singer, P. (2011). *Practical ethics*. Cambridge: Cambridge University Press.

Smart, J. J. C., & Williams, B. (1973). *Utilitarianism: For and against*. Cambridge: Cambridge University Press.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*(1), 76–105.

Stone, E. R., Yates, A. J., & Caruthers, A. S. (2002). Risk taking in decision making for others versus the self[1]. *Journal of Applied Social Psychology, 32*(9), 1797–1824.

Timmerman, J. (2004). The individualist lottery: How people count, but not their numbers. *Analysis, 64*(2), 106–112.

Truog, R. D., Mitchell, C., & Daley, G. Q. (2020). The toughest triage – allocating ventilators in a pandemic. *New England Journal of Medicine, 382*(21), 1973–1975.

Vroom, V. H., & Pahl, B. (1971). Relationship between age and risk taking among managers. *Journal of Applied Psychology, 55*(5), 399–405.