Human-Centric Intelligent Systems (2025) 5:418–430
https://doi.org/10.1007/s44230-025-00110-9

RESEARCH ARTICLE

Moral Attribute Elicitation for Automated Vehicles: Does Scenario Presentation Method Matter?

Chloe Gros[1] · **Leon Kester**[2] · **Marieke Martens**[2,3] · **Peter Werkhoven**[1,2]

Received: 25 April 2025 / Revised: 28 July 2025 / Accepted: 1 September 2025 / Published online: 12 September 2025

Abstract
The development of automated vehicles (AVs) presents significant ethical challenges, particularly in eliciting moral attributes from humans to guide AV decision-making in high-risk situations. This study explores whether two different elicitation methods—specifically, 3D animations and 2D schematic representations—lead to different moral attributes and priorities. Based on prior research, it was expected that more immersive methods, such as 3D animations with their dynamic and enriched visual presentation, may elicit stronger utilitarian responses by enhancing emotional engagement, as well as support participants' comprehension of the scenarios. The findings indicate that animations were associated with greater utilitarianism, with participants prioritising attributes like Physical Damage and Vulnerability while de-emphasising Car Preservation. The dynamic scenes also resulted in greater consensus on the completeness of the provided attributes. While the total number of newly proposed attributes remained similar, a larger proportion of participants suggested them compared to the 2D schematic representations. These results illustrate how methodological choices in moral attribute elicitation may influence the types of moral responses elicited, raising further questions about which specific features contribute to these effects and how presentation formats can be optimally designed for eliciting moral values that are optimally aligned with societal values.

Keywords Automated vehicles · Moral attribute elicitation · 3D animation · Scenario presentation · AI ethics · Autonomous driving systems

1 Introduction

In the development of automated vehicles (AVs), complex ethical challenges need to be addressed to ensure these systems align with societal values. One of the key challenges is the elicitation of moral attributes from society in order to understand the main factors that influence how AVs should make decisions. While sub-symbolic neural networks may be suited to specific tasks such as image recognition, they do not reason according to moral attributes, so explicitly studying these attributes is necessary to ensure that AVs make societally aligned ethical choices [14].

Human ethical decision-making can be modelled through various frameworks, including deontology, virtue ethics, and utilitarianism, as well as integrated frameworks [9, 11, 22]. The present study adopts Augmented Utilitarianism (AU), an integrated approach that draws on the strengths of each of these perspectives while dynamically adapting to evolving societal values and situational contexts [1, 2]. Unlike classical utilitarianism, which focuses solely on maximising aggregate utility, or deontological ethics, which applies predefined rules, AU incorporates harm minimisation as a central principle and allows for the weighting of moral attributes based on empirical social preferences. It draws from moral psychology, notably the theory of dyadic morality [22], to capture the complexity of human moral reasoning, considering not only outcomes but also the intentions and roles of agents and the perceptions of observers, thereby combining deontology, consequentialism, virtue ethics and adding the bystander perspective, combining this into one integrated societal perspective. A key operational feature of AU is its

✉ Chloe Gros
c.n.gros@uu.nl

[1] Faculty of Science, Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, the Netherlands

[2] TNO Netherlands, Intelligent Autonomous Systems, Postbus 96864, 2509 JG The Hague, the Netherlands

[3] Eindhoven University of Technology, Eindhoven, Netherlands

use of ethical goal functions, which encode societal values into a transparent, quantitative structure that can guide real-time AV decision-making. This enables automated systems to act consistently with human ethical expectations, even in novel or uncertain contexts. By enabling the construction of these adaptive and explainable goal functions, AU supports both accountability and public trust in AV systems, making it particularly well-suited for the challenges of societally aligned decision-making in automated driving.

A crucial aspect of applying AU is the method used to elicit human ethical judgments, meaning the process of presenting ethically complex scenarios to human participants and gathering their evaluations of which moral attributes—such as harm, responsibility, or fairness—should guide the automated vehicle's decisions. These attributes will ultimately be used to define the ethical goal function of the AV. Previous studies have used a range of elicitation methods, including text-based surveys, static images, video stimuli, and immersive virtual reality (VR) environments [3, 6–8, 10, 12, 16, 17, 19, 25]. However, the chosen elicitation method may significantly impact how participants engage with moral dilemmas, potentially influencing the outcomes of studies and, consequently, the design of AV decision-making models. Furthermore, the complexity of moral decision-making in AVs requires a robust methodological approach to accurately capture the diversity of human moral values. It is essential to account for individual differences in moral reasoning to ensure that the collected data reflect a broad and representative spectrum of ethical perspectives. Variations in scenario presentation, the level of engagement, and the degree of realism may affect the moral attributes that participants prioritise. Understanding these effects is essential for designing AVs that can make ethical decisions in real-world traffic situations which align with societal values.

The way ethical scenarios are visually presented in AV studies is of profound practical significance, as it can influence moral judgment and, ultimately, algorithmic decision-making in life-or-death situations. Research shows that people's ethical evaluations and preferences shift depending on the realism and immersion of the visual context. For instance, Cecchini et al. [9] argue for moving beyond abstract trolley dilemmas by using realistic traffic simulations to better reflect how humans make moral decisions in actual driving contexts [9]. Similarly, Francis et al. [13] demonstrate that virtual reality-based moral dilemmas elicit greater endorsement of utilitarian decision-making compared to passive observation, which can lead to different outcomes in ethical decision-making [13]. These differences are not trivial—how scenarios are perceived can influence which lives are prioritised in AV algorithms, as emphasised in the Ethical Valence Theory proposed by Evans et al. (2020), which treats moral claims as inputs for computational ethical reasoning (Evans et al., 2020). As such, the mode of representation directly shapes ethical intuitions and preferences, making it a crucial factor in the socially acceptable design of AV behaviour.

The scenarios used in this study were carefully constructed to reflect ethically relevant trade-offs that automated vehicles may face in real-world contexts. Each scenario was designed to instantiate moral attributes drawn from established biomedical ethics and moral psychology frameworks, such as fairness, utility, harm, and responsibility. By operationalising these attributes through concrete instances—such as differences in age, disability, legal behaviour, or moral responsibility—the scenarios model realistic dilemmas that AVs may encounter in dynamic environments. This design ensures that the moral judgments elicited from participants are grounded in ethically meaningful dimensions, rather than hypothetical or artificially simplified dilemmas. As such, the scenarios contribute directly to the ethical programming of AV decision-making systems by enabling the evaluation of moral preferences across a representative spectrum of social values.

The goal of eliciting moral attributes for AVs is to capture how human observers morally evaluate complex driving scenarios—evaluations that are used as inputs in shaping the behaviour of automated systems. Within the AU framework, these human preferences directly inform the ethical goal functions that govern AV decision-making. As such, the method used to elicit these values is not a neutral choice: different elicitation methods can lead to different prioritizations of moral attributes, which in turn affect how AVs act in practice. Assessing and comparing elicitation methods—such as 2D schematic representations versus 3D animations—is therefore essential to ensure that the moral data collected is both contextually accurate and aligned with societal expectations. Importantly, participants in this study were not asked what they themselves would do in each scenario, but rather what they believed the AV should do. This distinction reinforces the "experiencer" or bystander role central to the AU framework, which focuses on the societal perspective rather than individual self-interest. In this context, participants are not acting as drivers or agents in the scenario, but as societal observers/experiencers making normative judgments about what an automated vehicle should do. This distinction ensures that the elicited values reflect a third-person, impartial perspective aligned with the societal focus of the AU framework. The theoretical contribution of this study lies in its systematic examination of how presentation format influences the moral preferences elicited, and consequently, the ethical logic embedded in AV systems.

Prior research suggests that more immersive methods, such as VR, elicit stronger emotional responses, leading to an increased tendency toward utilitarian decision-making versus deontological decision-making [13, 18]. Greater

realism may also improve scenario comprehension, potentially supporting more extensive attribute elicitation.

This study investigates whether previously observed effects of immersive media on moral judgment and scenario comprehension extend to AV-related ethical decision-making. Specifically, 2D schematic representations (2DS) and 3D animations (3DA) are compared as methods for eliciting moral attributes in AV contexts.

2 Hypotheses

2.1 Greater Utilitarianism in Video

Utilitarianism is a normative ethical framework aimed at maximising the overall good. It can be divided into *positive utilitarianism*, which seeks to increase happiness and benefits for the greatest number, and *negative utilitarianism*, which focuses on preventing or reducing suffering as the primary moral goal. In the context of AV decision-making, negative utilitarianism would emphasise attributes such as Physical Damage, Psychological Harm, and Perceived Vulnerability, prioritising harm minimisation before considering other factors. In contrast, positive utilitarianism would consider attributes like Time Delay, focusing on enhancing overall happiness by prioritising factors that promote physical safety, psychological comfort, and an optimal driving experience.

This study examines the expectation that 3DA will prompt more negative utilitarian responses than 2DS because higher realism and immersion increase emotional engagement and perceived urgency in moral decision-making. Maćkiewicz et al. [18] suggest that more immersive environments heighten participants' sensitivity to potential harm, leading to stronger aversion to negative outcomes.

In critical scenarios, in which the probability of harm is high, a greater negative utilitarian response would prioritise attributes that focus on minimising overall harm. It is expected that attributes such as Physical Harm, Psychological Harm, and Perceived Vulnerability will be rated higher in relevance with 3D representations than 2DS. Negative utilitarian decision-making emphasises reducing suffering and protecting vulnerable individuals, which is crucial in urgent situations. Additionally, Moral Responsibility and Fair Innings are expected to increase in relevance, as utilitarianism seeks to distribute benefits and burdens equitably across society, ensuring that critical decisions consider the broader impact.

In non-critical scenarios, in which decisions have less immediate or severe consequences, the emphasis on negative utilitarian attributes might be less pronounced. However, a shift towards attributes that promote collective happiness is still expected, focusing on positive utilitarianism. Attributes like Car Preservation might be rated lower, as utilitarianism prioritises the greater good over individual interests, even in less urgent situations. Furthermore, Time of Arrival and Legality might also be influenced by the video representation, with utilitarian responses potentially placing less emphasis on efficiency or strict rule adherence if these conflict with achieving the best overall outcome.

2.2 Improved Elicitation of Moral Attributes with 3D Animations

As highlighted by Mackiewicz et al. [18], increased realism may enhance the elicitation of moral attributes by fostering a richer understanding of the situation. Experimental settings that are more realistic and immersive may be better suited for this purpose, as they enable participants to consider a wider array of contextual factors that might not be apparent in abstract formats such as text or 2D illustrations.

This hypothesis is tested by examining participants' responses to the scenarios, particularly the number of new attributes they propose and the nature of their feedback. It is expected that 3DA may lead to a higher number of newly proposed attributes or fewer comments indicating confusion.

3 Methodology

3.1 Scenario Representation

In this experiment, 2DS, meaning abstract drawings of a traffic scenario, is compared to 3DA, meaning a computer-generated video of a 3D traffic scenario, for eliciting moral attributes for AV decision-making.

3DA was selected instead of Virtual Reality (VR) to maintain a third-order perspective: instead of embodying a specific character, the participants were asked to observe the scenario from a detached, observational vantage point that allows for critical analysis and a holistic understanding of the scene's dynamics.

VR's first-person immersion can influence participants' perceptions by shaping their viewpoint within the scenario, placing them in a first-order perspective, experiencing the scenario as if they were directly involved, such as the driver or a pedestrian in the scene. In contrast, 3D video representations and 2DS present situations from a third-order perspective, aligning with John Rawls' ''veil of ignorance'' concept [21], which advocates for ethical decisions made from an impartial perspective to maximise societal happiness. By removing the individual's immediate self-interest and personal circumstances, this approach encourages participants to evaluate complex scenarios through an impartial lens, focusing on systemic

implications rather than subjective responses. Participants act as observers or "experiencers," fulfilling the normative role in moral situations [24].

3.2 Format Differences and Their Potential Effects

The two distinct presentation formats, 2DS and 3DA, differ across multiple dimensions. These differences create methodological challenges in isolating specific causal factors influencing moral judgments.

One of the most fundamental differences is that 3DA incorporate movement and temporal progression, allowing participants to observe the unfolding of events rather than inferring them from a static snapshot. This dynamic presentation enables observation of behavioural patterns, reaction times, and causal sequences that might influence vulnerability, responsibility, and risk attributions.

While the dynamic nature of 3D animations is a key distinguishing factor, the two presentation formats differ across several other dimensions that may influence participant responses. These include differences in colour (black-and-white in 2DS versus full-colour in 3DA), visual richness (minimalistic line drawings versus textured, realistic scenes), depth perception (flat, schematic layout versus three-dimensional perspective with spatial depth), and environmental context (limited background detail in 2DS versus immersive settings with ambient elements in 3DA). These multi-sensory features may contribute to enhanced situational awareness, emotional engagement, and comprehension of moral dynamics. For instance, improved depth cues in 3DA allow participants to more accurately assess distances, speeds, and spatial relations between road users—factors that are crucial in evaluating accident probability and perceived responsibility. Similarly, richer environmental details may increase perceived realism and prompt more intuitive moral judgments. The observed effects should be interpreted as the result of the combined influence of these presentation features.

Differences in presentation formats, despite their multifaceted nature, may influence moral judgments through a common pathway of increased emotional arousal. The enhanced realism, dynamic nature, and perceptual richness of 3DA likely intensify emotional engagement with the scenarios, which previous research has linked to more utilitarian decision-making in moral dilemmas [18]. However, it is important to note that emotional engagement does not linearly correlate with better decision-making,excessive emotional arousal can hinder rational processing and lead to suboptimal or biased judgments. Thus, while 3D scenarios may improve ecological validity and engagement, their influence on ethical decision-making must be interpreted with caution.

3.3 Attribute Selection

The objective of this experiment was to compare the differences between two different elicitation methods—specifically, 3DA and 2DS—in capturing moral attributes for AV decision-making. This study is grounded in the framework of Augmented Utilitarianism (AU), a non-normative ethical model that integrates elements of virtue ethics, deontology, and consequentialism, as well as moral psychology and neuroscience [1, 2].

Unlike traditional utilitarian approaches that focus solely on maximising overall happiness, AU introduces dynamic ethical goal functions that adapt to societal values and situational contexts. These goal functions quantify moral attributes and adjust decision-making processes in real-time, ensuring that AVs align with predefined ethical considerations.

A primary objective of AU involves defining and validating a set of moral attributes that contribute to an ethical goal function. This process includes systematically analysing participant responses across various elicitation methods to refine the weighting and applicability of these attributes in the context of AV decision-making.

The first set of attributes has been selected following the method based on AU described by Gros et al. [15]. Establishing these attributes in advance ensures a structured and systematic approach to analysing moral decision-making in AV scenarios. Without predefined attributes, participants might rely on highly subjective or inconsistent criteria, making comparisons across scenarios difficult. Using a well-defined set of attributes, as outlined in Table 1, establishes a foundation for evaluating the impact of different presentation formats on the perceived importance of these attributes.

Level 1 reflects the concept of harm as defined in the moral psychology theory of dyadic morality, involving an agent, an action, and a patient [22]. Levels 2, 3, and 4 are based on biomedical ethics principles adapted for mobility [5]. Level 2 distinguishes between Cost, Benefits (Utility), and Justice (Fairness), which are key principles in biomedical ethics [5]. Levels 3 and 4 further break down these principles into more specific attributes. Level 5 attributes represent various instances of Level 4 categories.

This experiment focuses on the relevance scores at Level 4, which are essential for explaining decision-making in specific scenarios. The attributes are defined as follows:

Physical Harm: Amount of physical damage done to an individual or a group of individuals. It includes all damage, from minor (bruise) to major (death), done to all concerned individuals.

Perceived Vulnerability: Vulnerability not directly linked to physical harm. For example, disabled people might be perceived as more vulnerable, even in situations

Table 1 Classification of attributes based on biomedical ethics and moral psychology

Level 1	Level 2	Level 3	Level 4	Level 5
Patient	Utility	Harm	Physical Harm	Severity of damage
				Type of damage (temporary/permanent)
			Psychological Harm	Family Status
			Perceived Vulnerability	Age
				Physical condition (disability)
				Gender
		Social Utility	Social Status	Age
				Physical Condition
				Profession
				Gender
				Family status
	Fairness	Liability	Moral responsibility	/
		Fair opportunity	Fair innings	Age
Action of AV	Fairness	Liability	Legality	/
		Fair opportunity	Lottery	/
Agent	Cost	Timeliness	Time Delay	/
		Financial Cost	Car preservation	/

where they would sustain the same amount of physical damage as someone not disabled.

Psychological harm: Any psychological damage that could result from a situation, to the potential victim and other parties (i.e. family and friends of the victim).

Social utility: The level of social or economic value a person is considered to possess. It is used to rank how different people benefit society as a whole.

Moral responsibility of the pedestrian: A person is morally responsible for creating, by their actions, a situation in which risk of harm is imposed on themselves and/or others. For example, by crossing the road without looking, one is responsible for creating a risk to oneself and potentially others.

Legality: An action is considered to be illegal if it is explicitly forbidden by the law. The legality of the pedestrians applies to the previous or current action of the pedestrians, such as crossing at a red light.

Fair innings: The fair innings argument takes the view that everyone is entitled to live a certain number of years. Everyone who does not get to live at least this number of years suffers the injustice of being cut off in their prime.

Lottery: A lottery system involves choosing randomly between different outcomes.

Time delay: Represents the importance of arriving on time. The arrival time can be influenced by any action of the AV. For example, slowing down to pass near a pedestrian would create a time delay, and the car would arrive later at its destination. Keep in mind that if the arrival time is never relevant, the car will never move.

Car preservation: Harm done to the owner of the vehicle following physical damage to the vehicle.

Lower-level attributes, such as those at Level 5, are broader and can encompass multiple Level 4 attributes, making them less useful for explanation-based decision-making. For example, the Level 5 attribute "Age" can be linked to both Fair Innings and Perceived Vulnerability, meaning it lacks the specificity needed for precise judgments in a given scenario.

To create the scenarios, one or more instances (Level 5) were selected for each Level 4 attribute. To cover all Level 4 attributes, 11 instances were tested: family status, gender, child, senior, disabilities, profession, the legality of the action, lottery, pedestrian's moral responsibility, car preservation, and time delay. In addition, for each attribute, both a critical (high probability of physical harm) and a non-critical (low or absent probability of physical harm) scenario were presented.

3.4 Procedure

Two separate studies were conducted: one in which participants evaluated 2DS and another in which they assessed 3DA. The first group of participants was presented with abstract picture scenarios (Fig. 1) with a descriptive text, while the second group was presented with 3DA. The videos reprised the schematic representations, showing a 3D-rendered version of it made with Unity. To maintain a neutral point of view, the video showed two perspectives: one on top of the AV and one behind the pedestrian. Both the 2DS

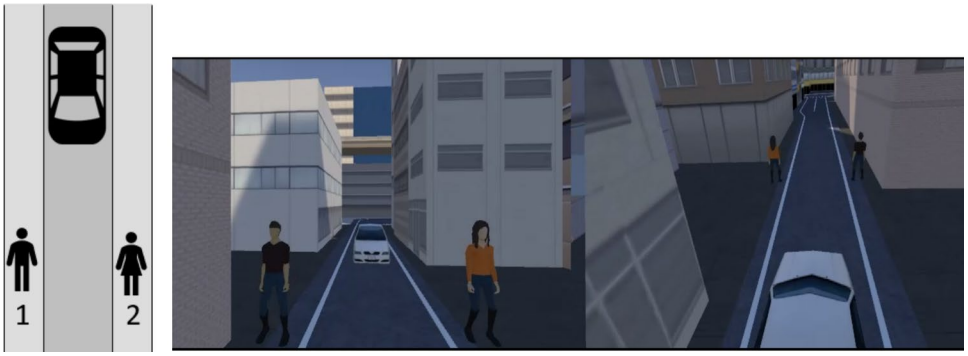

Fig. 1 Non-Critical "Gender" scenarios in 2DS and 3D video representation

and 3DA also had a text description giving some detail about the scenario, for example, for critical scenarios: "An empty AV is driving on a road. It comes across a crosswalk where 2 pedestrians are crossing, one of them is a small child (2). Apart from this, the two pedestrians look similar. The AV tries to stop; however, the brakes are not responsive. The connection wire to the brakes is defective. The AV has to make a trajectory decision." And for non-critical scenarios: "An empty AV is driving on a narrow one-way road. There are sidewalks on both sides, each with one pedestrian. The pedestrian on the right sidewalk is in a wheelchair (2). The AV has to position itself on the road (such as leaning more on one side or the other or slowing down)."

Participants were then asked to score the relevance of each of the attributes on a scale from 1 to 5, for each of the scenarios. All of the attributes were asked in each scenario to mitigate specific interaction effects.

As the experiment is explanation-based, after each question, the participants were asked to give explanations about their decision-making. This helped understand why certain attributes were ranked higher on certain scenarios, especially on scenarios that included multiple attributes, and to make sure that the question was understood correctly.

After each scenario, participants were asked if they perceived the scenario as critical, and if they thought the scenario was relevant for AV decision-making (Q1: *"Do you perceive this scenario as critical? A critical scenario is defined as a scenario where the probability of harm is high"*; Q2: "*Do you find this scenario to be relevant in identifying the attributes relevant in the population's decision-making for automated driving?")*. At the end of the questionnaire, they were also asked if they believed that these attributes fully described AV decision-making (Q3: *"Do you think that these attributes completely define the AV's decision-making? If not, what would be missing?")*.

The 11 scenarios were presented in random order. Participants did not have any time restrictions, and in the 3DA experiment, they were able to replay the videos at will.

Additionally, participants had multiple opportunities to propose new attributes. Firstly, after ranking each of the attributes after a scenario, they had the opportunity to add their own attribute to the list and to rank it as well, if they believed that another attribute might come into play in this particular scenario. Secondly, they had another opportunity at the end of the questionnaire, when they were asked if they believed that these attributes completely described the AV's decision-making and, if not, what was missing. They also had a final opportunity to add any other remarks that they believed could be relevant.

3.5 Participants

A total of 204 participants were recruited, comprising two distinct groups studied approximately one year apart. Participants were university students recruited through campus posters and announcements, with each participant receiving a 10€ compensation for completing the experiment.

3.5.1 2D Schematic Representations Group Demographics

The 2DS group (n = 103) consisted of 47 females, 52 males and 4 prefer not to say. Age distribution was as follows: 73 participants aged 18–24, 28 participants aged 25–34 and 2 older than 35. Employment status revealed 82 students, 13 part-time employees, 7 full-time employees, and 1 unemployed. The educational background showed 64 participants with bachelor's degrees, 21 with high school diplomas, 15 with master's degrees, and 3 others.

Regarding transportation habits, 74 participants held a driver's license. Primary modes of transportation were bicycles (51 participants), public transport (41 participants), and cars (8 participants). Driving frequency varied: 28 participants drove 2–4 times per week, 2 drove daily, 10 never drove, 26 drove once a month, and 19 drove once per week.

3.5.2 3D animations Group Demographics

The 3DA group (n = 101) comprised 45 males, 52 females, and 4 preferred not to say. Age distribution included 65 participants aged 18–24 and 32 participants aged 25–34.

Employment status showed 73 students, 17 full-time employees, 10 part-time employees and 1 unemployed. Educational levels consisted of 28 high school graduates, 38 bachelor's degree holders, 32 with master's degrees, and 3 others.

Similar to the 2DS group, 72 participants possessed a driver's license. Transportation preferences were: public transport (48 participants), bicycles (38 participants), and cars (11 participants). Driving frequency was reported as: 7 participants never drove, 20 drove once a month, 13 drove once per week, 14 drove 2–4 times per week, and 3 drove daily.

Both groups primarily consisted of university students and PhD candidates, which was an intentional choice aimed at maintaining a homogeneous sample, facilitating comparison of the experimental results between groups. This selection helped reduce inter-group variability and enhance internal consistency, particularly given the sample size, which, while limited, was sufficient for detecting statistically significant effects in key comparisons. As a result, the majority of participants were members of the general public with limited professional driving experience. As the aim of this study was to investigate whether differences in scenario representation influence the elicitation of moral attributes and not to use these results to build an ethical goal function for AV decision-making, this study does not claim that the attributes elicited are representative of the broader population. Ultimately, this method should be applied across a range of societal target groups to ensure the inclusivity and generalisability of the findings.

3.6 Data Analysis

The data analysis was structured to assess whether presentation format (2D schematic representations vs. 3D animations) significantly influenced the prioritisation of moral attributes in AV scenarios. Given the ordinal nature of the data (Likert-scale responses) and the lack of normal distribution, non-parametric statistical methods were used throughout.

To evaluate overall differences in attribute ratings across scenarios, the average Likert-scale score of each moral attribute was calculated, aggregated both within each scenario and across all scenarios in which the attribute was relevant. These aggregated scores were compared between experimental conditions using nonparametric multivariate analysis of variance (MANOVA), as implemented through the npmv package in R. This approach was selected due to its suitability for analysing multivariate, non-normally distributed data without relying on assumptions of homoscedasticity.

To examine attribute-level differences in more detail, Wilcoxon rank-sum tests (Mann–Whitney U tests) were conducted to compare the distributions of individual attribute scores across conditions. This test was chosen because it is robust to non-normal data and appropriate for comparing two independent groups.

For each scenario and attribute, only those considered relevant to the context were included in the analysis. For example, in the vulnerability scenario, only Physical Damage and Perceived Vulnerability were analysed. Attribute results were further separated by scenario type (critical vs. non-critical) to account for contextual variation in moral priorities.

To assess differences in participant comprehension between the two presentation formats, three binary and categorical response variables were analysed: perceived scenario criticality, perceived relevance for AV decision-making, and agreement on attribute completeness. Given the categorical nature of these variables, chi-square tests of independence were used to evaluate whether response distributions significantly differed between the 2D schematic and 3D animation conditions.

4 Results

The results were analysed by aggregating the average Likert-scale score of each attribute on its relevant scenarios, as well as across all its relevant scenarios. These results were then compared using the method (2DS or 3DA) as the predictor variable.

Results of the npmv package show significant differences in aggregated results for both critical (ANOVA-type test: $T = 25.408$, p-value < 0.001) and non-critical scenarios (ANOVA-type test: $T = 14.283$, p-value $= 0$).

4.1 Results per Scenario

The average ratings for each moral attribute across different scenarios were calculated. For each scenario, only the relevant attributes were taken into account. For example, for the vulnerability scenario (visually impaired man vs. non-visually impaired man), the attributes Physical Damage and Perceived Vulnerability are taken into account in the results. For some other scenarios, this did not play a role.

First, the main effect of scenario presentation (3DA-based versus picture-based) was analysed per scenario across attributes. The scenarios presenting significant differences were:

Car preservation Non-Critical (ANOVA-type test: $T = 15.147$, p-value < 0.001). The ratings of the attributes Physical Damage ($M_V = 3.38$, $M_I = 2.66$) and Car Preservation ($M_V = 3.11$, $M_I = 3.81$) were taken into account as these attributes were relevant for this scenario.

Senior Critical (ANOVA-type test: T = 2.594, p-value = 0.044). The attributes Physical Damage ($M_V = 4.51$, $M_I = 4.27$), Vulnerability ($M_V = 3.09$, $M_I = 2.59$), Social Status ($M_V = 1.77$, $M_I = 1.54$), Fair Innings ($M_V = 2.55$, $M_I = 2.42$) were taken into account.

Moral Responsibility Critical (ANOVA-type test: T = 4.588, p-value = 0.012). The attributes taken into account were Physical Damage ($M_V = 4.70$, $M_I = 4.46$) and Moral Responsibility ($M_V = 3.55$, $M_I = 3.12$).

Gender Critical (ANOVA-type test: T = 5.092, p-value = 0.003). The attributes Physical Damage ($M_V = 4.60$, $M_I = 4.60$), Vulnerability ($M_V = 2.60$, $M_I = 2.04$) and Social Status ($M_V = 1.66$, $M_I = 1.29$) were taken into account.

Gender Non-Critical (ANOVA-type test: T = 3.452, p-value = 0.022). The attributes Physical Damage ($M_V = 3.85$, $M_I = 3.68$), Vulnerability ($M_V = 2.34$, $M_I = 1.97$) and Social Status ($M_V = 1.61$, $M_I = 1.29$) were taken into account.

Vulnerability Non-Critical (ANOVA-type test: T = 110.033, p-value < 0.001). The attributes Physical Damage ($M_V = 3.92$, $M_I = 3.70$) and Vulnerability ($M_V = 3.72$, $M_I = 3.28$) were taken into account.

4.2 Results per Attribute

Second, the main effect of scenario presentation (3DA-based versus picture-based) was analysed per attribute across critical and non-critical scenarios.

In this analysis, all results of each attribute were considered within their relevant scenarios. For example, the moral responsibility attribute is pertinent only in the moral responsibility scenario, whereas the physical damage attribute is relevant across all scenarios. Additionally, a distinction was made between critical and non-critical scenarios, with each attribute result aggregated separately for these two types of scenarios.

To assess statistical differences per attribute between 2DS and 3DA, the Wilcoxon rank-sum test was used. The Wilcoxon rank-sum test is a nonparametric alternative to the two-sample t-test, which is based solely on the order in which the observations from the two samples fall. The predictor variable is the method (3DA or 2DS), and the dependent variables are the attributes.

Statistically significant differences in attribute rankings were found for the following attributes: *Physical Damage Non-Critical* ($M_V = 3.77$, $M_I = 3.58$, W = 547,345, p-value < 0.005), *Car Preservation Non-critical* ($M_v = 3.14$, $M_I = 3.81$, W = 6197, p-value < 0.0001), *Social Status Critical* ($M_v = 1.81$, $M_I = 1.56$, W = 40,455, p-value < 0.03), *Social Status Non-critical* ($M_v = 1.79$, $M_I = 1.47$, W = 38,107, p-value = 0.001214), *Vulnerability Critical* ($M_v = 3.51$, $M_I = 3.38$, W = 67,607, p-value < 0.003), *Vulnerability Non-Critical* ($M_v = 3.31$, $M_I = 3.19$, W = 69,961, p-value < 0.03), *Moral Responsibility Critical* ($M_v = 3.55$, $M_I = 3.12$, W = 3962.5, p-value < 0.02) (Tables 2 and 3).

Except for the Car Preservation attribute, all the statistically significant attributes have scored higher in 3DA than in 2DS. Figure 2 offers an overview of the scores of each attribute in 2DS and 3DA, in both critical and non-critical scenarios.

4.3 Overall Comprehension

4.3.1 Scenario Criticality Assessment

To evaluate how participants perceived the urgency of each scenario, they were asked to classify each scenario as either critical or non-critical following presentation (Q1: *"Do you perceive this scenario as critical? A critical scenario is defined as a scenario where the probability of harm is high)*. For scenarios designed as critical, 83% of participants in the 2DS condition classified them as critical, compared to 81% in the video-based condition. Conversely, for scenarios designed as non-critical, 60% of participants in the 2DS-based condition identified them as non-critical, while 50% of participants in the video-based condition made this assessment.

4.3.2 Scenario Relevance Assessment

Following each scenario presentation, participants evaluated the relevance of the situation for AV decision-making (Q2: "*Do you find this scenario to be relevant in identifying the attributes relevant in the population's decision-making for automated driving?")*. A noticeable difference was observed between presentation formats. In the 2DS-based scenarios, an average of 62% of participants judged the scenarios as relevant, 17% as potentially relevant, and 21% as not relevant for AV decision-making. In contrast, the video-based scenarios were perceived as relevant by 75% of participants, with just 12% considering them irrelevant. A chi-square test of independence indicated that this difference was statistically significant χ^2 (2, N = 4285) = 94.2944, p < 0.00001) where N represents the total number of scenario evaluations across all participants and conditions. This result suggests that video-based presentations were more frequently identified as relevant for eliciting moral attributes in AV decision-making.

4.3.3 Attribute Completeness Assessment

At the end of the experiments, participants were asked whether the provided attributes fully defined AV decision-making. In the 2DS-based scenarios, the responses were

Table 2 Average score of each attribute on Non-Critical Scenarios, with a 3D animations or 2D schematic representations media

Non-Critical	Physical damage	Car preservation	Psychological damage	Vulnerability	Social status	Moral responsibility	Legality of AV	Fair innings	Lottery	Arrival time
3D animations	**3.77**	**3.11**	3.12	**3.72**	**1.88**	3.23	3.15	2.52	2.03	2.27
2D schematic representations	**3.58**	**3.81**	2.82	**3.28**	**1.63**	2.92	3.08	2.63	2.03	2.34

Significant differences are highlighted in Bold

Table 3 Average score of each attribute on Critical Scenarios, with a 3D animations or 2D schematic representations media

Critical	Physical damage	Car preservation	Psychological damage	Vulnerability	Social status	Moral responsibility	Legality of AV	Fair innings	Lottery	Arrival time
3D animations	4.63	2.34	3.69	**3.67**	**1.98**	**3.55**	2.83	3.01	2.48	1.84
2D schematic representations	4.54	2.01	3.70	**3.54**	**1.86**	**3.12**	2.46	3.11	2.39	1.99

Significant differences are highlighted in Bold

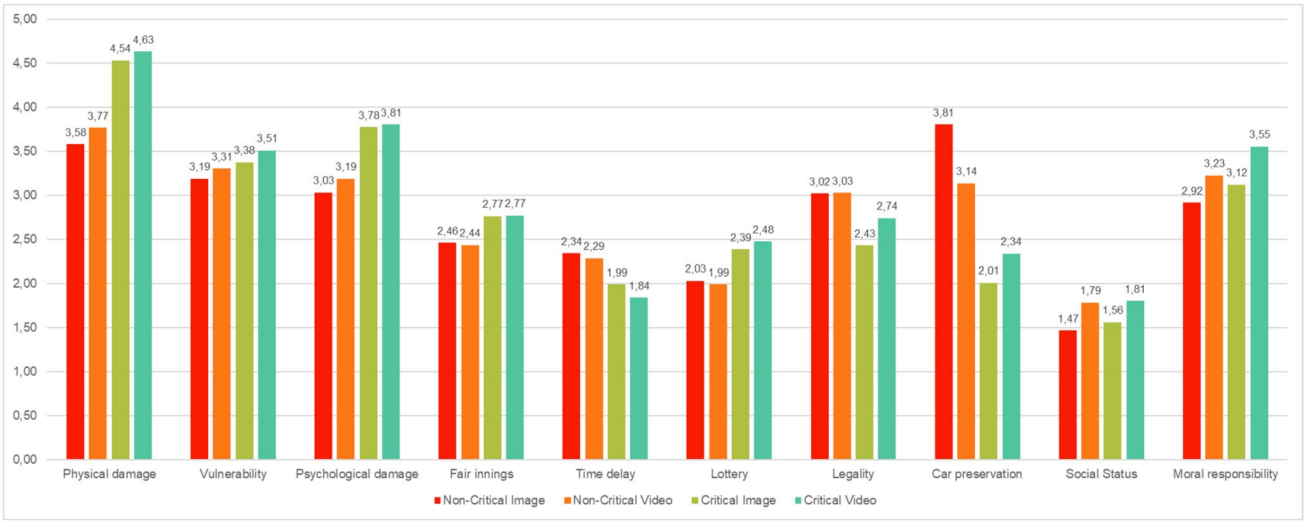

Fig. 2 Results of attribute scores distinguishing 2D schematic representations and 3D animations experiment, and critical and non-critical scenarios

approximately balanced, with 52 participants agreeing and 51 disagreeing. In contrast, in the video-based experiment, approximately one-third of the participants (65) agreed that the attributes completely defined AV decision-making, and only 36 disagreed. The results show a chi-square statistic of χ^2 (1, N = 204) = 4.0114, with a p-value of 0.045. This indicates that the difference in agreement rates between the two conditions is statistically significant at the 0.05 level.

4.3.4 Proposal of New Attributes

When asked about missing attributes Q3: *"Do you think that these attributes completely define the AV's decision-making? If not, what would be missing?")*, participants in both experiments proposed environmental damage (i.e., damage to the surroundings) and sustainability as additional attributes. In both experiments, 8 participants suggested relevant new attributes. In the 2DS experiment, more participants (14) were unsure about which attributes were missing after expressing dissatisfaction, compared to 10 in the video experiment. Additionally, a similar number of participants in each experimental format suggested variations of existing attributes, with 17 in the video experiment and 16 in the 2DS experiment.

5 Discussion

The results of this study provide evidence that supports the hypothesis of greater utilitarianism and comprehension in responses when participants are exposed to 3DA compared to 2DS. This section will discuss the key findings and their implications in the context of the hypotheses.

5.1 Greater Utilitarian Responses in 3D Animations

The hypothesis predicted that 3DA would induce greater utilitarianism in participant responses. The results indicate significant differences in attribute ratings across several scenarios.

The findings indicate that participants exposed to 3DA rated attributes such as Physical Damage, Vulnerability and Moral Responsibility as being higher in relevance compared to those exposed to 2DS, while rating the Car Preservation attribute as less relevant in 3D video. Time Delay was also rated lower in video scenarios than in 2DS. This difference is not necessarily due to participants perceiving the speeds in the scenarios as faster, as the time delay would occur after an incident happens, which was not shown in the video.

Physical Damage and Vulnerability: The higher ratings for Physical Damage and Vulnerability in 3DA suggest that participants are more attuned to the potential harm and vulnerability of individuals when presented with dynamic visual information. This heightened awareness likely influences their decisions towards minimising harm, a core negative utilitarian principle.

Car Preservation: The lower rating for Car Preservation in non-critical 3DA indicates that participants may prioritise human safety over material preservation when exposed to more immersive stimuli. This shift reflects a utilitarian focus on the greater good rather than individual or material interests.

Social Status and Moral Responsibility: The increased relevance of Social Status and Moral Responsibility in 3DA suggests that participants consider the broader societal impact and equitable distribution of benefits and burdens. This aligns with utilitarianism's goal of achieving the best overall outcome for society.

Unlike hypothesised, the Fair Innings attribute did not yield greater results with 3DA than 2DS. One possible

explanation is that its relevance may be less influenced by increased immersion. Unlike attributes such as Physical Damage or Perceived Vulnerability, which can be more directly visualised and emotionally engaged within dynamic scenarios, Fair Innings is an abstract concept based on life-course considerations rather than immediate situational cues. As a result, participants may rely on their pre-existing moral beliefs rather than being significantly influenced by the presentation format. Additionally, if 3DA heightened participants' focus on immediate harm and risk factors, attributes tied to long-term fairness considerations might have been comparatively deprioritised in their decision-making.

In conclusion, this study supports the hypothesis that 3DA induce greater utilitarianism in responses compared to 2DS in automated vehicle and human interaction scenarios. The dynamic nature of 3DA leads participants to prioritise attributes that focus on minimising harm, reflecting a utilitarian approach to decision-making.

5.2 Confidence and New Attributes

This study also supports the notion that increased realism in experimental settings may enhance the elicitation of moral attributes, making it easier for participants to recognise morally relevant factors, interpret social and environmental cues, and consider the real-world implications of their decisions.

5.2.1 Scenario Criticality Assessment

The results of the scenario criticality assessment (Q1) reveal two important patterns in scenario comprehension across presentation formats. First, the similar rates of correct identification for critical scenarios (83% 2DS vs. 81% 3DA) suggest that both 2D schematic and 3D video formats effectively communicate situations of high urgency or danger. This consistency indicates that critical scenarios maintain their salience regardless of presentation medium.

However, the substantial difference in correctly identifying non-critical scenarios (60% vs. 50%) points to a potential bias in the video-based presentation. Participants viewing 3DA were more likely to incorrectly classify non-critical situations as critical, suggesting that the dynamic nature of 3DA may heighten perceived risk or urgency even in relatively benign scenarios. This "criticality bias" in 3DA could be attributed to several factors:

- Temporal dynamics in videos may amplify perception of potential risks.
- Motion and sound cues might trigger heightened alertness.
- The immersive quality of video may increase emotional engagement.

This differential pattern between critical and non-critical scenario identification highlights an important methodological consideration: while 3DA may enhance ecological validity and perceived relevance, it may also introduce a tendency to perceive scenarios as more critical than intended.

This phenomenon has been documented in existing literature, where studies indicate that highly immersive environments can markedly amplify emotional responses, thereby affecting cognitive assessments of scenario criticality [20, 23]. This is further supported by the theory of constructed emotion, which posits that emotions are not hardwired responses but are constructed from core affective experiences and contextual interpretations [4]. In immersive settings, the vivid sensory input and heightened arousal can lead participants to "construct" emotional experiences that increase the perceived significance of events. Thus, while 3D animations can enhance engagement and memory, they may also increase the perceived criticality of certain scenarios [20]. This aligns with the understanding that our cognitive biases are deeply rooted in neuro-evolutionary processes that prioritise immediate, emotionally charged information—a mechanism that was advantageous for quick decision-making in ancestral environments. Thus, while 3D animations can enhance engagement and memory, they may also increase the perceived criticality of certain scenarios by tapping into these innate cognitive and emotional processes.

While it might be tempting to view the influence of emotions and cognitive biases as negative, they are essential to human cognition. Emotions and intuition provide the initial impetus for action, crucial in both everyday and complex scenarios. The bias lies not in the emotion itself but in its interpretation, which can vary across different cultural perspectives. Thus, these processes deepen our understanding of human cognition, even as we remain aware of their potential to skew perceptions.

5.2.2 Scenario Relevance Assessment

The notably higher relevance ratings (Q2) for video-based scenarios (75% vs. 62%) suggest that dynamic visual presentations significantly enhance participants' ability to contextualise automated driving challenges. This 13 percentage point increase in perceived relevance indicates that 3DA may better capture the complexity and temporal aspects of driving situations that static schematic representations cannot convey. While greater complexity could, in some cases, lead to lower relevance if it overwhelms participants or introduces ambiguity, the reduced uncertainty (fewer "maybe" responses) and decreased irrelevance ratings in the video condition suggest that, in this context, the added detail provided by 3DA enhances rather than hinders participants' ability to assess the scenarios.

5.2.3 Attribute Completeness Assessment and Proposal of New Attributes

The video format led to higher agreement among participants that the given attributes fully defined AV decision-making (Q3), indicating a clearer understanding of the decision-making process. Both formats saw participants propose environmental damage and sustainability as additional attributes, but the 3DA had fewer dissatisfied participants and proportionally more relevant new attribute suggestions. This aligns with the hypothesis that 3DA would lead to more new attributes or less confusion.

The similar number of new attributes proposed in both formats suggests that while both allowed for attribute refinement, the video format facilitated better comprehension, as fewer participants were unsure about missing attributes.

Overall, the 3DA' realism improved participants' ability to comprehend and propose moral attributes for AV decision-making, underscoring the value of immersive settings in eliciting nuanced moral judgments.

5.3 Limitations and Further Research

The 3DA, while more immersive than 2DS, still do not fully replicate the complexity of real-world driving situations. This methodological constraint stems from several key challenges in capturing the nuanced, dynamic nature of driving environments.

3DA, despite their enhanced visual fidelity, were not designed in this experiment to fully simulate the unpredictable variables of real-world driving, such as subtle environmental cues, peripheral awareness, and complex interactions between multiple road users. While incorporating these elements is possible, it would require significantly more effort and computational resources.

In addition, two-dimensional video presentations inherently restrict depth perception and spatial judgment. In addition, two-dimensional video presentations inherently restrict depth perception and spatial judgment. Although referred to as "3DA," these animations are still displayed on a flat, two-dimensional screen, meaning viewers do not experience true stereoscopic depth as they would in a real-world or virtual reality (VR) environment.

Future studies could explore the use of VR or augmented reality to create even more realistic environments. These technologies offer potential advantages, including immersive 360-degree perspectives, interactive scenario generation, and precise manipulation of environmental variables.

However, the increased realism of VR also presents challenges. VR environments can lead to heightened perspective-taking and emotional engagement, which may compromise the participants' ability to maintain a societal perspective [18]. Further VR studies should therefore incorporate mitigation strategies to control immersion and embodiment, as well as perspective-taking assessments.

Additionally, while the findings indicate that 3DA tend to induce more utilitarian responses in moral decision-making compared to 2DS, it is important to emphasise that this effect cannot be attributed solely to the 3D format itself. Rather, the observed differences result from the broader mode of presentation, which encompasses multiple interrelated factors. These include the increased level of perceptual detail, dynamic movement, realism of characters, perceived urgency through speed, and the salience of individual characteristics such as age, gender, or mobility. Each of these elements may independently or jointly influence participants' moral evaluations, making it difficult to isolate the effect of 3D as a single variable. Future research should aim to disentangle these components to better understand which aspects of immersive media most significantly shape moral judgments.

6 Conclusion

This study demonstrates that the method used to elicit moral attributes significantly influences the outcomes of AV decision-making models. 3D animations, with their dynamic and immersive nature, fostered a deeper understanding of moral dilemmas and led to more utilitarian responses than 2D schematic representations. Participants exposed to 3D animations prioritised attributes that focused on minimising harm, aligning with utilitarian principles. The video format also facilitated better comprehension and proposal of new attributes, indicating its effectiveness in capturing nuanced moral judgments. These findings highlight the importance of representation in moral elicitation and suggest that optimising scenario presentation can help ensure that AV decision-making models align more closely with societal values.

Building on these findings, the results suggest that 3D animations may serve as a more effective method for eliciting moral attributes in AV ethics research, due to their capacity to convey contextual detail and temporal dynamics. However, the immersive nature of such formats may also introduce systematic variation in participants' perceptions, such as an increased tendency to classify scenarios as critical. While this effect is not inherently negative, since emotions and intuition play a crucial role in human cognition and decision-making, to improve the reliability and comparability of future data collection efforts, researchers are encouraged to develop standardised protocols that include clear scenario design, control for potential presentation-induced biases, and document the influence of multisensory cues on moral evaluations. These measures can contribute to the development of ethically informed and socially responsive AV decision-making models.

Author Contributions Chloe Gros designed the study, conducted the data collection and analysis, and drafted the manuscript. Peter Werkhoven, Leon Kester and Marieke Martens contributed to the study

design, interpretation of results, and manuscript revisions. All authors read and approved the final manuscript.

Funding This research was supported by TNO. The funding body had no role in the design, execution, interpretation, or writing of the study.

Data Availability The datasets generated and analysed during the current study are available from the corresponding author on reasonable request. A curated version of the data package including anonymised responses is available via https://doi.org/10.24416/UU01-YD5C7I.

Declarations

Conflict of interest We have no known conflict of interest to disclose.

Ethics Approval and Consent to Participation All participants provided informed consent prior to participating in the study, in accordance with the Declaration of Helsinki.

Consent for Publication Not applicable. No identifying personal data is included in this manuscript.

References

1. Aliman N-M, Kester L. Requisite variety in ethical utility functions for AI value alignment. In: CEUR Workshop Proceedings (Eds), IJCAI 2019 AI safety workshop. CEUR WS (2019) https://doi.org/10.48550/arXiv.1907.00430
2. Aliman N-M, Kester L. Crafting a flexible heuristic moral meta-model for meaningful AI control in pluralistic societies. In: Wernaart B, editor. Moral design and technology. Wageningen Academic; 2022. p. 63–80. https://doi.org/10.3920/978-90-8686-922-0_4.
3. Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The moral machine experiment. Nature. 2018;563(7729):59–64. https://doi.org/10.1038/s41586-018-0637-6.
4. Barrett LF. The theory of constructed emotion: an active inference account of interoception and categorization. Soc Cogn Affect Neurosci. 2017;12(1):1–23. https://doi.org/10.1093/scan/nsw154.
5. Beauchamp TL, Childress JF. Principles of biomedical ethics. 8th ed. Oxford University Press; 2019.
6. Benvegnu G, Pluchino P, Garnberini L. Virtual morality: using virtual reality to study moral behavior in extreme accident situations. In: IEEE virtual reality and 3D user interfaces (VR). IEEE; 2021. p. 316–25. https://doi.org/10.1109/VR50410.2021.00054.
7. Bergmann LT, Schlicht L, Meixner C, König P, Pipa G, Boshammer S, et al. Autonomous vehicles require socio-political acceptance—an empirical and philosophical perspective on the problem of moral decision making. Front Behav Neurosci. 2018;12:31. https://doi.org/10.3389/fnbeh.2018.00031.
8. Bonnefon JF, Shariff A, Rahwan I. The social dilemma of autonomous vehicles. Science. 2016;352(6293):1573–6. https://doi.org/10.1126/science.aaf2654.
9. Cecchini D, Brantley S, Dubljević V. Moral judgment in realistic traffic scenarios: moving beyond the trolley paradigm for ethics of autonomous vehicles. AI Soc. 2023;40:1037–48. https://doi.org/10.1007/s00146-023-01813-y.
10. de Melo CM, Marsella S, Gratch J. Risk of injury in moral dilemmas with autonomous vehicles. Front Robot AI. 2021;7:572529. https://doi.org/10.3389/frobt.2020.572529.
11. Dubljević V, Racine E. The ADC of moral judgment: opening the black box of moral intuitions with heuristics about agents, deeds, and consequences. AJOB Neurosci. 2014;5(4):3–20. https://doi.org/10.1080/21507740.2014.939381.
12. Faulhaber AK, Dittmer A, Blind F, Wächter MA, Timm S, Sütfeld LR, et al. Human decisions in moral dilemmas are largely described by utilitarianism: virtual car driving study provides guidelines for autonomous driving vehicles. Sci Eng Ethics. 2019;25(2):399–418. https://doi.org/10.1007/s11948-018-0020-x.
13. Francis KB, Howard C, Howard IS, Gummerum M, Ganis G, Anderson G, et al. Virtual morality: transitioning from moral judgment to moral action? PLoS ONE. 2016. https://doi.org/10.1371/journal.pone.0164374.
14. Gros C, Kester L, Martens M, Werkhoven P. Addressing ethical challenges in automated vehicles: bridging the gap with hybrid AI and augmented utilitarianism. AI Ethics. 2024;5:2757–70. https://doi.org/10.1007/s43681-024-00592-6.
15. Gros C, Werkhoven P, Kester L, Martens M. A methodology for ethical decision-making in automated vehicles. AI Soc. 2025. https://doi.org/10.1007/s00146-025-02370-2.
16. Ju U, Kang J, Wallraven C. To brake or not to brake? Personality traits predict decision-making in an accident situation. Front Psychol. 2019;10:134. https://doi.org/10.3389/fpsyg.2019.00134.
17. Li J, Zhao X, Cho MJ, Ju W, Malle BF. From trolley to autonomous vehicle: perceptions of responsibility and moral norms in traffic accidents with self-driving cars. SAE Technical Paper 2016-01-0164; 2016. https://doi.org/10.4271/2016-01-0164.
18. Maćkiewicz B, Wodowski J, Andrusiewicz J. Why do people seem to be more utilitarian in VR than in questionnaires? Philos Psychol. 2023;38(4):1702–30. https://doi.org/10.1080/09515089.2023.2282060.
19. McManus RM, Rutchick AM. Autonomous vehicles and the attribution of moral responsibility. Soc Psychol Personal Sci. 2019;10(3):345–52. https://doi.org/10.1177/1948550618755875.
20. Parong J, Mayer RE. Learning about history in immersive virtual reality: does immersion facilitate learning? Educ Technol Res Dev. 2021;69(3):1433–51. https://doi.org/10.1007/S11423-021-09999-Y.
21. Rawls J. Justice as fairness. NY: Harvard University Press; 2001.
22. Schein C, Gray K. The theory of dyadic morality: reinventing moral judgment by redefining harm. Pers Soc Psychol Rev. 2018;22(1):32–70. https://doi.org/10.1177/1088868317698288.
23. Visch VT, Tan ES, Molenaar D. The emotional and cognitive effect of immersion in film viewing. Cogn Emot. 2010;24(8):1439–45. https://doi.org/10.1080/02699930903498186.
24. Wernaart B. Developing a roadmap for the moral programming of smart technology. Technol Soc. 2021;64:101466. https://doi.org/10.1016/j.techsoc.2020.101466.
25. Wilson H, Theodorou A. Slam the brakes: perceptions of moral decisions in driving dilemmas. AISafety. 2019;2019:2419.