Received: 26 July 2022 | Accepted: 7 August 2023

DOI: 10.1111/bjop.12679

ORIGINAL ARTICLE

the british psychological society
promoting excellence in psychology

Moral reasoning behind the veil of ignorance: An investigation into perspective-taking accessibility in the context of autonomous vehicles

Giovanni Bruno[1,2] | Andrea Spoto[1,2] | Michela Sarlo[3] | Lorella Lotto[4] | Alex Marson[1] | Nicola Cellini[1,2] | Simone Cutini[4,5]

[1]Department of General Psychology, University of Padua, Padua, Italy

[2]Mobility and Behavior Center, University of Padua, Padua, Italy

[3]Department of Communication Sciences, Humanities and International Studies, University of Urbino Carlo Bo, Urbino, Italy

[4]Department of Developmental Psychology and Socialization, University of Padua, Padua, Italy

[5]Padova Neuroscience Center, University of Padua, Padua, Italy

Correspondence
Giovanni Bruno, Department of General Psychology, University of Padua, Via Venezia 8, Padova 35131, Italy.
Email: giovanni.bruno.2@unipd.it

Abstract

Perspective-taking (PT) accessibility has been recognized as an important factor in affecting moral reasoning, also playing a non-trivial role in moral investigation towards autonomous vehicles (AVs). A new proposal to deepen this effect leverages the principles of the veil of ignorance (VOI), as a moral reasoning device aimed to control self-interested decisions by limiting the access to specific perspectives and to potentially biased information. Throughout two studies, we deepen the role of VOI reasoning in the moral perception of AVs, disclosing personal and contingent information progressively throughout the experiment. With the use of the moral trilemma paradigm, two different VOI conditions were operationalized, inspired by the Original Position theory by John Rawls and the Equiprobability Model by John Harsanyi. Evidence suggests a significant role of VOI reasoning in affecting moral reasoning, which seems not independent from the order in which information is revealed. Coherently, a detrimental effect of self-involvement on utilitarian behaviours was detected. These results highlight the importance of considering PT accessibility and self-involvement when investigating moral attitudes towards AVs, since it can help the intelligibility of general concerns and hesitations towards this new technology.

KEYWORDS

autonomous vehicles, ethics, moral dilemma, perspective-taking accessibility, utilitarian behaviour, veil of ignorance

BACKGROUND

In the process of making a critical decision, individuals – as rational agents – may struggle to admit the reduction of personal benefits in the name of collective advantage (e.g. Hardin, 1968; Impett & Gordon, 2008), and, when it comes to life and death situations, the concern for the greater collective goal typically led to intricate moral issues (Conway et al., 2018; Kahane et al., 2018). A clear example of this conflict is showcased in the intense debate on the acceptance of autonomous vehicles (AVs), which is fuelling the investigation of general attitudes towards this new AI technology (Awad et al., 2018; Fagnant & Kockelman, 2015; Jobin et al., 2019). Fully AVs (i.e. self-driving cars) are defined as vehicles in which driving operations occur without the direct involvement of a human driver, as the vehicle is completely capable to control its manoeuvres operating in a self-driving mode (National Highway Traffic Safety Administration, 2013). An emblematic example of this moral conflict has been described by Bonnefon et al. (2016) as the 'social dilemma of autonomous vehicles' (i.e. the 'AV dilemma'): in the unlucky event of a road accident with unavoidable death(s), people agree that aiming for the protection of the highest number of people would be the fairest choice for the collectivity, regardless of their traffic roles (AV passengers or pedestrians). Nonetheless, they would be willing to purchase an AV that prioritizes their personal safety and protection before anyone else's. The result is the definition of a fundamental psychological roadblocks to the implementation and the adoption of AVs (Shariff et al., 2017), that pairs with a general low level of trust and acceptance of this technology (Bergmann et al., 2018; Kaur & Rampersad, 2018; Zhang et al., 2019), with the demand of high standards of safety (Shariff et al., 2021), and with the need to overcome important technical, economical, and normative barriers (Fagnant & Kockelman, 2015; Malle et al., 2015). Clearly, the AV dilemma also raises an issue of moral nature, in which the utilitarian moral code is challenged by contrasting non-utilitarian and self-protective tendencies.

The utilitarian moral code[1] (Bentham, 1781/1996; Mill, 1861/2004) is the reference point in the investigation of moral reasoning since the definition of the traditional Switch (Foot, 1967) and Push (Thomson, 1985) dilemmas. Traditionally (Bruno et al., 2022; Greene et al., 2001; Lotto et al., 2014; Moore et al., 2008), moral dilemmas compare a utilitarian resolution for pursuing the greater good with an opposite non-utilitarian behaviour, often referred to the duty-based deontological moral code[2] (Kant, 1785), but generally representing an opposite alternative to utilitarianism. These dilemmas are traditionally sacrificial (Bartels & Pizarro, 2011) since the moral agent has to choose between two behaviours with the inevitable consequence of at least a human life loss. In some cases (e.g. the AV dilemma), the moral agent is framed as one of the potential victims (i.e. self-sacrificial dilemma), requiring a decision that may also lead to personal sacrifice. Interestingly, self-sacrificial dilemmas in the traditional sets (Greene et al., 2001; Moore et al., 2008) typically jeopardize the moral agent's life in the non-utilitarian alternative (e.g. 'Should you kill this person to save yourself and other five people?'; *The burning building dilemma*; Moore et al., 2008). This appears to be a fundamental difference with self-involvement AV dilemmas ('should you sacrifice yourself to save five pedestrians?', Awad et al., 2018; Bonnefon et al., 2016), where the moral agent plays the role of the AV passenger and personal sacrifice is required for pursuing the utilitarian outcome. Expectably, the self-sacrifice framing factor has been evidenced as fundamental in boosting the endorsement of utilitarian behaviour in moral dilemmas concerning AVs (Bruno et al., 2023). Interestingly, a non-dichotomous version of the self-sacrificial dilemma has been proposed and discussed (the "moral trilemma", Di Nucci, 2013; Thomson, 2008), delving into the permissibility to divert the trolley in the Switch dilemma if the moral agent is not willing to

[1]In normative ethics, an ethical theory according to which an action is right if it tends to produce positive effects not just for the moral agent but also for everyone else affected by the action. It can be assumed to be the paradigm case of consequentialism, which claims that an act is right if and only if it minimizes overall harm, denying that moral rightness depends on anything other than its consequences (Bentham, 1781/1996).

[2]Ethical theory that place special emphasis on the relationship between inviolable norms (i.e. duties) and human actions. This doctrine considers an action as morally good because of some characteristic of the action itself, not because the product of the action is good (Kant, 1785).

sacrifice him/herself for the greater good. Huebner and Hauser (2011) were the first to test the trilemma in an experimental setting, highlighting a high rate of utilitarian judgements independently from the presentation of the altruistic self-sacrifice option. Di Nucci (2013) further deepened this perspective, supporting Thomson's hypothesis (2008) on the advantage of moral trilemmas in revealing the desire to avoid self-sacrifice when pursuing the utilitarian solution.

Perspective-taking (PT) has been recognized as an important bias in affecting moral judgement (Kusev et al., 2016), and the AV dilemma demonstrates that the endorsement of a particular moral code can be negatively affected by personal involvement and required commitment (Bonnefon et al., 2016). A potential solution to this bias would be to investigate the AV dilemma both from the passenger's and the pedestrian's perspectives. Nonetheless, is rare to find such experimental applications in literature, leading to the underestimation of alternative perspectives (Borenstein et al., 2019). Kallioinen et al. (2019) investigated moral attitudes towards autonomous and non-AVs in a virtual reality environment, assuming several perspectives (car occupants, pedestrians and third-party observers). Results showed that moral judgement was influenced by PT, in favour of greater preferences towards self-protective outcomes. Later, Mayer et al. (2021) corroborated this tendency through a vignette-based study and varying the number of characters involved. An overall advantage of the utilitarian resolution was confirmed, but results revealed a general tendency towards non-utilitarian self-protective behaviours within a 1:5 sacrifice ratio. A potentially valuable strategy for the investigation of the role of PT in moral judgement has been retrieved from the theory of the political philosopher John Rawls, leveraging the amount of accessible information at disposal of the moral agent.

The AV dilemma behind the veil of ignorance

In the 1971, *A Theory of Justice*, the philosopher John Rawls describes his conception of 'social contract' as a form of fairness-based agreement between individuals, who would converge on common principles under specific circumstances (Rawls, 1971/2009; Sterba, 1989). For the derivation of these principles, Rawls suggested that individuals' reasoning has to take place in a hypothetical setting, the so-called 'original position', located behind the veil of ignorance (VOI). Behind the VOI, individuals (i.e. rational agents) are deprived of every contextual and personal information about the self and about others (e.g. gender, age, relationships, social and political positions). Rawls' belief is that only in this particular condition individuals are capable to reach a true agreement, assuming 'justice as fairness' (Maxcy, 2002; Moehler, 2018). Behind the VOI, Rawls suggests that the most appropriate decisional process should rely on the maximin strategy ('maximize the minimum'). This egalitarian decisional rule aims at ensuring the greatest possible benefit to the least-advantaged individual (i.e. the 'difference principle'; Rawls, 1973). In other words, following the maximin rule would result in sharing the benefits between the majority, at the cost of reducing for some. 'Decisions under ignorance' situations have been widely discussed in behavioural economics (Krug et al., 2020), suggesting the maximin criterion as a potential solution to optimization problems (Gorissen et al., 2015), but this decisional approach has never been tested in the context of moral psychology.

Rawslian theory has been reinterpreted by John Harsanyi's Equiprobability Model, stressing the role of impersonality over impartiality (Harsanyi, 1975, 1978). According to the philosopher, individuals are assumed to be *Bayesian* agents, aiming at maximizing their expected individual utility (on the basis of the Rational Choice Theory by Coleman, 1994). Nevertheless, when key contextual and personal information is concealed behind the VOI, they will likely follow the average utility principle, as an equal partition of resources between all the involved parts. In other words, when individuals cannot clearly favour themselves, they opt to prioritize collectivity in a more utilitarian rather than egalitarian sense. Importantly, both Rawls and Harsanyi assume that the nature of this decision is selfish, since the person is aware that she/he could end up being any member of the group (Ashford & Mulgan, 2013). Despite some fundamental similarities (Moehler, 2018), Rawls and Harsanyi support two different decisional processes in two defined states of ignorance, describing two types of veils with slightly different

features. Rawlsian VOI can be defined as a *Thick* Veil, where the agent has no reference points and is completely deprived of contextual and individual information on the self and the others (i.e. the 'No Knowledge Formula', Parfit, 2011). Divergently, Harsanyian VOI can be described as a *Thin* Veil (Harsanyi, 1975), where the agent is acquainted with at least limited information (e.g. the social position of the people behind the veil) but knows nothing more about the characteristics of each member and the self (i.e. the 'Equal Chance Formula', Parfit, 2011).

Recently, the conceptualization of the VOI has been operationalized and applied as an experimental setting in the investigation of moral reasoning, and also in the investigation of moral attitudes towards AVs. To our knowledge, the first application was proposed by Huang et al. (2019), which highlighted how VOI's impartial thinking may have the ability to boost utilitarian behaviours. Specifically, participants were asked whether an AV should be required to act utilitarian in response to a road accident involving nine characters and her/himself, given a 1-to-10 chance of being the single AV passenger and a 9-to-10 chance of being one of the nine pedestrians. Results showed that VOI reasoning was able to stimulate utilitarian decisions, interpreting this tendency as an attempt to maximize the odds of a self-beneficial outcome. Recently, Martin, Kusev, and Van Schaik (2021); Martin, Kusev, Teal, et al. (2021) delved into the topic, proposing a novel theoretical model for PT in moral dilemmas (i.e. PT accessibility theory), and disputing two previous methodological approaches (Bonnefon et al., 2016; Huang et al., 2019). First, the authors marked a flaw in the studies by Bonnefon et al. (2016). Indeed, the authors adopted a 'Partial PT accessibility' to the dilemma, exposing participants to only one perspective (AV passenger), and underestimating the role of other perspectives in affecting moral reasoning (Kallioinen et al., 2019; Mayer et al., 2021). Second, the authors criticized the features of the 'Full PT accessibility' paradigm presented by Huang et al. (2019), as flawed by the uneven odds of being each one of the involved characters. In fact, in nine of 10 cases, the participant was one of the pedestrians secured in the utilitarian option, which thereby automatically induce selfishness in the process (Martin, Kusev, Teal, et al., 2021). To fill these gaps, Martin, Kusev, and Van Schaik (2021) developed a between-subjects study comparing moral judgements with Partial PT accessibility (access to only one perspective) and Full PT accessibility (access to all the potential perspectives). Differently from Huang et al. (2019), this version of the AV dilemma with Full accessibility was more faithful to VOI principles (Rawls, 1971/2009), since it assumed even odds of being one of the characters involved in the crash scenario, and therefore eliminating the opportunity of making self-interest-driven decisions. Their findings confirmed that when contingent information is blurred in the Full PT moral scenario, the likelihood of the utilitarian resolution grows, coherently to an increased agreement between moral judgement and willingness to buy utilitarian AVs.

Aims and structure of the research

The present research aims to deepen the role of VOI reasoning in the process of moral evaluations, distinguishing the operationalization of Rawlsian and Harsanyian VOI and stressing the role of PT when facing problems of moral nature. The focus has been directed towards the context of autonomous transportation (Huang et al., 2019; Martin, Kusev, Teal, et al., 2021; Martin, Kusev, & Van Schaik, 2021), implementing a renewed application of VOI reasoning inspired by the moral trilemma paradigm (Thomson, 2008). It allows us to push the investigation beyond the typical utilitarian versus non-utilitarian dispute, considering a third potential solution (namely the maximin strategy). The effects of this decisional process are also tested in terms of moral acceptability and availability to share the autonomous technology, to bring new insights to the general perception of AVs' behaviours (Bonnefon et al., 2016). To this aim, two within-subjects studies were developed in order to test how variations in the availability of contingent information and the imposed perspective can affect moral judgements and perception of AVs (Kallioinen et al., 2019; Martin, Kusev, & Van Schaik, 2021). In Study 1 (S1) we investigate potential differences in the endorsement of different

AV behaviours through a 'funnel' within-subject approach, gradually moving from Full PT accessibility to Partial PT accessibility to increase the availability of contingent information scenario after scenario. Evidence collected in S1 has been further deepened in a follow-up study (Study 2, S2), with the aim of controlling for potential countereffects of the proposed methodology, as well as further testing the role of sacrificial framing in moral judgement (Bruno et al., 2023; Huebner & Hauser, 2011). Below are provided the specifics for each study, followed by a comprehensive discussion. Additional information about the experimental materials and data analysis are retrievable in the Supplementary Materials.

STUDY 1

Coherently with Martin, Kusev, and Van Schaik (2021), in S1 we aimed to test the effect of PT accessibility (partial vs. full) in shaping moral judgements but assuming a different approach. Indeed, Martin, Kusev, and Van Schaik (2021) opted for a between-subject design, comparing moral judgements in partial and full PT accessibility dilemmas. Participants assigned to the partial PT condition were asked to assume a specific perspective (the AV passenger), while in the Full PT condition, the moral agent had even odds of being the AV passenger or one of the pedestrians crossing the road. In our study, first we distinguish between two different Full PT scenarios, structured on the basis of VOI principles by Rawls (i.e. *Thick* Veil) and Harsanyi (i.e. *Thin* Veil). Consistently, to discern the decisional processes in 'decision under ignorance' situations, we have adapted the structure of moral trilemma to VOI reasoning and AV dilemma, directing the focus towards three potential resolutions: utilitarian, maximin and non-utilitarian. Importantly, we opted to provide increasing contingent information adopting a 'funnel' within-subject design, presenting Full PT accessibility scenarios before the disclosure of a specific perspective (AV passenger or pedestrian) in the final partial PT condition (i.e. *No* Veil). This design allows us to investigate moral consistency throughout the veils (i.e. 'moral profiles'), considering participants who have expressed equal or different moral judgements throughout the veils. To the best of our knowledge, this is the first applied research distinguishing between Rawslian-like and Harsanyian-like VOI reasoning, both in terms of information accessibility and in the separation of the two proposed decisional rules (maximin and utilitarian).

We hypothesized that:

H1.1. Coherently with the reference theories (Di Nucci, 2013; Harsanyi, 1975, 1978; Huebner & Hauser, 2011; Rawls, 1971/2009), different decision strategies are expected throughout different VOI conditions. Specifically, the favour towards the maximin decision rule should be higher behind the Rawslian *Thick* Veil, while moral agents should endorse the utilitarian rule more behind the Harsanyian *Thin* Veil.

H1.2. Consistently with previous studies (Bruno et al., 2023; Kallioinen et al., 2019; Mayer et al., 2021), when all personal and contingent information is disclosed in the partial PT accessibility condition (*No* Veil), a transition towards self-protective behaviours is expected.

H1.3. Assuming the within-subjects nature of this study, we expected a coherence between moral consistency and attitudes towards AVs, in terms of moral acceptability of the proposed AV behaviours and willingness to buy. For example, moral agents who have consistently favoured utilitarian behaviours would express better attitudes towards AVs programmed to follow this decisional rule.

Method

Participants

A priori power analysis has been computed on G-power (Faul & Erdfelder, 1992), assuming a medium effect size (Cohen's d = 0.20) and a correlation of .50 among repeated measures, with an alpha error probability of .05 and 0.95 of power. The system suggested a total number of 220 participants, and 251 Italian participants were recruited for the experiment. The final sample counted a total number of 239 participants (12 subjects were excluded as they failed to correctly answer two check questions, see Procedure). The final sample counted 50.21% females ($n = 120$). Overall, the mean age was 28.28 years ($SD = 8.26$, range = 18–63), the mean schooling age was 16.94 years ($SD = 2.78$), and 51.46% of the participants ($n = 123$) were enrolled in university courses. Most of the participants (88.28%, $n = 211$) had held diver licences, and 53.55% of the sample was involved in a road accident at least once in a lifetime ($n = 128$). The study was approved by the local ethics committee (ID No.: 4420).

Materials

The set of stimuli is composed of three sacrificial and self-involvement moral trilemmas, presented with both textual storyline and vignette and readapted by Martin, Kusev, and Van Schaik (2021). The trilemmas depict the same on-road situation involving an AV driving on an urban road with a single passenger on board, approaching an intersection. Because of a non-human-related failure, three pedestrians are now crossing the road in the direct path of the AV. The dynamic of the event does not allow the vehicle to brake safely, leading to an unavoidable crash.

Two trilemmas (Thick and Thin Veil) were framed consistently with VOI principles, having a full PT accessibility to contingent information. Consistently, in these scenarios participants were informed of having even odds of being the AV passenger or one of the pedestrians (Martin, Kusev, & Van Schaik, 2021). Differently, in the last dilemma (No Veil) participants were framed 'outside the veil' and assigned to a specific viewpoint in the scenario (AV passenger or one of the pedestrians), consistently to a partial PT accessibility to contingent information. Below is the storyline and the vignette (Figure 1) from the two VOI trilemmas (*Thick* and *Thin* Veils). The experimental material is retrievable in the Supplementary Materials.

> YOU could be the sole passenger (Pa) in an autonomous self-driving vehicle traveling at the speed limit down an urban road. OR you could be one of the three pedestrians now crossing the road. Pe1 and Pe2 are in the middle of the road, whereas Pe3 is just behind them. Because of a traffic light malfunction, the pedestrians are now in the direct path of the car. There is no more time to brake. Facing this event, the autonomous vehicle may be programmed to implement three different emergency maneuvers, resulting in different risks for the passenger and the pedestrians.

Each trilemma has three different resolutions, corresponding to three potential manoeuvres that the AV would be able to perform in that critical situation. The three manoeuvres were not expressed explicitly, but their outcomes were presented in the form of individual numerical chances of survival (Table 1): the higher the individual percentage, the higher the corresponding chance of survival. The decision to focus on the sole outcome of each manoeuvre was taken to emphasize the consequences of each outcome instead of the events' dynamics.

The selected strategy also allowed us to achieve a VOI-like representation of three decisional rules in the AV dilemma: to prioritize the AV passenger safety (following the non-utilitarian rule), to minimize the number of casualties (following the utilitarian rule), or to maximize the protection of the least-advantaged character (following the maximin rule). The specific chances of survival percentages were

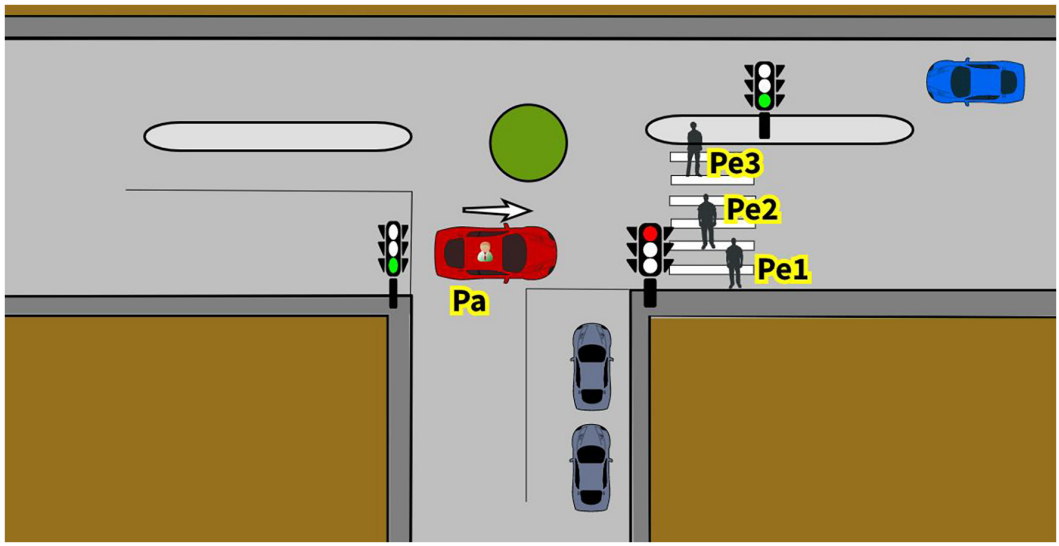

FIGURE 1 Vignette of the AV trilemma deployed in the two VOI conditions (Thick and Thin Veils).

defined to be as faithful as possible with the three AV behaviours, so as to communicate the number of characters at different levels of risk which was consistent with the correspondent decisional rule. The non-utilitarian behaviour favoured the lowest number of characters (two of four) and the lowest expected utility (Morgenstern & Von Neumann, 1944). In the No Veil scenario, this may turn out to be the self-protective option, depending on the experimental condition. The utilitarian rule allowed for the highest number of survivors (three of four) and the highest expected utility. Finally, the maximin behaviour resulted in an increased chance of survival of the character with the highest risk, but distributing it among the other characters, resulting in an expected utility value just below the utilitarian option.

The three dilemmas were presented in a fixed order, increasing the level of individual and contingent information available. When answering the first full PT scenario (Rawslian *Thick* Veil), the chances of survival were disjointed from characters and roles (i.e. no label provided), requiring participants to assume a decision on the sole basis of the final 'economic' outcome, but knowing nothing about their personal role and the role of other characters in the scene (Table 1, Thick Veil). Subsequently, the chances of survival were associated with each role in the second full PT scenario (Harsanyian *Thin* Veil), but participants still had no notion of their personal involvement. (Table 1, Thin Veil). Finally, all the information about the personal role and the role of others was disclosed to participants in the partial PT scenario (Table 1, No Veil), who were asked to assume the perspective of the AV passenger or of one of the pedestrians. Importantly, the chances of survival were kept constant throughout the three dilemmas to reduce the risk of confusion.

Procedure

S1 was programmed on the Qualtrics survey platform (Qualtrics, Provo, UT) and distributed online through Prolific, with a mean duration of 13.36 min ($SD = 6.37$ min). An hourly rate contribution of 12.70€ was provided, and on average, each participant was rewarded with 3.2€ for their participation. All the participants were requested to complete the survey using a laptop or a desktop computer and a mouse. Before the experimental session, each participant filled out the informed consent form, read the instructions about the experimental session, and provided personal information and driving habits. Additionally, two questions on numeracy ability were administered to control for basic knowledge of proportions and percentages. Twelve participants were excluded based on this control check. At this

TABLE 1 The three outcomes depict the three potential AV behaviours in the three scenarios (Thick, Thin and No Veil).

1. Thick Veil (full PT accessibility)				
AV behaviour	**Unknown character**	**Unknown character**	**Unknown character**	**Unknown character**
Non-utilitarian	99%	1%	1%	99%
Utilitarian	1%	99%	99%	99%
Maximin	42%	38%	38%	90%
2. Thin Veil (full PT accessibility)				
	Passenger	**Pedestrian 1**	**Pedestrian 2**	**Pedestrian 3**
Non-utilitarian	99%	1%	1%	99%
Utilitarian	1%	99%	99%	99%
Maximin	42%	38%	38%	90%
3.1 No Veil, AV passenger perspective (partial PT accessibility)				
	YOU	**Pedestrian 1**	**Pedestrian 2**	**Pedestrian 3**
Non-utilitarian	99%	1%	1%	99%
Utilitarian	1%	99%	99%	99%
Maximin	42%	38%	38%	90%
3.2 No Veil, pedestrian perspective (partial PT accessibility)				
	Passenger	**YOU**	**Pedestrian 2**	**Pedestrian 3**
Non-utilitarian	99%	1%	1%	99%
Utilitarian	1%	99%	99%	99%
Maximin	42%	38%	38%	90%

Note: The percentages in each cell indicate the chance of survival of each character (columns) in each outcome (rows). From top to bottom, the first table was recalled in the full Perspective Taking (PT) Thick Veil scenario (with hidden information about personal and others' roles); the second table in the full PT Thin Veil scenario (percentages attached to each role but no information about personal involvement); the third and fourth table in the partial PT No Veil condition (with available information about personal and the others' role). Trilemmas were presented in a fixed order: (1) Thick Veil, (2) Thin Veil and (3) No Veil. Here, half of the sample was requested to assume the AV passenger's perspective (3.1), while the rest assumed the pedestrian perspective (3.2, see Procedure). Importantly, during the experiment, the name of each decisional strategy (Non-utilitarian, Utilitarian, Maximin) was blurred and replaced with a more general label (namely: Behaviour 1, Behaviour 2, and Behaviour 3).

point, the three scenarios were presented in a fixed order. Starting the experimental section, a detailed explanation of the task was provided. Participants were informed about the modality of presentation of the upcoming moral scenarios, matched by descriptive vignettes. In the depicted event, the participants knew that they were involved in the scene, having even odds of being the AV passenger or one of the pedestrians. At this point, the three trilemmas were presented in fixed and increasing order of contingent informativeness, starting from the two full PT scenarios (Thick and Thin Veils), and concluding with the partial PT scenario (No Veil). Participants had unlimited time to read the storylines and watch the vignettes. Switching to the next screen, they were reminded of their involvement in the scenario. Then, they were required to choose the most appropriate AV manoeuvre among the three options (non-utilitarian, utilitarian and maximin) and consistently with the type of veil. Introduced to the last No Veil dilemma, the sample was branched into two gender-balanced groups with different PT accessibility, which requested them to assume the perspective of the AV passenger or of one of the pedestrians. Finally, participants were required to rate from 0 to 100 moral acceptance and willingness to buy AVs programmed to follow the three proposed behaviours (0 = completely unacceptable/ unwilling to buy, 100 = completely acceptable/ willing to buy). The evaluated AVs were programmed: to prioritize the AV passenger before anyone else (passenger-protective AV), to minimize the number of casualties

(utilitarian AV), and to maximize the protection of the least-advantaged character (maximin AV). Differently from the trilemma's options, the three AV behaviours were now explained and made explicit to participants.

Data analysis

The statistical analysis was conducted in the R environment (version 4.1.1; R Core Team, 2021). Given the hypotheses, a series of statistical models were implemented. For the investigation of H1.1 and H1.2, moral decisions to the three AV trilemmas (non-utilitarian, utilitarian and maximin) were assumed as three separate binomial dependent variables in three generalized mixed-effects linear models (M1 – M3), setting the dilemma type (Thick Veil, Thin Veil, No Veil with AV passenger perspective, No Veil with pedestrian perspective) as a fixed effect and participants a random intercept. Then, two further mixed-effects linear models (M4, M5), were implemented for the investigation of potential differences in terms of moral acceptability and willingness to buy AVs programmed to follow the three proposed behaviours. In these latter two models, moral consistency profiles were considered as a fixed effect to investigate H1.3. In these terms, we defined 'fully utilitarian', 'fully maximin' and 'fully non-utilitarian' individuals as the ones who gave the same answers throughout the three scenarios, while 'inconsistent' individuals changed their moral decision at least one time during the experimental session. The 'fully non-utilitarian' profile was excluded from the analysis, considering its scarce numerosity ($n = 3$). The models presented in the main analysis (M1 – M5) are the result of forward stepwise model comparisons. Post hoc pairwise comparisons were considered when requested using the R package 'emmeans' (Lenth et al., 2018), and Bonferroni correction was set as an adjustment method. Additional tables and figures are provided in the Supplementary Materials, together with the final datasets and the R script.

Results

Firstly, we investigated the role of the Full PT accessibility behind the veils in the endorsement of each decisional strategy (H1.1), across three generalized linear models (M1–M3). Surprisingly, the utilitarian moral code was mainly preferred in the Thick Veil trilemma (χ^2 (3) = 56.52, $p < .001$) when compared to the Thin Veil ($z = 3.45$, $p = .003$). The inversion of the predicted trend was also observed in the endorsement of the maximin decisional strategy, which was the more frequently selected outcome responding to the Thin Veil scenario (χ^2 (3) = 24.51, $p < .001$) when compared to the Thick Veil ($z = 2.95$, $p = .019$). These results are represented in Figure 2 and comprehensively in Table 2.

As expected, (H1.2), partial PT accessibility increases the favour towards the revealed self-protective behaviours. Indeed, participants tend to prefer the non-utilitarian (χ^2 (3) = 29.02, $p < .001$, $z = 3.06$, $p = .013$) or utilitarian behaviours (χ^2 (3) = 56.52, $p < .001$; $z = 6.74$, $p < .001$) when it matched with self-protective outcomes (Figure 3). Interestingly, under partial PT accessibility, the maximin strategy appeared to be highly preferred by AV passengers than pedestrians (χ^2 (3) = 24.51, $p < .001$; $z = 4.00$, $p < .001$), since in this case, the endorsement of the utilitarian AV manoeuvre would have led to the passenger's sacrifice.

Finally, the consistency between moral judgements and attitudes towards AVs (H1.3) was investigated with two mixed effects linear models (M4 – M5). The interaction effect between AV behaviour and Moral profile (χ^2 (4) = 67.32; $p < .001$) confirmed the hypothesized trend among morally consistent agents. Indeed, 'fully utilitarian' participants evaluated the utilitarian behaviour option as more moral than 'fully maximin' individuals ($z = 27.17$; $p < .001$), and – on the contrary – 'fully maximin' participants evaluated the maximin option as more moral than 'fully utilitarian' individuals ($z = 17.84$, $p = .001$). Interestingly, morally 'inconsistent' participants evaluated the passenger-protective AV as less immoral than 'fully utilitarian' ($z = 14.13$, $p = .001$) and 'fully maximin' moral agents ($z = 28.76$, $p < .001$). In terms of willingness to buy, 'fully utilitarian' participants would prefer to buy utilitarian AVs when

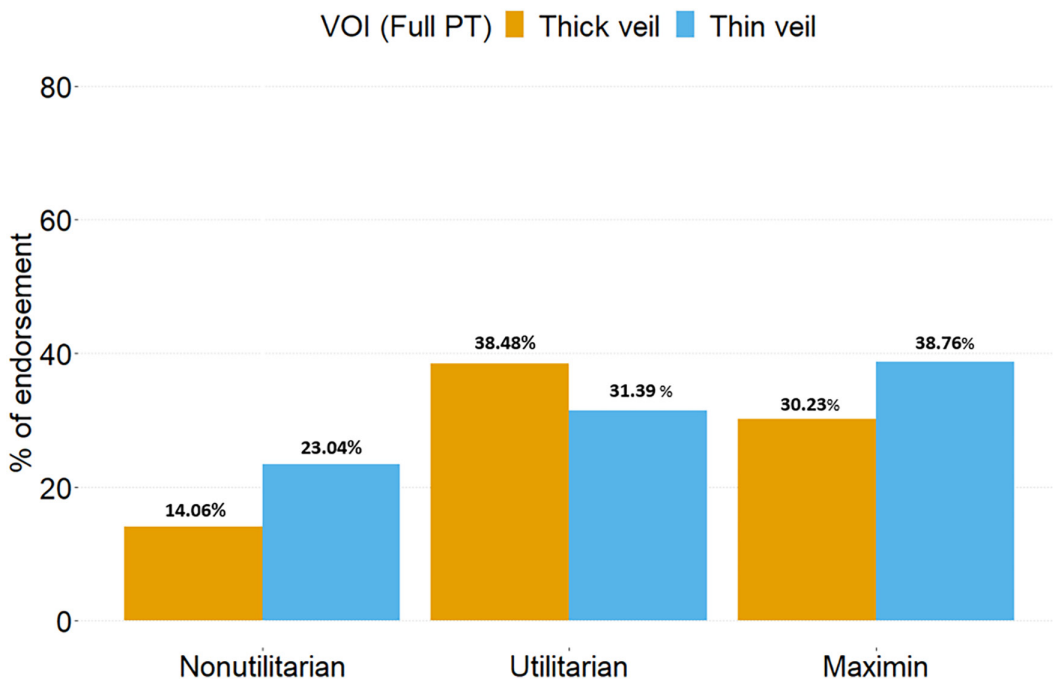

FIGURE 2 Bar chart of the relative percentage of endorsement of the three proposed AV manoeuvres (non-utilitarian, utilitarian and maximin), throughout the two Full PT accessibility conditions (Rawslian *Thick* Veil, Harsanyian *Thin* Veil).

compared to other algorithms (e.g., vs. maximin AVs: $z = 16.49$, $p < .001$) and different Moral profiles (vs. 'fully maximin': $z = 21.49$, $p = .001$). Overall, utilitarian AVs were perceived as more morally acceptable (χ^2 (2) = 342.63; $p < .001$) and more appealing in case of future purchase (χ^2 (2) = 43.19; $p < .001$), when compared to passenger-protective vehicles and AVs programmed to distribute the risk among characters in a maximin sense. The table of the descriptive scores of moral acceptability and willingness to buy is retrievable in the Appendix 1 (Table A1).

Discussion

S1 investigates if variations in PT accessibility and availability of contingent information have the potential to affect moral judgement and attitudes towards AVs, leveraging on the structure of the moral trilemma paradigm (Thomson, 2008) and through the operationalization of VOI reasoning. Coherently with the reference theories (H1.1; Harsanyi, 1975, 1978; Rawls, 1971/2009), we expected a stronger favour towards the maximin decisional strategy when both contingent and personal information were hidden behind the Rawslian *Thick* Veil. Conjunctively, a stronger favour towards the utilitarian code was expected behind the Harsanyian *Thin* Veil, when only contingent but no personal information was disclosed. Interestingly, evidence is in slight opposition with the theoretical framework: when answering to the two VOI scenarios, individuals show a higher preference for the utilitarian code behind the *Thick* Veil, and a favour towards the maximin strategy when responding to the *Thin* Veil trilemma. It appears that the disclosure of other characters' roles (i.e. labels) improves the interest in maximizing the protection of the least-advantaged individual, at the expense of the utilitarian resolution for the minimization of casualties. This solution appears to aim for a more 'democratic' distribution of risk among all the characters involved when compared to the more 'economical' utilitarian resolution for the minimization of the total number of casualties. Considering this result, we can assume that, when adding the individual labels as the simplest contextual

TABLE 2 Percentages of absolute endorsement of the three proposed AV behaviours, divided by moral scenario (Thick Veil, Thin Veil and No Veil) and – separated by the dotted line – controlled for the perspective assumed in the No Veil scenario (partial PT; AV passenger or pedestrian).

	Scenario			Perspective (No Veil)		
	Thick Veil	**Thin Veil**	**No Veil**	**AV passenger**	**Pedestrian**	**Overall**
Non-utilitarian (%)	3.76% ($n = 9$)	6.2% ($n = 15$)	16.73% ($n = 40$)	28.33% ($n = 34$)	5.04% ($n = 6$)	8.92% ($n = 64$)
Utilitarian (%)	63.61% ($n = 152$)	51.89% ($n = 124$)	49.80% ($n = 119$)	27.50% ($n = 33$)	72.27% ($n = 86$)	55.09% ($n = 395$)
Maximin (%)	32.63% ($n = 78$)	41.84% ($n = 100$)	33.47% ($n = 80$)	44.17% ($n = 53$)	22.69% ($n = 27$)	35.98% ($n = 258$)

Note: The overall percentages are presented in the last column.

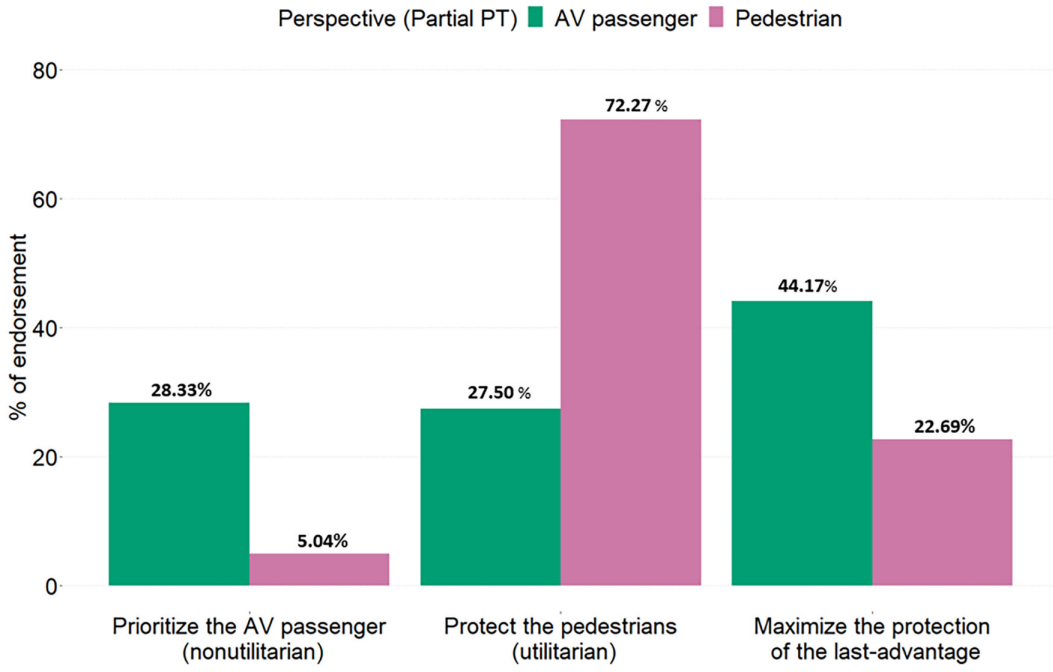

FIGURE 3 Bar chart on the total percentage of endorsement of the three proposed AV behaviours when the individual perspective (i.e. personal role) is revealed in the Partial PT accessibility condition (protect the AV passenger/non-utilitarian, protect the pedestrians/utilitarian and Maximin), divided by individual perspective (AV passenger, pedestrian).

information, the gap between the utilitarian and the maximin moral code is reduced, enhancing a more distributive approach to risk. The low appeal of the maximin rule behind the *Thick* Veil is predicted by Rawls himself (Rawls, 2001), which describes the use of this decisional strategy as a 'useful heuristic device' for the justification of the 'difference principle'. Moreover, Moehler (2018) suggests that, behind the Rawslian veil, rational agents may aim to maximize the individual expected utility independently of their personal role (Briggs, 2014), which would result in an improved likelihood of the utilitarian option.

As predicted (H1.2), the transition from full to partial PT accessibility – when personal involvement is finally disclosed to moral agents – led to a greater favour towards self-protective behaviours. Indeed, awarding the participant with a specific perspective (AV passenger or pedestrian) allowed them to discern the individual advantages (and the relative costs) from the main collective goal. Results show that individuals with the pedestrian perspective endorsed the utilitarian (and self-protective) option more frequently than the passenger-perspective counterpart. Indeed, a tendency towards self-protection was revealed also in this latter condition even if with considerably less intensity, assuming the non-utilitarian feature of the self-protective option. Interestingly, when personal life was endangered in the utilitarian option, the likelihood of the maximin AV behaviour consistently increased, confirming the potential role of this decisional rule as an intermediate moral solution in partial PT accessibility scenarios. The personal involvement effect was strongly anticipated by the traditional literature on moral decision-making (Greene et al., 2001; Lotto et al., 2014; Moore et al., 2008), and this result seems consistent with evidence collected on AV morality using partial PT accessibility scenarios (Kallioinen et al., 2019; Martin, Kusev, & Van Schaik, 2021; Mayer et al., 2021). Nonetheless, the authors opted to investigate this feature through the only AV perspective, highlighting a detrimental effect on utilitarian endorsement. The present study also focuses on the pedestrian perspective, strengthening the role of partial PT accessibility as a motivator towards self-protective tendencies.

Finally, we expected a potential coherence between moral judgements and attitudes towards AVs (H1.3). Three 'moral profiles' were derived from the experimental activity, describing moral agents who consistently followed the utilitarian moral code (i.e. 'fully utilitarian'), who consistently followed the maximin strategy (i.e. 'fully maximin'), and who changed their decisional rule at least one time (i.e. 'inconsistent'). As expected, consistent individuals evaluated the AVs programmed to follow the same moral goal as more acceptable, but this trend appears to be sustained by the availability of purchasing this vehicle only in the 'full utilitarian' individuals. This result suggests an effect of full and partial PT accessibility on 'fully utilitarian' individuals in improving the agreement between moral evaluation and shareability (Martin, Kusev, & Van Schaik, 2021), but overall, the pattern seems in line with the 'social dilemma' exposed by Bonnefon et al. (2016).

STUDY 2

In order to investigate the described hypotheses, in S1 a series of decisions were taken in order to define a methodology that was the most suitable to our research goals. In this sense, it appears important to give credit to potential comments on the eventual limitations of S1, testing it through a follow-up study (S2).

A primary question would be on a potential order effect. In S1, the three AV trilemmas were administered to each participant in a fixed order (Thick, Thin and No Veil). The transition from full to partial PT is a core feature of S1 and is plausible to expect different results when the personal role and contingent information are not disclosed progressively but randomizing the scenarios. A second discussion point is on the reliability of the evidence collected by administering a single trial per each VOI condition. This may be non-trivial, since this may have affected both moral judgement and moral consistency, categorizing as 'morally inconsistent' individuals on the basis of a limited number of evaluations.

Finally, a general concern may be referred to as the structure of moral alternatives. In this partial PT accessibility situation, participants make moral judgements being aware of their role in the scenario, and hence also knowing which of the alternatives will require self-sacrifice. Consequently, is admissible that self-sacrificial framing may impact the endorsement of the utilitarian act, making it more attractive ('kill one to save yourself and others') or more unpleasant ('kill yourself to save others'). This may be the case in S1, where individuals embracing the pedestrian perspective in the No Veil condition had the 'advantage' to protect themselves and two more at the expense of the AV passenger's life. The role of self-sacrifice has been deepened in previous research (e.g. Di Nucci, 2013; Greene et al., 2001; Huebner & Hauser, 2011; Moore et al., 2008), but little is known about its framing (Bruno et al., 2023).

Taking together, S2 aims at deepening these discussion points, investigating from a descriptive viewpoint how the experimental design from S1 may have affected the results. To fulfil this goal, the presentation of the three veils was randomized in a within-subjects study. Moreover, to investigate the role of sacrifice framing in the endorsement of the utilitarian moral rule, two different versions of the partial PT accessibility scenario were administered: in one case, pursuing the utilitarian behaviour would require the participant to accept a self-sacrificial act, in the opposite case this moral outcome is instead framed as self-protective. Additionally, the stimuli set was renewed with an additional scenario per condition, presenting eight AV trilemmas per participant.

Method

Participants

A priori power analysis has been computed on G-power (Faul & Erdfelder, 1992), assuming a conservative low-to-medium effect size (Cohen's $d = 0.15$) and a correlation of .50 among repeated measures, with an alpha error probability of .05 and 0.95 of power. The system suggested a total number of 62

participants, and 76 Italian participants were recruited for the experiment. The final sample counted 48.68% females ($n = 37$). Overall, the mean age was 27.04 years ($SD = 3.55$, range = 21–44), the mean schooling age was 17.68 years ($SD = 2.18$), and 42.10% of the participants ($n = 32$) were enrolled in university courses.

Materials

The set of stimuli is composed of eight sacrificial and self-involvement moral trilemmas. Three scenarios were retrieved by S1, and five new AV trilemmas were developed for the occasion. For each new scenario, the structure of each storyline and vignette was consistent with S1, depicting an AV approaching a crosswalk with three pedestrians crossing the road in the unlucky event of an unavoidable road accident. Four dilemmas were framed consistently with VOI principles and having full PT accessibility to contingent information (see Materials). Differently from S1, the partial PT accessibility scenarios (No Veil) only asked participants to assume the perspective of one of the three pedestrians. Two No Veil scenarios depicted the participants as the one right outside the AV trajectory (Utilitarian Self-Sacrificial, USS; Figure 4), while the remaining two asked the participants to assume the perspective of one of the two pedestrians right in front of the AV (Utilitarian Self-Protective, USP; Figure 4). Response modality was also coherent with S1, providing participants with three potential AV behaviours in the form of numerical individual chances of survival. Per each VOI condition, participants were provided with different levels of personal and contingent information (see Materials). In S2, the chances of survival of the No Veil scenario outcomes were adapted consistently with the adopted sacrifice framing in the utilitarian resolution. Accordingly, both USS and USP conditions required participants to assume the perspective of one of the pedestrians, but if in the first case (USS), the utilitarian decision required a self-sacrificial act in order to be accomplished, in the USP condition participant had the chance to protect the self and also the other two pedestrians (Table 3).

Procedure

S2 was programmed on the Qualtrics survey platform (Qualtrics, Provo, UT). The programme provided an anonymous link to the survey, which was then distributed via social networks and institutional communication channels following a snowball non-probabilistic sampling technique (Chandler et al., 2019; Goodman, 1961; Parker et al., 2019). Each participant signed an informed consent form before participation, which was voluntary and unremunerated.

The experimental procedure was consistent with S1 (see Procedure). Differently from the previous experiment, the dilemma set was randomized within-subjects following a Latin square design. This technique allowed for randomization of the three VOI conditions (Thick, Thin and No Veil) but having the full set of possible order with equal numerosity. To reduce the risk of confusion due to the switch from full to partial PT accessibility scenarios, the presentation of AV trilemmas belonging to the same VOI condition was randomized pairwise, and the four No Veil trilemmas were randomized as a single block. After the administration of AV trilemmas and consistently with S1, participants were required to rate from 0 to 100 moral acceptance and willingness to buy AVs programmed to follow the three proposed behaviours (0 = completely unacceptable/ unwilling to buy, 100 = completely acceptable/ willing to buy).

Data analysis

The statistical analysis was conducted in the R environment (version 4.1.1; R Core Team, 2021). Assuming the follow-up nature of the present study, the data analysis procedure from S1 was replied for S2

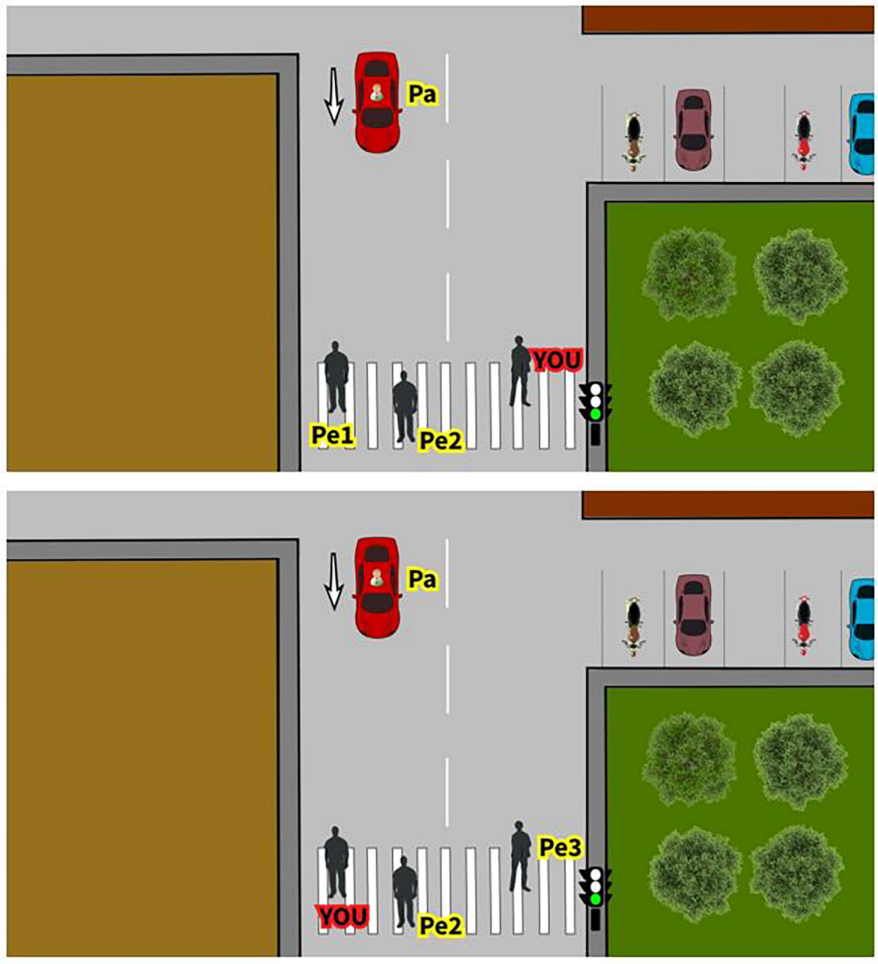

FIGURE 4 Vignette of the No Veil AV trilemma, in the utilitarian self-sacrificial framing (USS, above) and in the utilitarian self-protective framing (USP, below).

(see Data analysis). Again, three moral consistency profiles ('fully utilitarian', 'fully maximin' and 'inconsistent') were derived for the investigation of attitudes towards AVs. Consistent individuals were the ones who gave the same answers throughout all eight scenarios, while 'inconsistent' individuals changed their moral decision at least one time during the experimental session. For descriptive purposes, two additional profiles were considered in Table A2 (Appendix 1), gathering participants with at least six of eight consistent moral judgements ('mainly utilitarian' and 'mainly maximin'). Additional information is provided in the Supplementary Materials, together with the final datasets and the R script.

Results

Table 4 describes the percentage of endorsement of the three available AV behaviour behind the Thick and Thin Veils (full PT) and outside the veil (partial PT), controlling for the sacrifice framing of the utilitarian option. Overall, results seem consistent with S1 in the predictable low interest towards the non-utilitarian option (Table 4), and, descriptively, a preference towards a particular decisional approach behind the *Thick* (utilitarian) and the *Thin* (maximin) Veils is still partially observable (Figure 5). Nevertheless, randomization appears to have significantly affected moral judgement in the

TABLE 3 The three outcomes depict the three potential AV behaviours in the No Veil scenarios (utilitarian self-sacrificial and utilitarian self-protective).

No Veil, utilitarian self sacrificial (USS – partial PT accessibility)				
AV behaviour	**Passenger**	**Pedestrian 1**	**Pedestrian 2**	**YOU**
Non-utilitarian	1%	1%	99%	99%
Utilitarian	99%	99%	99%	1%
Maximin	42%	38%	90%	38%
No Veil, utilitarian self-protective (USP – partial PT accessibility)				
	Passenger	**YOU**	**Pedestrian 2**	**Pedestrian 3**
Non-utilitarian	99%	1%	1%	99%
Utilitarian	1%	99%	99%	99%
Maximin	42%	38%	38%	90%

Note: The percentages in each cell indicate the chance of survival of each character (columns) in each outcome (rows) for the two partial PT accessibility conditions. Here is possible to observe the different nature of the utilitarian AV behaviour in terms of individual involvement. In the case of self-sacrificial framing (USS), the utilitarian option is satisfied only in the case of self-sacrifice. In the case of self-protective framing (USP), the utilitarian option is satisfied also protecting the participant. Importantly, during the experiment, the name of each decisional strategy (non-utilitarian, utilitarian, and maximin) was blurred and replaced with a more general label (namely: Behaviour 1, Behaviour 2 and Behaviour 3).

full PT accessibility conditions. Indeed, results indicate a non-significant difference between the two VOI veils in the likelihood of utilitarian (χ^2 (3) = 46.60, $p < .001$; $z = 1.34$, $p = 1$) and maximin decisions (χ^2 (3) = 25.56, $p < .001$; $z = 0.21$, $p = 1$). This effect may be due to an improvement in the endorsement of the maximin approach behind the two veils (Table 4).

Importantly, the trend towards favouring self-protecting behaviours was also confirmed in S2, where participants assumed the pedestrian's perspective in two different sacrificial framings. Indeed, participants tend to prefer the non-utilitarian (χ^2 (3) = 42.24, $p < .001$, $z = 5.76$, $p < .001$) or utilitarian behaviours (χ^2 (3) = 46.60, $p < .001$; $z = 6.29$, $p < .001$) when it matched with self-protective outcomes (Figure 6). Interestingly – and consistently with S1 – when facing the AV trilemma with the perspective of the pedestrian at risk in the utilitarian option (USS), the likelihood of the maximin decisional rule was higher than in the opposite sacrifice framing (USP; χ^2 (3) = 25.56, $p < .001$; $z = 3.83$, $p < .001$).

Finally, also in S2 three moral profiles were derived from the expressed moral judgements, so to re-evaluate the consistency between moral judgements and attitudes towards AVs. Consistently with S1, Moral acceptability of the proposed AV algorithms was evaluated differently from different moral profiles (χ^2 (4) = 28.59; $p < .001$), with higher evaluations from 'fully utilitarian' and 'fully maximin' individuals of the corresponding AV behaviour when compared to other algorithms and moral profiles. Finally, a generally low interest in purchasing AVs was confirmed independently from the moral profile. The table of the descriptive scores of moral acceptability and willingness to buy is retrievable in the Appendix 1: Table A2.

Discussion

S2 is assumed as a follow-up investigation, with the aim of testing a series of methodological decisions assumed in the implementation of S1. Importantly, results confirm the relevance of the 'funnel' within-subject approach followed in S1 in revealing characteristic approaches in the resolution of the AV trilemma behind the VOI. Indeed, presenting personal and contingent information about the AV scenario in a randomized order changed the approaches towards the dilemmas, slightly improving the likelihood of the maximin resolution. This was predictable. In the development of S1, we sought to maintain the AV trilemma presentation in a fixed order since we aimed to disclose

TABLE 4 Percentages of absolute endorsement of the three proposed AV behaviours, divided by moral scenario (Thick Veil, Thin Veil and No Veil) and – separated by the dotted line – controlled for the sacrifice framing in the No Veil scenario (utilitarian self-sacrificial option, USS; utilitarian self-protective option, USP).

	Scenario			Sacrifice framing (No Veil)		
	Thick Veil	Thin Veil	No veil	Utilitarian self-sacrificial	Utilitarian self-protective	Overall
Non-utilitarian (%)	3.29% ($n = 5$)	4.63% ($n = 7$)	13.81% ($n = 42$)	25.00% ($n = 38$)	2.63% ($n = 4$)	8.89% ($n = 54$)
Utilitarian (%)	46.05% ($n = 70$)	38.41% ($n = 58$)	45.39% ($n = 138$)	25.26% ($n = 40$)	64.47% ($n = 98$)	43.82% ($n = 266$)
Maximin (%)	50.65% ($n = 77$)	56.95% ($n = 86$)	40.79% ($n = 124$)	48.68% ($n = 74$)	32.89% ($n = 50$)	47.28% ($n = 287$)

Note: The overall percentages are presented in the last column.

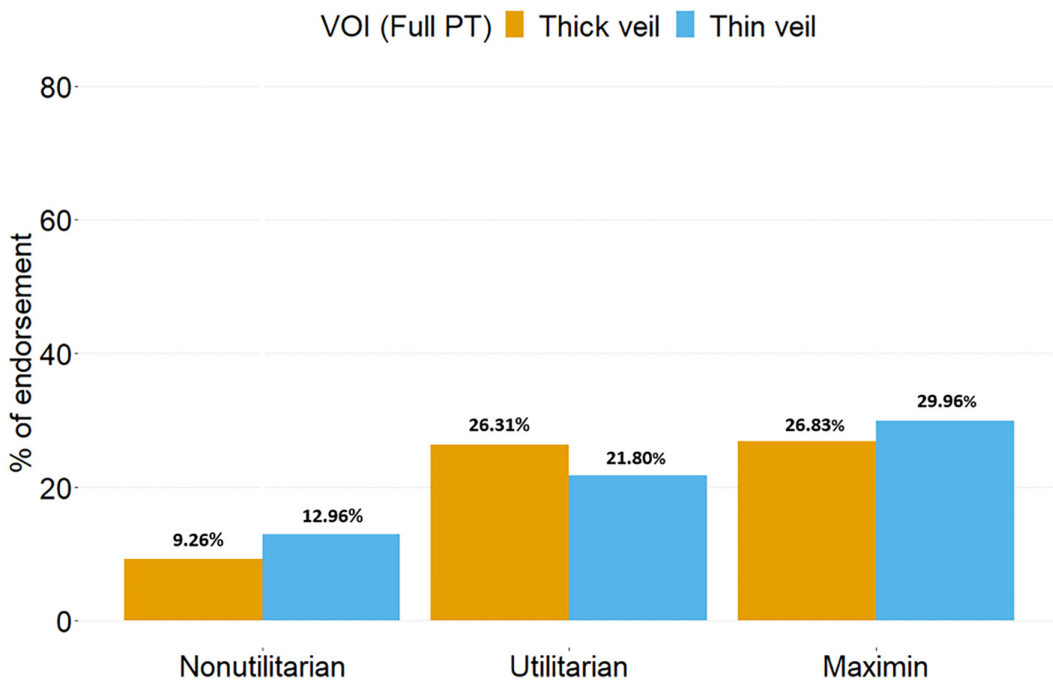

FIGURE 5 Bar chart of the relative percentage of endorsement of the three proposed AV manoeuvres (non-utilitarian, utilitarian and maximin), throughout the two Full PT accessibility conditions (Rawslian *Thick* Veil, Harsanyian *Thin* Veil).

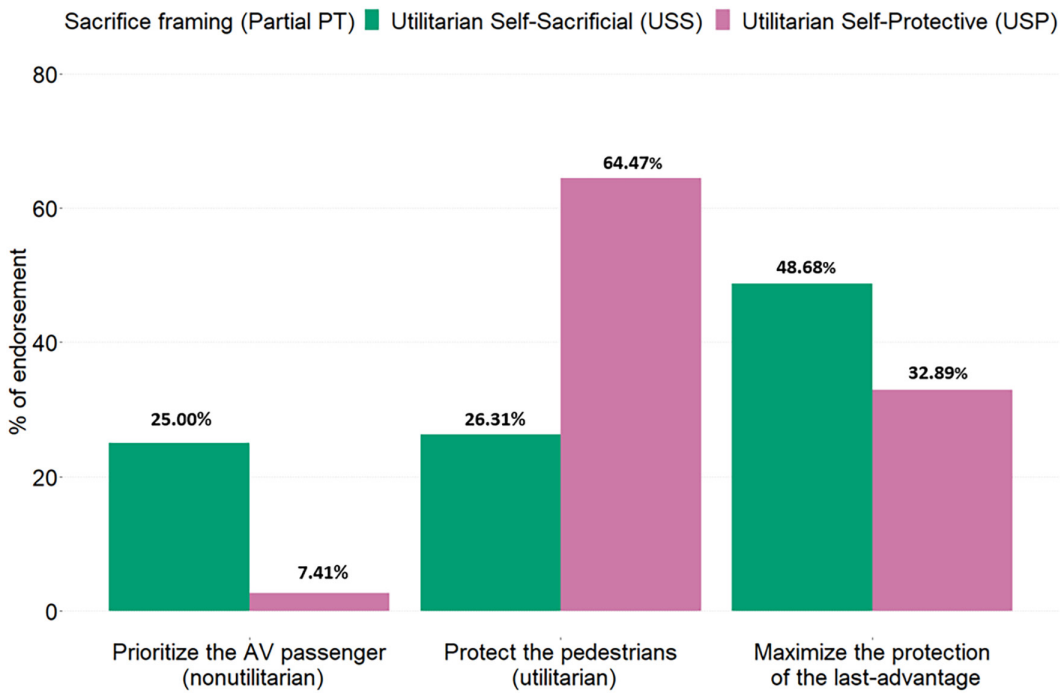

FIGURE 6 Bar chart on the total percentage of endorsement of the three proposed AV behaviours when the individual perspective (i.e. personal role) is revealed in the Partial PT accessibility condition (protect the AV passenger/non-utilitarian, protect the pedestrians/utilitarian and maximin), divided by sacrifice framing. In the utilitarian self-sacrificial framing (USS) the utilitarian option is satisfied only in the case of self-sacrifice. In the case of self-protective framing (USP), the utilitarian option is satisfied also protecting the participant.

progressive information about personal involvement and contingent features of the scenario step-by-step. Randomizing the accessibility to PT does not seem a suitable solution to enhance VOI reasoning, since moral agents are no longer bereft of personal and contingent information (Harsanyi, 1975, 1978; Moehler, 2018; Rawls, 1971/2009), which may assumingly influence the decisional processes (Di Nucci, 2013).

Nonetheless, is possible to admit how results from S2 may suggest a good consistency with trends evidenced in S1. A low overall interest towards non-utilitarian AV behaviours was confirmed, as well as the consistency between attitudes towards this technology and the expressed moral judgements, assumed in different 'moral profiles'. Preliminarily, these findings seem independent of the number of stimuli administered to each participant, which seems important in order to control the time and cognitive requests of participants (Broeders et al., 2011; Bruno et al., 2022). Importantly, S2 confirmed the hypothesis of a non-trivial role of sacrifice framing in moral reasoning, which consistently affects moral judgement in relation to how the utilitarian option is framed. Indeed, S2 supports S1 and the literature on self-protective tendencies in partial PT accessibility conditions (Bruno et al., 2023; Kallioinen et al., 2019; Mayer et al., 2021). Here, we have framed the utilitarian option only using a single role (the pedestrian), suggesting how this factor might be independent of the role played in the AV dilemma.

GENERAL DISCUSSION

The present study aimed at deepening the utility of VOI principles and PT accessibility in investigating moral reasoning in the AV dilemma, taking advantage of the collected evidence on the role of full PT accessibility in investigating moral preferences towards this technology (Huang et al., 2019; Martin, Kusev, Teal, et al., 2021; Martin, Kusev, & Van Schaik, 2021). Differently from previous studies, we decided to study PT accessibility following a new 'funnel' approach, disclosing additional information about the moral scenario through a progressive transition from full to partial PT. Importantly, we distinguished between two different Full PT scenarios, operationalizing VOI reasoning principles by Rawls (1971/2009) and Harsanyi (1975, 1978) in the form of moral trilemmas (Di Nucci, 2013; Thomson, 2008), so to dispose of two different levels of informativeness behind the veil and to disentangle the advantage of utilitarian attitudes providing a new decisional approach (i.e. maximin) to the process.

S1 and S2 confirm that full PT can actively influence individual moral attitudes when compared to partial PT accessibility and that this is not independent of the order in which information is presented. Consistently with the literature (Martin, Kusev, Teal, et al., 2021), a general selfless interest in the prosocial benefits of utilitarian behaviours behind the VOI is confirmed, assuming (i) the impossibility of taken selfish decisions (even odds) and (ii) the avoidance of any 'anchoring effect' presenting this condition as first (Di Nucci, 2013). Assumingly, the interest in self-protective actions has been evidenced to have a detrimental role on prosocial behaviours, operating when personal roles are disclosed in partial PT conditions (Bruno et al., 2023; Kallioinen et al., 2019; Mayer et al., 2021), and influencing moral agents towards a moral solution that prevents or limits individual risks. In the vision of these results, it appears clear that personal interest cannot be underestimated when evaluating moral attitudes under partial PT accessibility situations. We stress the recommendation for both researchers and practitioners to consider PT accessibility and personal roles when investigating moral concerns and attitudes, especially when towards AVs (Huang et al., 2019; Martin et al., 2017; Martin, Kusev, & Van Schaik, 2021) since it can help the intelligibility of general concerns and hesitations limiting and slowing the adoption of this technology (Kaur & Rampersad, 2018; Shariff et al., 2017, 2021).

The operationalization of two VOI conditions (Thick and Thin Veil) and the 'funnel' within-subjects approach have been important in revealing a significant preference towards utilitarian resolutions when moral agents have limited information about the self and the other characters involved in the scenario (Thick Veil). Oppositely to our hypothesis, the sole disclosure of other characters' roles (Thin Veil) improved the interest to maximize the protection of the least-advantaged individual, reducing the

likelihood of the utilitarian decisional rule. Considering this result, we can assume that, when adding the individual labels as the simplest contextual information, the gap between the utilitarian and the maximin moral code decrease, enhancing a more distributive approach to risk. The reduced likelihood of the maximin rule behind the Rawslian veil was plausible for Rawls (2001) and potentially explained by Moehler (2018), who suggests that rational agents aim to maximize the expected individual utility also when unaware of their personal position. Importantly, Moehler (2018) pointed out that the dispute for the appropriate decisional rule behind the veil is originally conceived only at a normative level. However, formal methods alone cannot offer a full description of moral ideas behind specific moral decision situations, and other principles of justice and morality are conceivable other than Rawls' original position and Harsanyi's equiprobability model (Moehler, 2018).

Overall, the utilitarian endorsement seems to be negatively affected by the amount of contextual and personal information. This result deserves further investigation, by focusing – for example – on how the progressive disclosure of additional personal and contextual features of the dilemma may affect moral reasoning (e.g. gender, age, social position and potential negligence in the traffic). It must be acknowledged that the present operationalization of Rawslian and Harsanyian VOIs may seem rather simplistic, as it condenses two complex and tangled theories in a simpler experimental setting. Additionally, the employment of percentages in risk communication is controversial (e.g. De Melo et al., 2021; Gigerenzer et al., 2007; Gigerenzer & Galesic, 2012; Peters et al., 2006, 2011). Despite our care in following indications on their use, potential limitations have to be taken into account. In this sense, is it worth pointing out that this study does not aim at stressing the discrepancies between the two theories (e.g. Frohlich et al., 1987; Gaus & Thrasher, 2015; Moehler, 2018), but the VOI principles have served as an inspiration to deepen the impact of PT accessibility in the moral perception and general attitudes towards AVs. Further studies may opt for manipulating the numerical risk, revealing for instance the existence of numerical thresholds for switching decisional strategies or even propose a more effective method to investigate VOI reasoning that dispenses with a numerical representation of risk.

Interestingly, the present research is one of few applications that tested moral trilemma as a useful experimental tool in the investigation of morality, especially in the field of moral perception of AVs. Further studies may stress the role of PT in the resolution of moral trilemmas, especially towards self-protective behaviours. Di Nucci (2013) empirically tested the role of a third moral solution in affecting moral reasoning in the traditional Switch problem (Foot, 1967; Thomson, 2008). Despite some methodological flaws, Di Nucci showed that requiring the moral agent to deal with a three-option dilemma has the effect of increasing the impermissibility of the utilitarian (and self-protective) alternative in a following dichotomous Switch problem. Additionally, moral judgement in moral dilemmas appears to be affected by the dilemma structure itself (e.g. Greene et al., 2001; Lotto et al., 2014; Schein, 2020), but everyday situations like road driving may affect moral reasoning (Bruno et al., 2022). We believe that future studies may continue to investigate the potentiality of moral trilemmas also in the field of AI ethics, which could help overcome a series of limitations regarding the stark distinction between the utilitarian and non-utilitarian decisional criteria (e.g. Evans et al., 2020; Rhim et al., 2021) and disclosing self-interested or egoistic motives towards self-protective behaviours (Bruno et al., 2023; Di Nucci, 2013; Gino et al., 2016). Importantly, leveraging on multiple alternatives and a probabilistic risk formulation, the three-option format can become a valid experimental tool to investigate attitudes towards AVs other than the classic life and death edge cases employed in the trolley problem (Goodall, 2016). It will allow to consider additional layers of information, potentially more applicable with the AV's ethics of risk (Contissa et al., 2017; Geisslinger et al., 2021; Gogoll & Müller, 2017).

Finally, the employment of a sequential paradigm – typical of within-subjects experiments – allowed us to define 'moral profiles' on the basis of individuals' moral judgements of the presented dilemmas. Sequential behaviour paradigms are tasks in which the same individual has to face a choice in consideration of relevant prior behaviour (Mullen & Monin, 2016). Throughout this task, individuals can be consistent or inconsistent with their initial behaviour (e.g. Conway & Peetz, 2012; Fanggidae et al., 2022). Moral consistency has been detected mainly when individuals think abstractly, transcending the actual event, and focusing on superordinate goals and values (e.g. Cornelissen et al., 2013; Trope

& Liberman, 2010). Overall, the progressive introduction of AV technology in the traffic system has to deal with a number of psychological roadblocks shared with other AI systems (Shariff et al., 2017), such as AI literacy (Wang et al., 2022), trust in automation (Jessup et al., 2019), acceptance of autonomous systems (Zhang et al., 2019). Results show that investigating moral consistency can actually improve the intelligibility of general attitudes towards AVs, as the acceptability and the shareability of AV algorithms seems to converge with morally consistent profiles. Further studies may continue in this direction, integrating the investigation of individual ethics, attitudes, trust and intentionality towards the adoption of AVs (Martinho et al., 2021; Panagiotopoulos & Dimitrakopoulos, 2018; Qian et al., 2023).

In conclusion, the present study brings novel findings in the multifaceted discussion about the morality of AVs' behaviours, proposing a new approach for investigating how PT accessibility can affect perception and attitudes towards this new technology.

AUTHOR CONTRIBUTIONS

Giovanni Bruno: Conceptualization; formal analysis; investigation; methodology; resources; writing – original draft. **Andrea Spoto:** Formal analysis; methodology; supervision; writing – review and editing. **Michela Sarlo:** Methodology; supervision; writing – review and editing. **Lorella Lotto:** Methodology; supervision; writing – review and editing. **Alex Marson:** Investigation. **Nicola Cellini:** Supervision; writing – review and editing. **Simone Cutini:** Supervision; writing – review and editing.

ACKNOWLEDGEMENTS

This work was carried out within the scope of the project 'use-inspired basic research', for which the Department of General Psychology of the University of Padova has been recognized as 'Dipartimento di Eccellenza' by the Ministry of University and Research.

FUNDING INFORMATION

This research was supported and partially funded by the Psychological Sciences PhD course of the Department of General Psychology at the University of Padua.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://osf.io/vxe8w/.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Open Science Framework and Supplemental materials for this article can be found online at: https://osf.io/vxe8w/.

ETHICS APPROVAL

The study was approved by the local ethics committee (ID No.: 4420).

CONSENT TO PARTICIPATE

Informed consent was obtained from all individual participants included in the study.

ORCID

Giovanni Bruno https://orcid.org/0000-0002-2526-880X
Andrea Spoto https://orcid.org/0000-0002-7580-544X
Michela Sarlo https://orcid.org/0000-0001-6652-7604

Lorella Lotto https://orcid.org/0000-0002-0303-3014
Alex Marson https://orcid.org/0000-0002-2760-2154
Nicola Cellini https://orcid.org/0000-0003-0306-4408
Simone Cutini https://orcid.org/0000-0001-6332-3219

TWITTER

Giovanni Bruno GioBluno

REFERENCES

Ashford, E., & Mulgan, T. (2013). Contractualism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (summer 2018 edition)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/contractualism/

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.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*, 154–161.

Bentham, J. (1781/1996). *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press.

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*, 1–12.

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

Borenstein, J., Herkert, J., & Miller, K. W. (2019). Autonomous vehicles and the ethical tension between occupant and non-occupant safety. *Computer Ethics-Philosophical Enquiry (CEPE) Proceedings*, *2019*(1), 6.

Briggs, R. A. (2014). *Normative theories of rational choice: Expected utility*. https://plato.stanford.edu/entries/rationality-normative-utility/

Broeders, R., Van den Bos, K., Müller, P. A., & Ham, J. (2011). Should I save or should I not kill? How people solve moral dilemmas depends on which rule is most accessible. *Journal of Experimen-Tal Social Psychology*, *47*(5), 923–934.

Bruno, G., Sarlo, M., Lotto, L., Cellini, N., Cutini, S., & Spoto, A. (2022). Moral judgment, decision times and emotional salience of a new developed set of sacrificial manual driving dilemmas. *Current Psychology*, *42*, 13159–13172. https://doi.org/10.1007/s12144-021-02511-y

Bruno, G., Spoto, A., Lotto, L., Cellini, N., Cutini, S., & Sarlo, M. (2023). *Framing self-sacrifice in the investigation of moral judgment and moral emotions in human and autonomous driving dilemmas [manuscript submitted for publication]*. Department of General Psychology, University of Padua.

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*, 2022–2038.

Coleman, J. S. (1994). *Foundations of social theory*. Harvard University Press.

Contissa, G., Lagioia, F., & Sartor, G. (2017). The ethical knob: Ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, *25*(3), 365–378.

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition*, *179*, 241–265.

Conway, P., & Peetz, J. (2012). When does feeling moral actually make you a better person? Conceptual abstraction moderates whether past moral deeds motivate consistency or compensatory behavior. *Personality and Social Psychology Bulletin*, *38*(7), 907–919.

Cornelissen, G., Bashshur, M. R., Rode, J., & Le Menestrel, M. (2013). Rules or consequences? The role of ethical mind-sets in moral dynamics. *Psychological Science*, *24*(4), 482–488.

De Melo, C. M., Marsella, S., & Gratch, J. (2021). Risk of injury in moral dilemmas with autonomous vehicles. *Frontiers in Robotics and AI*, *7*, 572529.

Di Nucci, E. (2013). Self-sacrifice and the trolley problem. *Philosophical Psychology*, *26*(5), 662–672.

Evans, K., de Moura, N., Chauvier, S., Chatila, R., & Dogan, E. (2020). Ethical decision making in autonomous vehicles: The AV ethics project. *Science and Engineering Ethics*, *26*(6), 3285–3312.

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, *77*, 167–181.

Fanggidae, J. P., Jin, H. S., Kim, H. J., & Kelly, L. (2022). When altruistic decisions shape egoistic motives: Motivation shift in sequential charitable support. *International Journal of Advertising*, *42*, 1110–1143.

Faul, F., & Erdfelder, E. (1992). *GPOWER: A priori, post-hoc, and compromise power analyses for MS-DOS [computer program]*. Bonn University, Department of Psychology.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.

Frohlich, N., Oppenheimer, J. A., & Eavey, C. L. (1987). Choices of principles of distributive justice in experimental groups. *American Journal of Political Science*, *31*, 606–636.

Gaus, G., & Thrasher, J. (2015). *Rational choice and the original position: The (many) models of Rawls and Harsanyi*. Chapman University Digital Commons.

Geisslinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous driving ethics: From trolley problem to ethics of risk. *Philosophy & Technology*, *34*, 1033–1055.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96.

Gigerenzer, G., & Galesic, M. (2012). Why do single event probabilities confuse patients? *BMJ*, *344*, e245.

Gino, F., Norton, M. I., & Weber, R. A. (2016). Motivated Bayesians: Feeling moral while acting egoistically. *Journal of Economic Perspectives*, *30*(3), 189–212.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics*, *23*, 681–700.

Goodall, N. J. (2016). Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, *30*(8), 810–821.

Goodman, L. A. (1961). Snowball sampling. *The Annals of Mathematical Statistics*, *32*, 148–170.

Gorissen, B. L., Yanıkoğlu, İ., & den Hertog, D. (2015). A practical guide to robust optimization. *Omega*, *53*, 124–137.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*(3859), 1243–1248.

Harsanyi, J. C. (1975). Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American Political Science Review*, *69*(2), 594–606.

Harsanyi, J. C. (1978). Bayesian decision theory and utilitarian ethics. *The American Economic Review*, *68*(2), 223–228.

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(48), 23989–23995.

Huebner, B., & Hauser, M. D. (2011). Moral judgments about altruistic self-sacrifice: When philosophical and folk intuitions clash. *Philosophical Psychology*, *24*, 73–94.

Impett, E. A., & Gordon, A. (2008). For the good of others: Toward a positive psychology of sacrifice. *Positive Psychology: Exploring the Best in People*, *2*, 79–100.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In J. Chen & G. Fragomeni (Eds.), *Virtual, augmented and mixed reality. Applications and case studies: 11th International Conference, VAMR 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21* (pp. 476–489). Springer International Publishing.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164.

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. (1785). *Groundwork of the metaphysics of morals*. Yale University Press.

Kaur, K., & Rampersad, G. (2018). Trust in driverless cars: Investigating key factors influencing the adoption of driverless cars. *Journal of Engineering and Technology Management*, *48*, 87–96.

Krug, Z., Guillaume, R., & Battaïa, O. (2020). Decision under ignorance: A comparison of existing criteria. In *International conference on information processing and management of uncertainty in knowledge-based systems* (pp. 158–171). Springer.

Kusev, P., Van Schaik, P., Alzahrani, S., Lonigro, S., & Purser, H. (2016). Judging the morality of utilitarian actions: How poor utilitarian accessibility makes judges irrational. *Psychonomic Bulletin & Review*, *23*(6), 1961–1967.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). *emmeans: Estimated marginal means, aka least-squares means*. (R package, Version 1.4) [Computer software].

Lotto, L., Manfrinati, A., & Sarlo, M. (2014). A new set of moral dilemmas: Norms for moral acceptability, decision times, and emotional salience. *Journal of Behavioral Decision Making*, *27*(1), 57–65.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). *Sacrifice one for the good of many? People apply different moral norms to human and robot agents*. In 2015 10th ACM/IEEE international conference on human-robot interaction (HRI) (pp. 117–124). IEEE.

Martin, R., Kusev, I., Cooke, A. J., Baranova, V., Van Schaik, P., & Kusev, P. (2017). Commentary: The social dilemma of autonomous vehicles. *Frontiers in Psychology*, *8*, 808.

Martin, R., Kusev, P., Teal, J., Baranova, V., & Rigal, B. (2021). Moral decision making: From Bentham to veil of ignorance via perspective taking accessibility. *Behavioral Science*, *11*(5), 66.

Martin, R., Kusev, P., & Van Schaik, P. (2021). Autonomous vehicles: How perspective-taking accessibility alters moral judgments and consumer purchasing behavior. *Cognition*, *212*, 104666.

Martinho, A., Herber, N., Kroesen, M., & Chorus, C. (2021). Ethical issues in focus by the autonomous vehicles industry. *Transport Reviews*, *41*(5), 556–577.

Maxcy, S. J. (2002). *Ethical school leadership*. R&L Education.

Mayer, M. M., Bell, R., & Buchner, A. (2021). Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLoS One*, *16*(12), e0261673.

Mill, J. S. (1861/2004). *Utilitarianism and other essays*. Penguin Books.

Moehler, M. (2018). The Rawls–Harsanyi dispute: A moral point of view. *Pacific Philosophical Quarterly*, *99*(1), 82–99.
Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, *19*(6), 549–557.
Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology*, *67*(1), 363–385.
National Highway Traffic Safety Administration. (2013). *Preliminary statement of policy concerning automated vehicles*. Washington, DC, NHTSA 14-13.
Panagiotopoulos, I., & Dimitrakopoulos, G. (2018). An empirical investigation on consumers' intentions towards autonomous driving. *Transportation Research Part C: Emerging Technologies*, *95*, 773–784.
Parfit, D. (2011). *On what matters*. Oxford University Press.
Parker, C., Scott, S., & Geddes, A. (2019). *Snowball sampling*. SAGE research methods foundations.
Peters, E., Hart, P. S., & Fraenkel, L. (2011). Informing patients: The influence of numeracy, framing, and format of side effect information on risk perceptions. *Medical Decision Making*, *31*(3), 432–436.
Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*, 407–413.
Qian, L., Yin, J., Huang, Y., & Liang, Y. (2023). The role of values and ethics in influencing consumers' intention to use autonomous vehicle hailing services. *Technological Forecasting and Social Change*, *188*, 122267.
R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/
Rawls, J. (1971/2009). *A theory of justice*. Harvard University Press.
Rawls, J. (1973). Some ordinalist-utilitarian notes on Rawls's theory of justice. *The Journal of Philosophy*, *70*(9), 245–263.
Rawls, J. (2001). *Justice as fairness – A restatement (E. Kelly, Ed.)*. Harvard University Press.
Rhim, J., Lee, J. H., Chen, M., & Lim, A. (2021). A deeper look at autonomous vehicle ethics: An integrative ethical decision-making framework to explain moral pluralism. *Frontiers in Robotics and AI*, *8*, 632394.
Schein, C. (2020). The importance of context in moral judgments. *Perspectives on Psychological Science*, *15*(2), 207–215.
Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*(10), 694–696.
Shariff, A., Bonnefon, J. F., & Rahwan, I. (2021). How safe is safe enough? Psychological mechanisms underlying extreme safety demands for self-driving cars. *Transportation Research Part C: Emerging Technologies*, *126*, 103069.
Sterba, J. P. (1989). Toulmin to Rawls. In R. Cavalier, J. S. Gouinlock, & J. P. Sterba (Eds.), *Ethics in the history of western philosophy*. Springer.
Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*, 1395–1415.
Thomson, J. J. (2008). Turning the trolley. *Philosophy and Public Affairs*, *36*, 359–374.
Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463.
Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behaviour*. Princeton University Press.
Wang, B., Rau, P. L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, *42*(9), 1324–1337.
Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R., & Zhang, W. (2019). The roles of initial trust and perceived risk in public's acceptance of automated vehicles. *Transportation Research Part C: Emerging Technologies*, *98*, 207–220.

How to cite this article: Bruno, G., Spoto, A., Sarlo, M., Lotto, L., Marson, A., Cellini, N., & Cutini, S. (2023). Moral reasoning behind the veil of ignorance: An investigation into perspective-taking accessibility in the context of autonomous vehicles. *British Journal of Psychology*, *00*, 1–25. https://doi.org/10.1111/bjop.12679

APPENDIX 1

TABLE A1 Mean and standard deviation of moral acceptance (0 = completely unacceptable) and willingness to buy (0 = unwilling to buy) of the three main consistency profiles (consistent utilitarian, consistent maximin and inconsistent).

		AV behaviour					
		Prioritize the passenger		Minimize the number of casualties		Maximize the protection of the least-advantaged	
	N	Moral acceptance	Willingness to buy	Moral acceptance	Willingness to buy	Moral acceptance	Willingness to buy
Fully utilitarian	83	31.73 (24.35)	50.37 (31.28)	85.69 (14.74)	66.87 (27.12)	46.55 (23.28)	37.60 (25.91)
Fully maximin	45	28.49 (39.07)	39.07 (29.81)	58.51 (28.97)	45.38 (29.87)	64.40 (22.18)	46.42 (25.17)
Inconsistent	108	45.87 (25.63)	57.33 (28.26)	77.54 (19.74)	55.87 (29.02)	57.55 (27.17)	46.34 (26.12)
Overall	239	37.88 (26.04)	51.73 (30.32)	76.77 (22.33)	57.77 (29.33)	54.73 (25.72)	43.27 (26.02)

Note: The 'consistent non-utilitarian profile' was not added to the profile for the scarce numerosity ($n = 3$). The overall information is presented in the last row.

TABLE A2 Mean and standard deviation of moral acceptance (0 = completely unacceptable) and willingness to buy (0 = unwilling to buy) of the three main consistency profiles (consistent utilitarian, consistent maximin and inconsistent).

		AV's behaviour					
		Prioritize the passenger		Minimize the number of casualties		Maximize the protection of the least-advantaged	
	N	Moral acceptance	Willingness to buy	Moral acceptance	Willingness to buy	Moral acceptance	Willingness to buy
Fully utilitarian	14	32.3 (26.5)	45.6 (41.1)	88.4 (12.4)	55.1 (38.0)	45.6 (18.0)	35.1 (24.3)
Mainly utilitarian	7	50.0 (35.1)	54.0 (31.9)	69.1 (34.3)	50.4 (23.4)	45.7 (30.5)	33.3 (29.8)
Fully maximin	12	15.4 (17.60)	45.8 (40.8)	58.8 (35.4)	46.9 (41.3)	62.6 (36.1)	45.6 (38.4)
Mainly maximin	14	25.4 (26.5)	21.3 (26.6)	70.1 (25.8)	29.1 (36.0)	66.2 (23.1)	34.1 (31.8)
Inconsistent	29	50.5 (21.0)	51.9 (28.4)	71.1 (22.5)	60.9 (29.3)	53.8 (27.0)	36.6 (23.8)
Overall	76	35.2 (26.4)	44.3 (34.1)	72.3 (26.1)	50.9 (34.9)	55.1 (27.1)	39.9 (27.9)

Note: Compared to Study 1, two further profiles were added ('mainly utilitarian' and 'mainly maximin'), assuming at least six moral judgements coherent with the corresponding moral code. Overall information is presented in the last row.