Motivation and Emotion (2023) 47:781–794
https://doi.org/10.1007/s11031-023-10024-3

ORIGINAL PAPER

Framing self-sacrifice in the investigation of moral judgment and moral emotions in human and autonomous driving dilemmas

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

Accepted: 25 April 2023 / Published online: 4 May 2023

Abstract

In the investigation of moral judgments of autonomous vehicles (AVs), the paradigm of the sacrificial dilemma is a widespread and flexible experimental tool. In this context, the sacrifice of the AV's passenger typically occurs upon enactment of the utilitarian option, which differs from traditional sacrificial dilemmas, in which the moral agent's life is often jeopardized in the non-utilitarian counterpart. The present within-subject study (n = 183) is aimed at deepening the role of self-sacrifice framing, comparing autonomous- and human-driving text-based moral dilemmas in terms of moral judgment and intensity of four moral emotions (shame, guilt, anger, and disgust). A higher endorsement of utilitarian behavior was observed in human-driving dilemmas and for self-protective utilitarian behaviors. Interestingly, the utilitarian option was considered less moral, shameful, and blameworthy in the case of concurrent self-sacrifice. The present study collects novel information on how different levels of driving automation shape moral judgment and emotions, also providing new evidence on the role of self-sacrifice framing in moral dilemmas.

Keywords Sacrifice framing · Moral dilemma · Moral emotions · Utilitarianism · Driving behavior

In the past two decades, the growing emphasis on the behavior of autonomous vehicles (AVs) has had an important influence on moral judgment investigation. Interest in this technology escalated quickly with the vision of the upcoming revolution in transportation, which will most likely induce a gradual reduction of human decision-making abilities in driving (Society of Automotive Engineers [SAE], 2021). Researchers have made arguments regarding the advantages and challenges of autonomous transportation (Elliott et al., 2019; Fagnant & Kockelman, 2015; Martinez-Díaz & Soriguera, 2018; Maurer et al., 2016; Rahwan et al., 2019), and several studies have focused on individuals' attitudes and motivation regarding this new technology (Haboucha et al., 2017; Jobin et al., 2019; Othman, 2021). Among these factors, the perceived risk of losing control of driving operations is considered a significant psychological roadblock to the adoption of this technology (Shariff et al., 2017), especially considering how AVs will manage possible harm in dangerous situations. This obstruction to the full endorsement of AVs has been treated as a moral issue that overcomes autonomous transportation's reported advantages (Meyer et al., 2017), and researchers have explored the ethical concerns regarding human-independent decisions (Hidalgo, 2021).

Based on its suitability, the sacrificial dilemma has been widely considered a flexible tool in the investigation of moral judgments of AVs' behavior (Martí-Vilar et al., 2021; Unger, 1996). Traditionally, moral dilemmas have been used to compare two opposing moral doctrines: utilitarianism, aimed at minimizing harmful consequences (Bentham, 1781), and deontologism, aimed at adhering to categorical norms and duties (e.g., "The ends never justify the means"; Kant 1785). The trolley problem is the most popular

✉ Giovanni Bruno
giovanni.bruno.2@unipd.it

[1] Department of General Psychology, University of Padua, Via Venezia 8, 35131 Padova, Italy

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

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

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

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

example of a sacrificial dilemma (Foot, 1978), depicting a brakeless running trolley that is about to run over five track workers. The only way to save the workers is by pulling a lever that will divert the trolley onto a secondary track, where only one worker will be sacrificed. Trolley-like problems are described as *incidental* (Lotto et al., 2014), where harm is permissible only as a foreseen but unintended side effect to save the largest number of people, on the basis of Aquinas's (1952) doctrine of the double effect. On the contrary, when harm is intentionally caused in pursuit of a greater good, the dilemma is defined as *instrumental*, an example of which is the footbridge problem (Thomson, 1985). Here, the utilitarian moral code is respected only by physically pushing a large man off an overpass to stop the trolley with his body. In this context, the unwillingness to endorse the utilitarian behavior is traditionally interpreted as non-utilitarian choice, reflecting the prohibition of a personal moral violation despite its utilitarian value (Gleichgerrcht & Young, 2013; Greene et al., 2004; Cushman et al., 2012). In the dual-process model framework (Greene et al., 2001, 2008), intentional and personal harm typically elicit non-utilitarian judgments as a consequence of a stronger emotional response (e.g., the footbridge problem).

Incidental and instrumental moral dilemmas can be structured as sacrificial dilemmas, in which at least one life must be sacrificed to fulfill the selected moral code (Kahane 2015). In time, sacrificial dilemmas have been applied to a wide variety of barely realistic circumstances (e.g., Greene et al., 2001; Moore et al., 2008), leading to questioning their ecological validity (Bauman et al., 2014; Gold et al., 2014; Kahane et al., 2018). Additionally, traditional dilemma sets have been developed assuming moral judgment to follow a 'structure-based' interpretation rule (e.g., incidental vs. instrumental), which results in downplaying the importance of contextualization (Schein, 2020). Although a 'structure-based' moral reasoning is widely endorsed in the relevant literature (Greene et al., 2001; Lotto et al., 2014; Moore et al., 2008), several studies demonstrated that more plausible and lifelike storylines may actually prompt utilitarian moral reasoning (Bruno et al., 2022; Körner et al., 2019).

Sacrificial dilemmas' reliability has been intensively debated in recent years (Bartels et al., 2014, Bauman et al., 2014), highlighting scenarios in which characters' lives are in jeopardy. Traditionally, when a sacrificial act saves the life of the moral agent or those of other people (self-involvement dilemmas), the preference for self-protection has always been assumed as part of the utilitarian option (e.g., "Should you kill this man to save yourself and the other five people?"; *The burning building dilemma*; see Moore et al., 2008).

Utilitarian judgment of AVs' moral behavior

The experimental deployment of sacrificial dilemmas has grown exponentially in recent years because of the intensive investigation of the cognitive and emotional basis of human morality (Awad et al., 2020; Cushman et al., 2006; Greene et al., 2001, 2008; Sarlo et al., 2012). In this framework, applying the sacrificial and self-involving version of the trolley problem in the driving context seems straightforward (Bruno et al., 2022), and in numerous studies, researchers have opted to adopt its basic structure—specifically readapted—following different approaches. Indeed, this version of a moral dilemma has been used to investigate the association between morality and AVs' behavior in immersive driving simulators (Frison et al., 2016; Samuel et al., 2020) in virtual reality (VR) settings (Faulhaber et al., 2019; Kallioinen et al., 2019; Riegler et al., 2021; Sütfeld et al., 2019, 2017) and—mainly—responding to image- or text-based moral scenarios (Bonnefon et al., 2016; Huang et al., 2019; (Martin et al., 2021a). Among some variations (e.g., number of characters involved, decision maker's perspective, and risk level), the autonomous version of the trolley problem typically proposes an AV that is driving n passenger(s) and unexpectedly encounters m pedestrians who are crossing the road ($n < m$). The only way to avoid the accident—following the utilitarian moral code—is for the AV to steer off the street suddenly, sacrificing its own passenger(s). Traditionally, the alternative solution—always interpreted as non-utilitarian—is for the AV to continue straight, sacrificing the pedestrians and protecting its passenger(s). Usually, in the autonomous-driving version of the sacrificial dilemma, the sacrifice of the moral agent as the AV's passenger occurs upon enactment of the proactive utilitarian option (e.g., swerving off to the side of the road, where the car will impact a barrier, killing the passenger/s but leaving the pedestrians unharmed; Bonnefon et al., 2016), which differs from traditional sacrificial dilemmas, in which the agent's life is usually jeopardized in the non-utilitarian counterpart (Greene et al., 2001, 2004; Lotto et al., 2014; Moore et al., 2008).

In the development of AVs' moral investigation through text-based dilemmas, the seminal work of Bonnefon et al. (2016) lit the fuse. The authors observed that the support of utilitarian AVs for the minimization of overall harm had decreased at the individual level, where participants preferred self-protective AVs for themselves. Importantly, the Moral Machine project (Awad et al., 2018) massively investigated moral preferences regarding AVs from a global and cross-cultural perspective, confirming the preference for sparing the largest possible number of lives and assessing the moderating role of other important factors (e.g., age of lives, whether lives are human, and whether road users' behavior is lawful). De Melo et al. (2021) demonstrated the mediation effect of total perceived risk and other drivers' behavior on the likelihood of utilitarian AV maneuvers, confirming the role of probabilistic outcomes in the resolution of moral dilemmas (Bazerman & Greene, 2010). Moreover, a limited number of studies have confirmed that decision-makers' perspectives effectively influence moral judgment (Huang et al., 2019; Kallionen et al., 2019, Mayer et al., 2021). Interestingly, following the *veil of ignorance* (Rawls, 2009) reasoning seems to favor utilitarian resolutions, decreasing the inconsistency between moral judgment and willingness to buy utilitarian AVs (Martin et al., 2021a, b).

Despite the relevant attention directed toward the investigation of moral judgments related to autonomous technology, very few studies have focused on traditional human driving or the comparison with its autonomous counterpart. Bruno et al. (2022) detected relative ease in the endorsement of utilitarian behavior in human-driving dilemmas when compared to traditional sacrificial scenarios (e.g., Lotto et al., 2014), interpreting the plausibility of the on-road context as a facilitator in moral judgments. Li et al. (2016) recognized the utilitarian moral code as the default norm for human and autonomous driving vehicles, with a stronger expectation of utilitarianism in autonomous agents (Malle et al., 2015) but ascribing less responsibility to them in the case of mistakes. This utilitarian expectation of AV behavior has also been recently confirmed in VR settings (Kallionen et al., 2019). In terms of attribution of moral responsibility, McManus and Rutchick (2019) detected a positive relationship between agency and blame: In the case of negative consequences, AVs are considered less blameworthy than human drivers—because they cannot act deliberatively (Pizarro et al., 2003). Nonetheless, evidence on the allocation of moral responsibility to Artificial Intelligence (AI) systems still appears somewhat contrasting. Hong et al. (2020) observed a higher level of blaming towards AI drivers than human drivers, proportionate to the severity of the damage. Also, Bennett et al. (2020) found an inverse relationship between driving automation and human blaming in case of a road accident (at the expense of AV manufacturers), even though the attribution of final responsibility being mainly addressed to human drivers. Gill (2021) also investigated this topic, claiming a reduction in direct human responsibility when aboard an AV. Data confirmed this hypothesis, showing a reduction of frequency and moral permissibility of self-sacrifice solutions in a one-to-one AV dilemma, compared to human-driving dilemmas. Despite these preliminary results, the roles of moral responsibility and agency in autonomous and human-driving sacrificial dilemmas require further investigation, considering prior knowledge about the intentionality of harm (Cushman et al., 2006; Greene et al., 2004).

Emotions and moral judgment

Altogether, several individual and contextual factors are involved in moral judgment. Among them, emotion is probably one of the most widely discussed factors (Haidt, 2001; McHugh et al., 2022), despite its role in the mechanisms of moral cognition is still debated (Byrd & Conway, 2019; Crockett, 2013; Cushman, 2013; Greene et al., 2016) and remains somewhat unclear (Huebner et al., 2009; Landy & Goodwin, 2015). In the evaluation of a moral issue, individual moral behavior is insidiously influenced by moral emotions, which serve as mediators between individual moral principles (i.e., norms and conventions) and moral decisions (Tangney et al., 2007). Moral emotions arise because of daily events that motivate people to engage in—or avoid—righteous or wrong moral actions (Kroll & Egan, 2004), linked to the interest of one or more individuals other than themselves (Greenbaum et al., 2020; Haidt, 2003). Importantly, moral emotions can affect a moral agent before the actuation of the decision—during the evaluation of the potential alternatives (Tangney et al., 2007).

Moral emotions can be divided into two main categories: self-conscious emotions (e.g., shame and guilt) and other-condemning emotions (e.g., anger and disgust; Haidt 2003). In the first case, negative feelings are directed to the self in the form of self-evaluation in the violation of moral standards (Gehm & Scherer, 1988; Tangney & Dearing, 2002). On the one hand, shame is conceived as a public-oriented emotion ("What others will think of me?"; Buss 1980) and involves a global negative evaluation of the inner self and the event as objects of disapproval. On the other hand, guilt is perceived as a private-oriented emotion ("I did a bad thing") condemning the negative behavior only and not the self as a whole (Lewis, 1971). Shame seems to be a more powerful moral experience than guilt (Behrendt & Ben-Ari, 2012), and empirical evidence suggests that although shame leads to defensiveness and distance, guilt promotes constructive responses (Tangney et al., 2007). In contrast, anger and disgust are the two main other-condemning moral emotions, depicted as negative feelings in response to a third party's moral violation. Anger takes the form of indignation regarding mistreatments of and injustices affecting the self or others, spawning an immediate response against the immoral event (Hutcherson & Gross, 2011). In contrast, disgust represents a repulsion that arises against filthy moral conduct and seemingly correlates with the moral judgment's severity (Huebner et al., 2009; Rozin et al., 1999; Schnall et al., 2008). These two emotions are elicited by different cues in the moral context (Russell & Giner-Sorolla, 2011), and like self-conscious emotions, they are experienced when the moral decision is made and when it is observed (Haidt, 2003). Referring to self-referred and other-referred moral emotions may be useful in disentangling moral judgments of AVs' behavior, assuming the intrinsic difference between a traditional human-driven (hands-on-the-wheel) vehicle or an autonomous (hands-off-the-wheel) vehicle and when considering how the decrease in agency induces less responsibility attribution to and blaming of the AV's harmful actions (Malle et al., 2014; McManus & Rutchick, 2019).

The present study

In the field of moral psychology, the investigation of the moral perception of AV technology is typically investigated through sacrificial moral dilemmas. In a typical sacrificial AV dilemma, when the moral agent's life is at stake, the endorsement of the utilitarian option corresponds to the acceptance of the moral agent's own sacrifice (i.e., "I die, but many survive"). This is an important structural difference from traditional self-involved sacrificial dilemmas (e.g., Greene 2001; Lotto et al., 2014), in which the endorsement of the utilitarian behavior matches the self-protective choice (i.e., "I live, and many survive"). In this context, we investigated the potential role of the agent's sacrifice in the acceptance of the AV's utilitarian behavior, by testing the following hypothesis:

> H1: When the utilitarian moral behavior leads to the moral agent's self-sacrifice (i.e., "I die, and many survive"), we expect a lower endorsement of the utilitarian behavior and a reduction of its moral acceptability, as compared to life-saving sacrificial scenarios (i.e., "I live, and many survive").

Furthermore, the comparison between traditional human driving and autonomous driving has been often overlooked in the literature, especially when using text-based moral dilemmas. Several behavioral studies have investigated general attitudes, emotional activation, and moral perception in the evaluation of sacrificial AV dilemmas (e.g., Bonnefon et al., 2016), but no evidence has been collected in terms of differences from traditional manual driving. In this study, we aimed to compare moral judgments in moral dilemmas applied to these two fundamentally different modes of transportation, with the formulation of the following hypothesis:

> H2: In line with the role played by lifelike storylines in the enhancement of utilitarian moral reasoning (Bruno et al., 2022; Körner et al., 2019), we predict a higher endorsement of utilitarian behavior in the case of human-driving moral dilemmas. At the same time, since the driving dilemmas share the same incidental structure – and considering the evidence in favor of a 'structure-based' interpretation of moral dilemmas - we expect the endorsement of the utilitarian moral code to be considerably high in both autonomous- and human-driving scenarios.

Finally, for the first time in the context of autonomous and non-autonomous transportation, we investigated the role of two self-referred (shame and guilt) and two other-referred (anger and disgust) moral emotions. In this context, we formulated the following hypotheses:

> H3: A higher intensity of self-referred emotions will be reported after moral decisions concerning traditional human-driven vehicles than after those involving autonomous vehicles.

> H4: A higher intensity of other-referred emotions will be reported after moral decisions concerning autonomous vehicles than after those involving traditional human-driven vehicles.

> H5: Regardless of the level of automation, a higher intensity of self-referred emotions will be reported after the endorsement of self-protective behaviors, and a higher intensity of other-referred emotions will be reported after the endorsement of self-sacrificial behaviors.

Methods

Participants

Before analyzing any data, we tested a baseline equation, assuming a small effect size (Cohen's $d = 0.10$) and a correlation of 0.50 among repeated measures, with a bidirectional hypothesis and an alpha error probability of 0.05 with 0.90 power, calculated with the G*Power statistical software (Faul & Erdfelder, 1992). The analysis suggested a minimum of 140 subjects, and we recruited 183 participants (94 women). Their mean age was 27.82 years ($SD = 10.55$, range: 18–66), and their mean education duration was 16.7 years of formal schooling ($SD = 2.03$, range: 11–24). Of the participants, 65.22% ($n = 120$) were enrolled in university courses, with 50.55% ($n = 93$) enrolled in human sciences degree programs (e.g., psychology or sociology). Most of the participants (90.21%, $n = 166$) held driver's licenses, and most of them (75.54%, $n = 139$) drove a maximum of 15,000 km per year. Half of the sample (48.91%, $n = 90$) had been involved in at least one car accident in their lives, and only 4.35% ($n = 8$) had had a collision in the past 12 months.

Of the participants, 87.5% (n = 161) stated that they had heard about AVs. Additionally, we administered the Positive and Negative Affect Schedule scale (PANAS) to assess the participants' positive affect and negative affect during the 7 days leading up to the survey date (Terracciano et al., 2003; Watson et al., 1988). The participants had a mean positive-affect score of 31 (SD = 7.20) and a mean negative-affect score of 22.80 (SD = 8.03), below the normative thresholds and with no differences between men and women.

Stimuli

Following the structure of the validated set of Lotto et al. (2014), we designed 12 self-involvement sacrificial moral scenarios specifically for this study, adapted from the sacrificial human-driving set by Bruno et al. (2022). Coherently with the previous studies and based on the doctrine of double effect (Aquinas, 1952), we structured our dilemmas as incidental, interpreting the sacrifice as a predicted — albeit undesired — consequence that is mandatory to protect the highest number of lives. Specifically, the dilemma set comprised 12 moral scenarios (see Table 1 for samples) and is included in the supplementary material (https://osf.io/pb3xc/?view_only=4ae203cc39e24d68859da3f6b67591a5). The study was structured as a two-by-two repeated measures factorial design. In six of the scenarios, the participant was requested to identify with the actual driver of a traditional human-driven vehicle whereas in the remaining six, they had to identify with a passenger of a completely autonomous and self-driving car (Level 5 of automated driving; SAE, 2021). In the traditional car storyline, the participant alone was in charge of the driving decision, whereas in the AV scenarios they had to go with the vehicle's decision.

In each dilemma, the driver or AV faced a particular traffic situation (e.g., overtaking a slow vehicle) with the participant and one other passenger on board. Suddenly, a critical problem arose from an unpredicted event, forcing the driver or AV to choose between the passengers' safety and that of a larger number of pedestrians. We presented each moral dilemma as a textual description of the situation and two possible solutions: a utilitarian behavior and a non-utilitarian behavior. Consistently with the utilitarian doctrine, opting for the utilitarian maneuver always resulted in an active action (i.e., steering) aiming at safeguarding the highest number of characters and so minimizing the overall harmful consequences. Oppositely, the endorsement of the non-utilitarian option resulted in holding the current track, rejecting harm as an unintended side effect of collective welfare (Bruno et al., 2022; Cushman et al., 2012; Lotto et al., 2014).

To investigate the life-threatening factor's effect on the moral agents' endorsement of the utilitarian code, we further framed the dilemma set for this factor. With this aim, three scenarios per driving typology (n = 6) depicted the moral

Table 1 Sample Autonomous and Human-Driving Dilemmas (Text Translated from Italian)

Dilemma	Scenario	Outcomes
Human-driving Utilitarian Sacrifice Framing (Human-USF, n = 3)	You are driving a car with a passenger on a two-lane highway. It's late night, and there is only a car coming in the opposite direction. Suddenly you notice a small van on the side of the road, and 4 workers a few meters from you in the middle of the road, dealing with road maintenance work. You begin to slow down when you realize that the brakes are not working.	A. You let the car proceed straight, running over the four workers, who will die. B. You suddenly steer left. The four workers will be unhurt, but your car will crash against the guardrail, where you and your passenger will die.
Human-driving Non-utilitarian Sacrifice Framing (Human-NSF, n = 3)	You are a taxi driver; it's night and you are driving a passenger. As in the last nights, a thick fog has descended on your city and the visibility is strongly compromised. You can notice two pedestrians on the right sidewalk. Suddenly you notice two cyclists crossing the road right in front of you. Because of the thick fog, you did not notice him and now there is no more time to brake.	A. You let the car proceed straight, running over the two cyclists, who will die. Your taxi will swerve crushing against a building, and you and your passenger will die. B. You suddenly steer left, running over the two pedestrians on the sidewalk, who will die, but you, your passenger, and the two cyclists will be saved.
Autonomous driving Utilitarian Sacrifice Framing (AV-USF, n = 3)	You and another person are the passengers of a fully autonomous vehicle, driving on a tree-lined avenue. A truck is proceeding in front of you, which is now slowing down for no apparent reason. The road lanes are separated by a dotted line, so you decide to overtake them. During the overtaking, four runners suddenly cross the road appearing from behind the truck. There is no more time to brake. The autonomous vehicle did not perceive them in time, and now there is no more time to brake.	A. Proceed straight, running over the four runners, who will die. B. Suddenly steer to the left. The four runners will be unhurt, but the autonomous vehicle will crash against a big tree, where you and the other passenger will die.
Autonomous driving Non-utilitarian Sacrifice Framing (AV-NSF, n = 3)	You and another person are the passengers of a fully autonomous taxi vehicle. A violent storm has hit your city for a few hours, it is still raining, and the visibility is strongly compromised. You can notice two pedestrians on the right sidewalk. Suddenly two cyclists appear from the right, now standing in the middle of the road. The autonomous vehicle did not perceive them in time, and now there is no more time to brake.	A. Proceed straight, running over the two cyclists, who will die. The autonomous vehicle will swerve crushing against a streetlamp, and you and the other passenger will die. B. Suddenly steer left, running over the two pedestrians on the sidewalk, who will die, but you, the other passenger, and the two cyclists will be saved.

agent's sacrifice in the utilitarian outcome ("I die, but many survive"; utilitarian sacrifice framing [USF]) whereas in the remaining six dilemmas, the non-utilitarian act involved the moral agent's sacrifice ("I die, and many die"; non-utilitarian sacrifice framing [NSF]).

We controlled all 12 of the driving dilemmas for a number of factors that may play a role in the moral decision-making process (Awad et al., 2018; Bruno et al., 2022). We maintained a 1:2 ratio between lives saved and lives sacrificed, and we provided no additional information about or characterization of the other road users. To avoid a mediated allocation of responsibility, we avoided mentioning traffic rule violations and using leading language. Furthermore, we adopted scrupulous control of the number of words used to avoid intraindividual differences in reading times. The mean reading time was 31.31 s (SD = 63.75), and the mean decision time was 9.45 s (SD = 16.51).

Experimental procedure

Each participant signed an informed consent form before participation, which was voluntary and unremunerated. The local ethics committee approved the study (ID number 3514). We programmed and distributed the task as an online survey via Qualtrics software (Qualtrics, Provo, UT). To avoid device compatibility issues (Krebs & Höhne, 2021), we required the participants to complete the experiment using a laptop or computer.

At the beginning of the task, the participant had the opportunity to read and sign the informed consent form and then read the experimental procedure and instructions. Following the administration of the PANAS, the participants received a careful explanation of the dilemma presentation mode. At this point, the 12 moral scenarios were randomly administered to the participants. For each of them, the textual description remained on the screen for the entire time the participant needed to understand the situation. Then, the first alternative, randomly selected between the two, appeared on the screen. After 7 s, the second moral option appeared on the screen. Finally, after 7 more seconds, the participant selected their morally preferred outcome (see Bruno et al., 2022). After each moral dilemma, the participant had to evaluate the moral acceptability of the two proposed moral outcomes on an 8-point Likert scale (0 = *completely immoral*, 7 = *completely moral*). Subsequently, the participant had to rate the perceived intensity of the four moral emotions, two of which were self-referred (shame and guilt) and two of which were other-referred (anger and disgust). The evaluation referred to the present time, after the moral decision, and was expressed on an 8-point Likert scale (0 = *no intensity*, 7 = *maximum intensity*) (Fig. 1).

Statistical analysis

We conducted the statistical analysis in the R environment (version 4.1.1). As a first step, we preliminarily tested data distributions for the variables of interest (fitdistrplus package in R; Delignette-Muller & Dutang 2015). To test the experimental hypothesis, we set the moral decision choices, the moral acceptability (for the utilitarian and non-utilitarian options), and the four moral emotions (shame, guilt, anger, and disgust) as dependent variables for the statistical investigation. We coherently fit one generalized mixed-effect linear model—for the dichotomous moral decision—and six linear mixed models to the data, setting the participants as random intercepts (lme4 package in R; Bates et al., 2015). We selected the models after a specific onward stepwise model selection procedure, considering in each case all the predictors of interest (moral decision, sacrifice framing, driving style, experimental order, gender, and car accidents experienced in the past) and their interactions, in the described order. Final models were selected accordingly to the Akaike Weights comparison procedure (Wagenmakers & Farrell, 2004). When needed, we conducted post hoc pairwise comparisons to investigate interlevel differences (emmeans package in R; Length, 2020). Tables 2 and 3 present the statistical analysis's descriptive results. In the Results section and for each of the seven models, all the selected predictors have been listed coherently with their order of implementation. We include full, detailed predictors based on the seven model selection procedures as well as the complete data set and further detailed information on our statistical approach in the supplementary materials.

Results

The binomial distribution was set for implementing the generalized mixed linear model *m1* on moral decisions (utilitarian, non-utilitarian). Following the model comparison, we set driving style (autonomous, manual), sacrifice framing (USF, NSF), and gender (female, male) as fixed effects, together with the interaction between driving style and sacrifice framing. Consistently with H1, when we framed the agent's sacrifice in the non-utilitarian option (NSF), we observed a greater frequency of utilitarian behavior ($\chi^2_1 = 88.97$, $p < .001$). As expected in H2, we observed a slightly higher endorsement of the utilitarian behavior in the human-driving condition (78.71%) than with the autonomous version (74.59%; $\chi^2_1 = 6.78$, $p = .009$). Descriptively, in the USF, if an autonomous vehicle performed it, the utilitarian maneuver was selected less often (Fig. 2). Nonetheless, we observed no significant interaction between the two factors ($p = .14$).

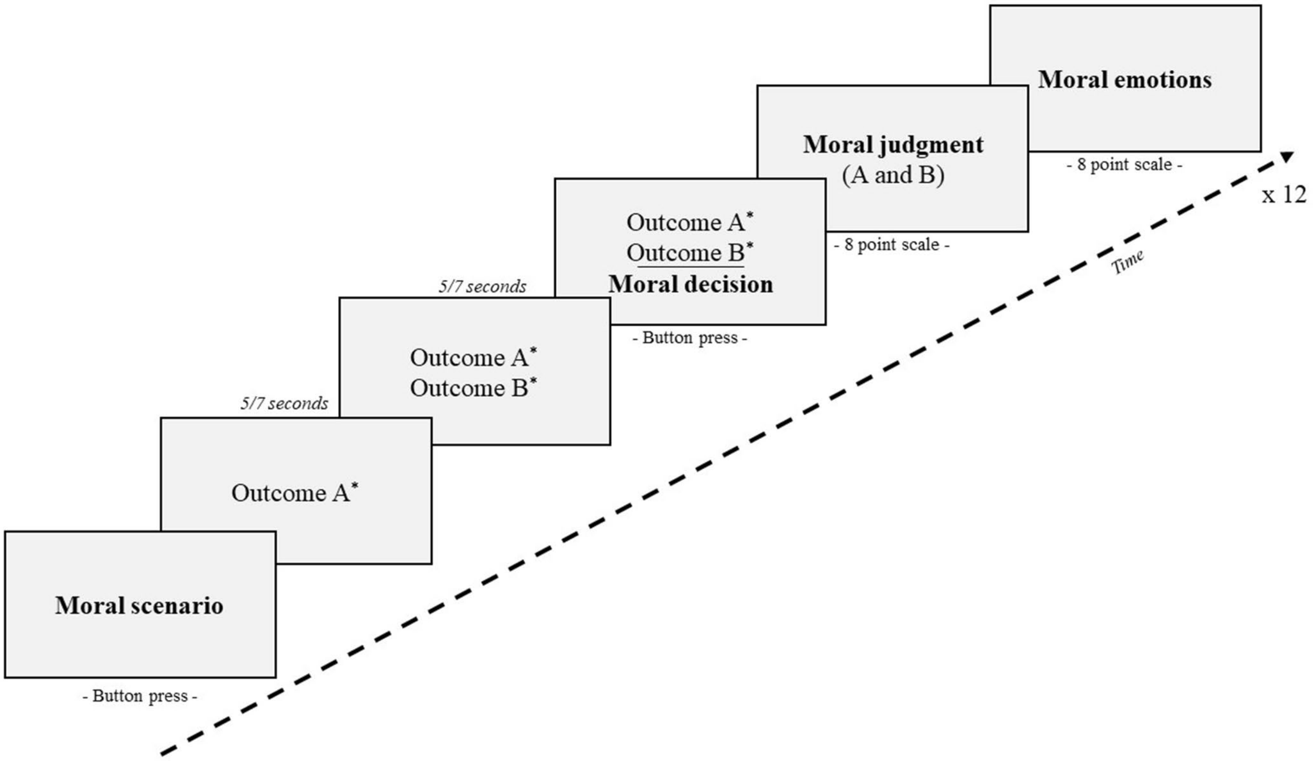
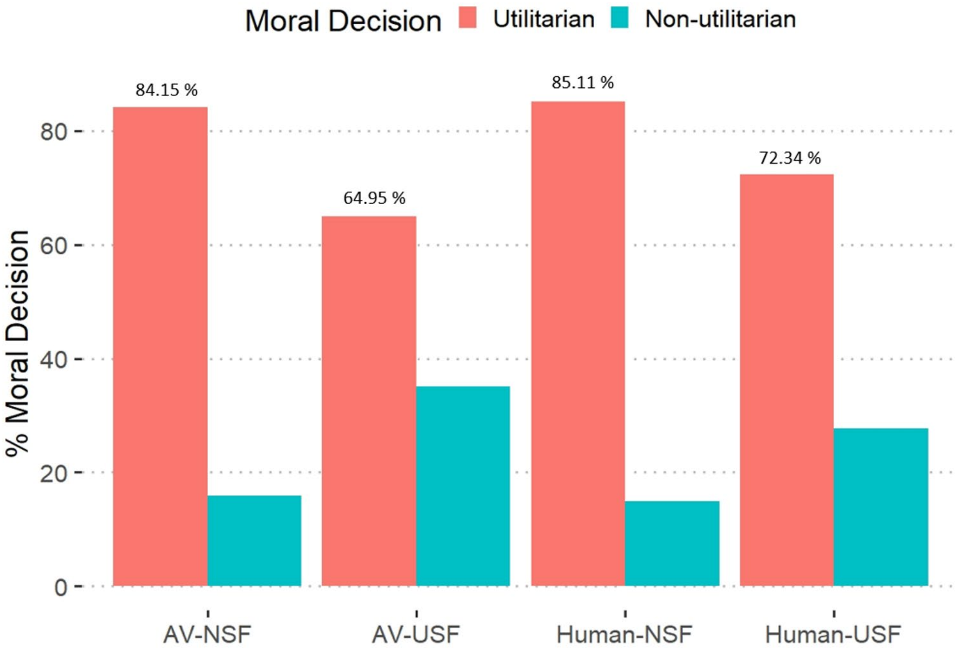


Fig. 1 The Experimental Procedure of the Present Study. The Sequence Was Repeated 12 Times, One per Each Administered Dilemma. The Presentation of the Two Outcomes (A and B) was Randomized Between-Subject

Fig. 2 Bar Chart of Moral Decision Percentage Frequencies, Divided by Driving Style (AV: Autonomous-Driving; Human: Human-Driving) and Sacrifice Framing (NSF: Non-utilitarian Sacrifice Framing, USF: Utilitarian Sacrifice Framing)



Focusing on the evaluated moral acceptability of the two proposed options (utilitarian and non-utilitarian), we fitted two linear mixed models to the data setting the moral acceptability of the utilitarian option (*m2*) and the moral acceptability of the non-utilitarian option (*m3*) as dependent variables. In both cases, the resulting final models considered the fixed effects of driving style, moral decision, sacrifice framing, experimental order, and gender, together with

Table 2 Mean and Standard Deviation (between brackets) of the Considered Dependent Variables, Divided by Driving Style (Human, Autonomous) and Sacrifice Framing (NSF: Non-utilitarian Sacrifice Framing, USF: Utilitarian Sacrifice Framing). For Each Dilemma Category, the Total Percentage of Utilitarian Decisions is Reported in the First Row

	Human-Driving	A-Driving	USF	NSF
Moral Decision: Utilitarian (%)	78.71	74.59	68.65	84.63
Moral Acceptance: Utilitarian	2.41 (2.01)	2.28 (2.00)	2.74 (2.04)	1.96 (1.90)
Moral Acceptance: Non-utilitarian	1.32 (1.62)	1.45 (1.79)	1.43 (1.74)	1.35 (1.66)
Shame (Self-Referred)	2.43 (2.37)	2.22 (2.28)	2.09 (2.25)	2.56 (2.38)
Guilt (Self-Referred)	3.51 (2.47)	3.18 (2.48)	3.08 (2.45)	3.61 (2.48)
Anger (Other-Referred)	3.59 (2.45)	3.64 (2.50)	3.66 (2.48)	3.57 (2.47)
Disgust (Other-Referred)	2.41 (2.41)	2.55 (2.50)	2.44 (2.45)	2.52 (2.46)

Table 3 Mean and Standard Deviation (between brackets) of the Considered Dependent Variables, Divided by Driving Style (Human, Autonomous) and Sacrifice Framing (NSF: Non-utilitarian Sacrifice Framing, USF: Utilitarian Sacrifice Framing). For Each Dilemma Category, the Total Percentage of Utilitarian Decisions is Reported in the First Row

	Human-USF	Human-NSF	AV-USF	AV-NSF
Moral Decision: Utilitarian (%)	72.34	85.11	64.95	84.15
Moral Acceptance: Utilitarian	2.85 (2.05)	1.98 (1.88)	2.63 (2.02)	1.94 (1.92)
Moral Acceptance: Non-utilitarian	1.33 (1.54)	1.32 (1.68)	1.52 (1.92)	1.38 (1.64)
Shame (Self-Referred)	2.16 (2.30)	2.71 (2.28)	2.02 (2.20)	2.42 (2.34)
Guilt (Self-Referred)	3.20 (2.44)	3.82 (2.46)	2.95 (2.46)	3.40 (2.48)
Anger (Other-Referred)	3.64 (2.51)	3.55 (2.46)	3.68 (2.51)	3.59 (2.49)
Disgust (Other-Referred)	2.34 (2.39)	2.48 (2.43)	2.55 (2.51)	2.56 (2.50)

the interaction between moral decision and driving style. Participants rated the moral acceptability of both alternatives as low (Table 2), with lower scores for the non-utilitarian option. As expected, in the case of endorsement of the utilitarian behavior, we found the non-utilitarian option was less acceptable than its counterpart ($\chi^2_1 = 162.48$, $p < .001$; $\bar{x}_{\text{non-utilitarian}} = 2.02$; $\bar{x}_{\text{utilitarian}} = 2.45$). We observed no statistical effects of the driving style ($p = .10$) or the interaction with the moral decision ($p = .78$), but we detected a significant reduction in moral acceptability during the course of the experiment ($\chi^2_{11} = 43.53$, $p < .001$). Interestingly, the participants rated the utilitarian behavior as more acceptable when it was coupled with the agent's self-sacrifice ($\bar{x}$ = 2.74, Fig. 3) than the self-protective utilitarian option ($\bar{x}$ = 1.96; $\chi^2_1 = 296.98$, $p < .001$). We detected no significant effects when we set the non-utilitarian behavior as the dependent variable (*m3*).

Subsequently, we fitted four linear mixed models setting the two self-referred (shame and guilt) and the two other-referred moral emotions (anger and disgust) as dependent variables. In both the self-referred emotions cases (shame *m4* and guilt *m5*), the model comparison procedure indicated to consider the models including moral decision, driving style, sacrifice framing, experimental order, and gender as fixed effects, as well as the interactions between (i) moral decision and driving style and (ii) moral decision and sacrifice framing. Interestingly, the results showed a similar trend between the two self-referred emotions after the moral decision, with overall consistently lower scores in the case of shame (see Tables 2 and 3). As expected in H3 on the intensity of self-referred moral emotions between levels of automation, shame and guilt were perceived as less intense after decisions in AV dilemmas (*m4*: $\chi^2_1 = 23.43$, $p < .001$; *m5*: $\chi^2_1 = 44.84$, $p < .001$), in USF dilemmas (*m4*: $\chi^2_1 = 106.85$, $p < .001$; *m5*: $\chi^2_1 = 113.67$, $p < .001$), and following the utilitarian moral decision (*m4*: $\chi^2_1 = 52.91$, $p < .001$; *m5*: $\chi^2_1 = 36.05$, $p < .001$). With reference to the hypothesized difference of intensity of self-referred moral emotions in presence of the endorsement of self-protective behaviors (H5), the moral decision showed a significant interaction with the sacrifice framing factor (USF, NSF), highlighting greater intensities of shame and guilt when one pursues self-protection, compared to the self-sacrificial option (*m4*: $\chi^2_1 = 91.99$, $p < .001$; *m5*: $\chi^2_1 = 104.67$, $p < .001$; Fig. 3). Additionally, men reported significantly lower emotion intensities than women in terms of shame (men: $\bar{x}$ = 1.78; women: $\bar{x}$ = 2.84) and guilt (men: $\bar{x}$ = 2.65; women: $\bar{x}$ = 4.00).

Finally, in both the other-referred moral emotion cases (anger *m6* and disgust *m7*), the model comparison procedure indicated to consider the models including moral decision, driving style, sacrifice framing, and experimental order as fixed effects, as well as the interactions between (i) moral decision and driving style and (ii) moral decision and sacrifice framing.

We observed no effects in the anger *m6* model, with the exception of a significant interaction between sacrifice framing and moral decision ($\chi^2_1 = 13.65$, $p < .001$), showing greater anger intensity when the moral decision corresponded to self-sacrifice (Fig. 3) and consistently with

H5. Model *m7* partially endorsed H4 on the intensity of other-referred moral emotions between levels of automation, showing only a significant effect of driving style on disgust intensity ($\chi^2_{11} = 7.30, p = .006$), with lower scores in the case of human-driving dilemmas (human driving: $\bar{x}$ = 2.41; AV: $\bar{x}$ = 2.55). Furthermore, anger and disgust intensities increased during the experimental procedure (*m6*: $\chi^2_{11} = 37.20, p < .001$; *m7*: $\chi^2_{11} = 11.66, p < .001$).

Discussion

Sacrificial moral dilemmas are widely considered a widespread, flexible, and powerful instrument to investigate morality in autonomous transportation (Bonnefon et al., 2016; Huang et al., 2019; (Martin et al., 2021a). When facing life-threatening dilemmatic situations involving an AV's passenger, the moral agent needs to face a non-utilitarian resolution that leads to their protection regardless of the number of consequential casualties, as well as a utilitarian resolution that allows for the preservation of the largest possible number of characters (e.g., pedestrians), but possibly at the expense of their own life (i.e., "I die, but many survive"). This tradeoff appears fundamentally different from the moral juxtaposition typically depicted in the traditional versions of sacrificial self-involvement dilemmas (e.g., Greene et al., 2001, 2004; Lotto et al., 2014; Moore et al., 2008). Indeed, in nondriving dilemmas, the moral agent can typically rely on a utilitarian resolution that provides the opportunity to protect the largest number of people *and* themselves (i.e., "I live, and many survive"). As compared to the AV dilemmas, such utilitarian outcome may mitigate the moral issue at stake, allowing for a resolution that is suitable at both the individual (i.e., "I can protect myself"*)* and collective (i.e., "I can also protect the largest possible number of people") levels. We aimed to investigate this structural difference, focusing on the potential role of self-sacrifice in affecting the endorsement of utilitarian behavior when facing autonomous- and human-driving moral scenarios.

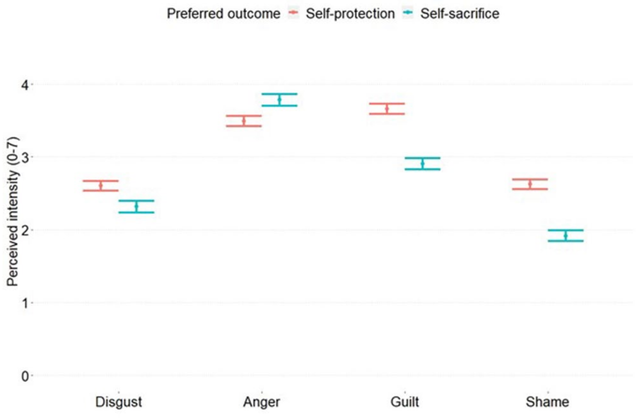


Fig. 3 Error Bars Plot Representing Means and Standard Errors of the Intensity of Other- and Self-Referred Emotions (Anger and Disgust, Shame and Guilt), Divided by Preferred Outcome (Self-Protection, Self-Sacrifice), Despite the Sacrifice Framing

As expected (H1), when self-sacrifice was framed in the utilitarian options (USF), we observed a clear decrease in the endorsement of the utilitarian moral code, as compared to the non-utilitarian framing (NSF). From a merely rational or evolutionary perspective, this result seems intuitive (i.e., "I prefer to live rather than to die"; e.g., Petrinovich et al., 1993) but reveals the importance of accounting for the self-sacrifice framing in the moral investigation (Huebner & Hauser, 2011; Thomson, 2008). In this context, the role of self-protection in human morality has been slightly elaborated in the literature. Despite Haidt's (2007) claims regarding the suppression of self-interest in moral reasoning, there is evidence that the moral agent places greater value on their own life than on a stranger's life (Huebner & Hauser, 2011; Moore et al., 2008). Nonetheless, Sachdeva et al. (2015) showed that the endorsement of self-sacrifice in moral dilemmas is perceived as more morally praiseworthy than sacrificing a third person. Overall, the evidence we collected appears to fit this conclusion, in that participants clearly endorsed the utilitarian behavior mainly when it allowed to protect themselves but judged this outcome as more immoral than the self-sacrificial choice. Interestingly, albeit only descriptively, when self-sacrifice was framed in the utilitarian option (USF), more people endorsed this behavior in the traditional human-driven vehicle than in the AV. This trend deserves further investigation, deepening the role of direct vs. indirect agency in the evaluation of moral behavior (Gill, 2021) and user preferences for AVs (i.e., Awad et al., 2018; Bonnefon et al., 2016; Haboucha et al., 2017) when the moral agent's life is at stake.

Additionally, considering the few studies focusing on moral judgment when human beings are at the wheel of a traditional human-driven vehicle (cf. Bruno et al., 2022), we checked for substantial differences in the endorsement of humans and autonomous drivers' moral behavior. In support of our hypothesis (H2), we observed a higher endorsement of the utilitarian behavior in human-driving scenarios, compared to those with their autonomous counterparts. This result is in line with the evidence Bruno et al. (2022) provided on the advantage of lifelike moral situations in bringing out the utilitarian moral code and with the potential distortion of moral judgment caused by the description of highly implausible events (Körner et al., 2019). Here, an AV hitting the road and opting between two maneuvers may seem a clear example of an unlikely event. The AV technology is surely developing quickly, but its rise and implementation seem still distant, especially in the general public's perspective (e.g., Guo et al., 2021). Indeed, the definition of AVs' public acceptance and public perception remains a work in progress for experts and policymakers (e.g., Hilgarter & Granig 2020; Othman, 2021). In contrast to what observed in terms of moral decisions, no differences were obtained between levels of automation in moral acceptability of driving maneuvers, suggesting that controlling for the level of automation may have a role in endorsing a certain moral behavior, but not in evaluating its moral acceptability. This result seems coherent with the finding of low acceptability of utilitarian AVs reported in previous studies (Bonnefon et al., 2016; (Martin et al., 2021a), allowing for a preliminary generalization of this tendency to both autonomous and non-autonomous vehicles. Clearly, these results deserve further testing in order to gain additional insights into the differences underlying moral perception of autonomous and human driving. Nonetheless, the observed predominance of the utilitarian moral code in both driving conditions seems in line with a 'structure-based' interpretation of moral dilemmas, claiming for the independence of moral choice from the specific contextualization.

In the present study, morality in driving-based sacrificial and self-involvement dilemmas was also examined from an emotional perspective, by measuring the intensity of four moral emotions, two self-referred (shame and guilt) and two other-referred (anger and disgust), experienced after decision-making. Our findings suggest that the consideration of these moral emotions is a valuable integration of previous investigations on moral judgment. Indeed, the results confirmed our third hypothesis (H3), as self-referred moral emotions (shame and guilt) were reported as more involved in traditional human-driving dilemmas, when the moral agent is in charge of the vehicle maneuvers, physically controlling the wheel. Specifically, a higher intensity of guilt in human-driving dilemmas was predictable, as this moral emotion aims for constructive responses (Tangney et al., 2007) and is more specifically linked to personal transgressions in the moral realm (e.g., Sabini & Silver 1997; Smith et al., 2002). As expected, the participants perceived disgust - as a negative feeling in response to a third party's moral violation - as more intense in response to critical moral events depicting AVs, partially confirming our hypothesis (H4) that other-referred moral emotions would be experienced with higher intensity in the AV- than human-driving dilemmas, and coherently with AI attribution of blame described by Hong et al. (2020). The inconsistency observed between disgust and anger is not surprising, as these moral emotions are known to be elicited by different cues in a moral situation (Gutierrez & Giner-Sirolla, 2007; Russell & Giner-Sirolla, 2011), with anger evoked by contextual cues of harm and intentionality, which are typical features of sacrificial and incidental dilemmas.

Interestingly, the decision to sacrifice or protect the self in the endorsement of the utilitarian behavior differentially affected the intensity of self-referred and other-referred moral emotions experienced after decision-making (H5). Indeed, the intensity of shame and guilt was higher when a self-protective decision was taken (i.e., "I live, and many survive"; NSF), as compared to a self-sacrificial decision (i.e., "I die, but many survive"; USF). Additionally, participants reported a higher intensity of anger when opting for self-sacrificial behavior (USF). This outcome is reasonable, since shame and guilt may be elicited by public and private selfish behaviors (Buss, 1980; Dillenberger & Sadowski, 2012; Gehm & Scherer, 1988), and injustice toward the self - or others - is able to trigger anger towards unspecified third-party in response to an immoral event (Haidt, 2003; Hutcherson & Gross, 2011). Overall, we can infer that people prefer to pursue utilitarianism while protecting the self, but this solution is perceived as more morally unacceptable, shameful, and blameworthy than the more praiseworthy self-sacrifice for a greater utilitarian goal (Sachdeva et al., 2015). This result seems consistent for manual and autonomous-driving vehicles, suggesting that the level of automation does not affect the evaluation of moral acceptability or the perceived intensity of moral emotions in sacrificial self-involvement dilemmas.

We acknowledge some limitations of the present study. Despite the experimental flexibility of text-based moral dilemmas, their application to on-road situations seems limited, especially in describing and deploying complex driving decisions and intricate traffic dynamics. Nevertheless, Sütfeld et al. (2019) compared VR-based and text-based AVs' moral scenarios, confirming the reliability of abstract contextualization in comparison with the more ecological - and more expensive - VR assessment. In this sense, combining several approaches is advisable to enhance the reliability of results with the help of immersive and realistic traffic environments. In addition, in the present study only two levels of automation were investigated, Level 0 (no automation) and Level 5 (full automation), similar to the opposite poles of the SAE's classification (2021). Our goal was to compare these two opposite means of transportation, in which the human has a completely different role in the selection of moral behavior. Nevertheless, autonomous driving features still require the driver's active involvement, and future applications on AV's morality may focus on more actual, intermediate levels of automation (e.g., Level 3), at which the human takeover (i.e., control of the vehicle being transferred from human driver to AV or vice versa) plays a key role in the allocation of driving responsibilities (Bellet et al., 2019). Finally, two features of moral alternatives' structure are worthy of attention. First, to maintain a constant and realistic ratio between lives saved and lives sacrificed, we added to the scenarios an additional passenger who shared the moral agent's fate. In several studies, research focused on the role, features, and number of characters involved in AV dilemmas (Awad et al., 2018; Bonnefon et al., 2016; Faulhaber et al., 2019), suggesting slight variations in terms of agreement with the utilitarian AV behavior or in the willingness to purchase this technology when more people are involved. Secondly, the non-utilitarian outcome in the USF dilemmas offered only partial accessibility, making explicit pedestrians' but not passengers' fate. Accessibility has been investigated in AV dilemmas leveraging on philosophical theories (Huang et al., 2019; (Martin et al., 2021a) and on the role of context in moral judgment (Schein, 2020). Kusev et al. (2016) revealed the role of full accessibility to the consequences of moral actions in boosting the endorsement of utilitarian behaviors, albeit only in the context of other-involvement scenarios. Further studies may confirm these results by manipulating the number of characters involved or controlling for the explicit availability of all the resulting consequences.

Overall, the present study has focused on self-sacrifice framing as a structural characteristic of sacrificial self-involvement dilemmas. We demonstrated that even if the pursuit of utilitarianism and self-protection is still conceived as the best possible outcome, obtaining a personal benefit from the utilitarian option is perceived as more immoral and shameful than the more praiseworthy self-sacrifice in the vision of the protection of the most. We collected this evidence by applying the moral dilemma tool to human vs. autonomous driving, detecting a preference for utilitarianism when the moral agent is in direct control of a traditional non-autonomous vehicle. Regarding the emotional impact of moral decision-making, take a stand in human-driving dilemmas seem to strongly elicit self-referred moral emotions (shame and guilt), indicating different emotional reactions to human- and autonomous-driving moral behaviors.

In conclusion, these results suggest taking into careful consideration the self-sacrifice factor in the development of sacrificial self-involvement dilemmas, as it seems to shape moral judgment and moral reactions to the utilitarian moral code. This finding is particularly important in the investigation of moral perception of driving behaviors, where, until today, the utilitarian behavior and the self-sacrifice factor have been combined differently than in typical sacrificial self-involvement dilemmas.

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. The authors would like to thank Dr. Paolo Buttazzoni and Dr. Fausto De Rossi for the help with data collection.

Authors' contributions Conceptualization: Giovanni Bruno; Methodology: Giovanni Bruno, Andrea Spoto, Lorella Lotto, Michela Sarlo Formal analysis and investigation: Giovanni Bruno, Andrea Spoto; Writing - original draft preparation: Giovanni Bruno; Writing - review and editing: Andrea Spoto, Lorella Lotto, Simone Cutini, Nicola Cellini, Michela Sarlo; Resources: Giovanni Bruno, Lorella Lotto, Michela Sarlo; Supervision: Andrea Spoto, Lorella Lotto, Nicola Cellini, Simone Cutini, Michela Sarlo.

Funding Open access funding provided by Università degli Studi di Padova within the CRUI-CARE Agreement.

Data Availability The data that support the findings of this study are openly available in Open Science Framework at: https://osf.io/pb3xc/?view_only=4ae203cc39e24d68859da3f6b67591a5.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethics approval The study was approved by the local ethics committee (ID No.: 3514).

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

References

Aquinas, T. (1952). The summa theologica (fathers of the english dominican province, trans.). In W. Benton (Series Ed.), Great Books Series: Vol. 19. Chicago: Encyclopedia Britannica, Inc. (Original work published 1274).

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.

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J. F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5), 2332–2337.

Bartels, D. M., Bauman, C. W., Cushman, F., Pizarro, D. A., & McGraw, A. P. (2014). Moral judgment and decision making. The Wiley Blackwell Handbook of Judgment and decision making. Chichester, UK: Wiley. 478–515.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. Journal of Statistical Software, 67(1), 1–48.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. Social and Personality Psychology Compass, 8(9), 536–554.

Bazerman, M. H., & D. Greene, J. (2010). In favor of clear thinking: Incorporating moral rules into a wise cost-benefit analysis - Commentary on Bennis, Medin, & Bartels (2010). *Perspectives on Psychological Science*, 5(2), 209–212.

Behrendt, H., & Ben-Ari, R. (2012). The positive side of negative emotion: The role of guilt and shame in coping with interpersonal conflict. Journal of Conflict Resolution, 55(6), 1116–1138.

Bellet, T., Cunneen, M., Mullins, M., Murphy, F., Pütz, F., Spickermann, F., … Baumann, M. F. (2019). From semi to fully autonomous vehicles: New emerging risks and ethico-legal challenges for human-machine interactions. Transportation research part F: traffic psychology and behaviour, 63, 153–164.

Bennett, J. M., Challinor, K. L., Modesto, O., & Prabhakharan, P. (2020). Attribution of blame of crash causation across varying levels of vehicle automation. Safety Science, 132, 104968.

Bentham, J. (1781). An introduction to the principles of morals and legislation. McMaster University Archive for the History of Economic Thought.

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

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, 1–14.

Buss, A. H. (1980). Self-consciousness and social anxiety. WH freeman.

Byrd, N., & Conway, P. (2019). Not all who ponder count costs: Arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontological and utilitarian tendencies. Cognition, 192, 103995.

Crockett, M. J. (2013). Models of morality. Trends in Cognitive Sciences, 17(8), 363–366. – 725.

Cushman, F. A. (2013). Action, outcome, and value a dual-system framework for morality. Personality and Social Psychology Review, 17(3), 273–292.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. Emotion, 12(1), 2.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. Psychological science, 17(12), 1082–1089.

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. Journal of statistical software, 64, 1–34.

De Melo, C. M., Marsella, S., & Gratch, J. (2021). Risk of injury in moral dilemmas with autonomous vehicles. Frontiers in Robotics and AI, 213.

Dillenberger, D., & Sadowski, P. (2012). Ashamed to be selfish. Theoretical Economics, 7(1), 99–124.

Elliott, D., Keen, W., & Miao, L. (2019). Recent advances in connected and automated vehicles. Journal of Traffic and Transportation Engineering (English Edition), *6*(2), 109–131.

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.

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

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., … 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.

Foot, P. (1978). The problem of abortion and the doctrine of double effect. In Virtues and vices. Blackwell.

Frison, A. K., Wintersberger, P., & Riener, A. (2016, October). First person trolley problem: Evaluation of drivers' ethical decisions in a driving simulator. In Adjunct proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications (pp. 117–122).

Gehm, T. L., & Scherer, K. R. (1988). Relating situation evaluation to emotion differentiation: Nonmetric analysis of cross-cultural questionnaire data. In K. R. Scherer (Ed.), Facets of emotion: Recent research (pp. 61–77). Lawrence Erlbaum Associates, Inc.

Gill, T. (2021). Ethical dilemmas are really important to potential adopters of autonomous vehicles. Ethics and Information Technology, 23(4), 657-673.

Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. PloS one, 8(4), e60418.

Gold, N., Pulford, B. D., & Colman, A. M. (2014). The outlandish, the realistic, and the real: Contextual manipulation and agent role effects in trolley problems. Frontiers in Psychology,5, 35.

Greenbaum, R., Bonner, J., Gray, T., & Mawritz, M. (2020). Moral emotions: A review and research agenda for management scholarship. Journal of Organizational Behavior, 41(2), 95–114.

Greene, J. D. (2016). Why cognitive (neuro) science matters for ethics. In S. M. Liao (Ed.), Moral brains: The neuroscience of morality (pp. 119–149). Oxford University Press.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. Cognition, 107(3), 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. Neuron, 44(2), 389–400.

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.

Guo, Y., Souders, D., Labi, S., Peeta, S., Benedyk, I., & Li, Y. (2021). Paving the way for autonomous Vehicles: Understanding autonomous vehicle adoption and vehicle fuel choice under user heterogeneity. Transportation Research Part A: Policy and Practice, 154, 364–398.

Gutierrez, R., & Giner-Sorolla, R. (2007). Anger, disgust, and presumption of harm as reactions to taboo-breaking behaviors. Emotion, 74, 853–868.

Haboucha, C. J., Ishaq, R., & Shiftan, Y. (2017). User preferences regarding autonomous vehicles. Transportation Research Part C: Emerging Technologies, 78, 37–49.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological review, 108(4), 814.

Haidt, J. (2003). The moral emotions. In R. J. Davison, K. R. Scherer, & H. H. Goldsmith (Eds.), Handbook of affective sciences (pp. 852–870). Oxford, UK: Oxford University Press.

Haidt, J. (2007). The new synthesis in moral psychology. Science, 316(5827), 998–1002.

Hidalgo, C. A., Orghiain, D., Canals, J. A., De Almeida, F., & Martín, N. (2021). How humans Judge Machines. MIT Press.

Hilgarter, K., & Granig, P. (2020). Public perception of autonomous vehicles: A qualitative study based on interviews after riding an autonomous shuttle. Transportation research part F: traffic psychology and behaviour, 72, 226–243.

Hong, J. W., Wang, Y., & Lanz, P. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. International Journal of Human–Computer Interaction, 36(18), 1768–1774.

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good. *Proceedings of the national academy of sciences*, 116(48), 23989–23995.

Huebner, B., Dwyer, S., & Hauser, M. D. (2009). The role of emotion in moral psychology. Trends in Cognitive Sciences, 13(1), 1–6.

Huebner, B., & Hauser, M. D. (2011). Moral judgments about altruistic self-sacrifice: When philosophical and folk intuitions clash. Philosophical Psychology, 24(1), 73–94.

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social–functionalist account of anger, disgust, and contempt. Journal of Personality and Social Psychology, 100(4), 719–737.

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

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. Social neuroscience, 10(5), 551–560.

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2018). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. Cognition, 134, 193–209.

Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Pipa, G., Ste- phan, 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, 1–15.

Kant, I. (1785). Groundwork of the metaphysics of morals.

Krebs, D., & Höhne, J. K. (2021). Exploring scale direction effects and response behavior across PC and smartphone surveys. Journal of Survey Statistics and Methodology, 9(3), 477–495.

Körner, A., Joffe, S., & Deutsch, R. (2019). When skeptical, stick with the norm: Low dilemma plausibility increases deontologi-cal moral judgments. Journal of Experimental Social Psychology,84, 103834

Kroll, J., & Egan, E. (2004). Psychiatry, moral worry, and the moral emotions. Journal of Psychiatric Practice, 10(6), 352–360.

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.

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. Perspectives on Psychological Science, 10(4), 518–536.

Lenth, R. (2020). emmeans: Estimated marginal means, aka least-squares means (Version 1.5. 2–1) [R package].

Lewis, H. B. (1971). Shame and guilt in neurosis. Psychoanalytic review, 58(3), 419–438.

Li, J., Zhao, X., Cho, M. J., Ju, W., & Malle, B. F. (2016). From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. *SAE Technical paper*, 10, 2016-01.

Lotto, L., Sarlo, M., & Manfrinati, A. (2014). A New Set of Moral Dilemmas: Norms for Moral Acceptability, decision Times, and emotional salience. Journal of Behavioral Decision Making, 16(20), 6513–6525.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. Psychological Inquiry, 25, 147–186.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). 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, P., Teal, J., Baranova, V., & Rigal, B. (2021b). Moral decision making: From Bentham to veil of ignorance via perspective taking accessibility. Behavioral Sciences, 11(5), 66.

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

Martínez-Díaz, M., & Soriguera, F. (2018). Autonomous vehicles: Theoretical and practical challenges. Transportation Research Procedia, 33, 275–282.

Martí-Vilar, M., Escrig-Espuig, J. M., & Merino-Soto, C. (2021). A systematic review of moral reasoning measures. Current Psychology, 1–15.

Maurer, M., Gerdes, J. C., Lenz, B., & Winner, H. (2016). Autonomous driving: Technical, legal and social aspects. *Springer Nature*.

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.

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2022). Moral judgment as categorization (MJAC). Perspectives on Psychological Science, 17(1), 131–152.

McManus, R. M., & Rutchick, A. M. (2019). Autonomous vehicles and the attribution of moral responsibility. Social psychological and personality science, 10(3), 345–352.

Meyer, J., Becker, H., Bösch, P. M., & Axhausen, K. W. (2017). Autonomous vehicles: The next jump in accessibilities?. Research in transportation economics, 62, 80–91.

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.

Othman, K. (2021). Public acceptance and perception of autonomous vehicles: A comprehensive review. AI and Ethics, 1(3), 355–387.

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. Journal of personality and social psychology, 64(3), 467.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. Journal of Experimental Social Psychology, 39, 653–660.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., … Wellman, M. (2019). Machine behaviour. Nature, 568(7753), 477–4

Rawls, J. (2009). A theory of justice. Harvard University Press (Original publication 1971).

Riegler, A., Riener, A., & Holzmann, C. (2021). A systematic review of virtual reality applications for automated driving: 2009–2020. Frontiers in human dynamics, 48.86.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). Journal of personality and social psychology, 76(4), 574.

Russell, P. S., & Giner-Sorolla, R. (2011). Moral anger, but not moral disgust, responds to intentionality. Emotion, 11(2), 233.

Sabini, J., & Silver, M. (1997). In defense of shame: Shame in the context of guilt and embarrassment. Journal for the Theory of Social Behaviour, 27(1), 1–15.

Sachdeva, S., Iliev, R., Ekhtiari, H., & Dehghani, M. (2015). The role of self-sacrifice in moral dilemmas. PloS one, 10(6), e0127409.

SAE International (2021), Taxonomy and definitions for terms related to driving automation Systems for On-Road Motor Vehicles. J3016_201806, https://www.sae.org/standards/content/j3016_201806/

Samuel, S., Yahoodik, S., Yamani, Y., Valluru, K., & Fisher, D. L. (2020). Ethical decision making behind the wheel – a driving simulator study. Transportation Research Interdisciplinary Perspectives, 5, 100147.

Sarlo, M., Lotto, L., Manfrinati, A., Rumiati, R., Gallicchio, G., & Palomba, D. (2012). Temporal dynamics of cognitive–emotional interplay in moral decision-making. Journal of Cognitive Neuroscience, 24(4), 1018–1029.

Schein, C. (2020). The Importance of Context in Moral Judgments. Perspectives on Psychological Science, 15(2), 207–215.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. Personality and social psychology bulletin, 34(8), 1096–1109.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. Nature Human Behaviour, 1(10), 694–696.

Smith, R. H., Webster, J. M., Parrott, W. G., & Eyre, H. L. (2002). The role of public exposure in moral and nonmoral shame and guilt. Journal of personality and social psychology, 83(1), 138.

Sütfeld, L. R., Ehinger, B. V., König, P., & Pipa, G. (2019). How does the method change what we measure? Comparing virtual real- ity and text-based surveys for the assessment of moral decisions in traffic dilemmas. PLoS ONE, 14(10), 1–14.

Sütfeld, L. R., Gast, R., König, P., & Pipa, G. (2017). Using virtual reality to assess ethical decisions in road traffic scenarios: Applicability of value-of-life-based models and influences of time pres- sure. Frontiers in Behavioral Neuroscience, 11, 122.

Tangney, J. P., & Dearing, R. L. (2002). Shame and guilt Guilford Press. New York, NY.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. Annu. Rev. Psychol., 58, 345–372.

Terracciano, A., McCrae, R. R., & Costa Jr, P. T. (2003). Factorial and construct validity of the italian positive and negative affect schedule (PANAS). European journal of psychological assessment, 19(2), 131.

Thomson, J. J. (1985). The trolley problem. Yale Law Journal, 94, 1395–1415.

Thomson, J.J. (2008). Turning the trolley. Philosophy and Public Affairs, 36, 359–374.

Unger, P. (1996). Living high and letting die. New York: Oxford University Press.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. Psychonomic bulletin & review, 11(1), 192–196.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of personality and social psychology, 54(6), 1063.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.