AI & SOCIETY
https://doi.org/10.1007/s00146-025-02370-2

OPEN FORUM

# A methodology for ethical decision-making in automated vehicles

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

Received: 14 January 2025 / Accepted: 11 April 2025


**Abstract**
Despite significant advancements in AI and automated driving, a robust ethical framework for AV decision-making remains undeveloped. Such a framework requires clearly defined moral attributes to guide AVs in evaluating complex and ethically sensitive scenarios. Existing frameworks often rely on a single normative ethical theory, limiting their ability to address the nuanced nature of human decision-making and leading to conflicting outcomes. Augmented Utilitarianism (AU) offers a promising alternative by integrating elements of virtue ethics, deontology, and consequentialism into a non-normative framework. Grounded in moral psychology and neuroscience, AU employs mathematical ethical goal functions to capture societally aligned attributes. In this study, we propose and evaluate a method to elicit these attributes for AV decision-making. One hundred participants were presented with traffic scenarios, including critical and non-critical situations, and tasked with evaluating the relevance of an initial set of 11 attributes (e.g., physical harm, psychological harm, and moral responsibility) while suggesting additional relevant attributes. Results identified two new attributes—environmental harm and energy efficiency—and revealed that four attributes (physical harm, psychological harm, legality of the AV, and self-preservation) varied significantly between critical and non-critical scenarios. These findings suggest that the weight of attributes in ethical goal functions may need to adapt to situational criticality. The method was validated based on key evaluation criteria: it demonstrated sensitivity by producing varying relevance scores for attributes, was deemed relevant by participants for eliciting AV decision-making attributes, and allowed for the identification of additional attributes, enhancing the robustness of the framework. This work contributes to the development of a dynamic and context-sensitive ethical framework for AV decision-making.

**Keywords** Automated vehicles · Morality · Ethics · Self-driving cars · Artificial intelligence ethics

## 1 Introduction

The responsible application of automated vehicles[1] (AVs) relies on their capability to include moral values in their ethical decision-making such that decisions are 'value aligned' with societal values. Theoretical and experimental research about ethical AV decision-making and its so-called moral dilemmas (i.e., the relative weight of moral attributes for a specific situation) are multiplying (Awad et al. 2018; de Melo et al. 2021; Faulhaber et al. 2019; Kallioinen et al.

✉ Chloe Gros
c.n.gros@uu.nl

Peter Werkhoven
werkhoven@wxs.nl

Leon Kester
leon.kester@tno.nl

Marieke Martens
m.h.martens@tue.nl

[1] Utrecht University, Utrecht, Netherlands

[2] Netherlands Organisation for Applied Scientific Research, Delft, Netherlands

[3] Eindhoven University of Technology, Eindhoven, Netherlands

[1] 1 Regulation (EU) 2019/2144 of the European Parliament and of the Council of 27 November 2019 on type-approval requirements for motor vehicles (…) defines a "fully automated vehicle" as a motor vehicle that has been designed and constructed to move autonomously without any driver supervision (REGULATION (EU) 2019/2144 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL 2019). In the SAE classification, we focus on automation levels 4 and 5, where human drivers are not required to intervene by the automated system (On-Road Automated Driving (ORAD) Committee 2021).

Published online: 30 April 2025

2019; Sütfeld et al. 2017). While these contributions provide valuable insights, existing methods still face challenges in defining a comprehensive set of moral attributes that capture societal values in AV decision-making.

While some studies aim at a first description of such an ethical framework (Awad et al. 2018; de Melo et al. 2021; Faulhaber et al. 2019; Kallioinen et al. 2019; Sütfeld et al. 2017), the attributes that they describe seem to lack scientific grounding. For example, in the Moral Machine experiment (Awad et al. 2018), thousands of participants chose their preferred outcome in numerous unfortunate events. Examples of attributes used for decision-making in this experiment are the number of individuals killed, their gender, age, and social status. However, the lack of scientific grounding makes it unclear how this set of attributes was chosen, which leads to a lack of understanding of the participants' decision-making process. For example, are children advantaged because they seem more vulnerable and unable to get out of the way in time? Or following a 'fair innings' philosophy of having the right to live a certain number of years? In addition, the scenarios randomly compared attributes, making it difficult to understand how each attribute influenced the participants' decision-making. Finally, it did not allow participants to contribute their own attributes nor tried to have an exhaustive set of attributes.

One notable framework is the Agent-Deed-Consequences (ADC) model of moral judgment developed by Cecchini and Dubljević (2025) and Pflanzer et al. (2022). This model posits that moral evaluations are derived from concurrent assessments of three components: the character of the agent (Agent), the nature of the action (Deed), and the outcomes produced (Consequences). While the ADC framework provides a valuable starting point by outlining these key components, it often lacks the granularity needed to specify the precise attributes that should be considered within each component.

In this paper, we develop a method for eliciting societally aligned moral attributes for AV-decision making based on a descriptive framework called augmented utilitarianism (AU) (Aliman and Kester 2019, 2022; Gros et al. 2024). Similarly to the ADC framework, AU is a non-normative framework grounded in moral psychology, cognitive neuroscience, and philosophy. It is compatible with Gray and Shein's theory of dyadic morality (Schein and Gray 2018) as applied in the medical domain. Dyadic morality postulates that perceived harm is based on a thinking agent causing damage to a vulnerable patient.

As such, AU combines virtue ethics (thinking agent perspective), deontology (action and damage perspective), and consequentialism (vulnerable patient perspective). In addition, AU introduces the concept of experiencer as a flexible moral authority that fills in the normative aspects of ethical decision-making, adapting guidelines to reflect contemporary societal values. This role enables AU to remain context-sensitive and responsive to evolving ethical standards, avoiding rigid or absolute moral paradigms. AU positions itself as a descriptive and explanatory framework that aims at capturing the nuances in human morality's pluralism rather than aiming at a normative framework.

Thus, AU aims at increasing public trust in AVs by ensuring that their ethical decision-making is transparent, explainable, and aligned with societal values. Public acceptance of AVs is not only contingent on their technical reliability but also on how their ethical choices are perceived and understood. However, recent research on generative AI highlights a growing challenge: the spread of misinformation and its influence on user information processing (Shin et al. 2024). Misconceptions about how AVs make ethical decisions could erode trust, particularly if the public is exposed to misleading narratives about AV behaviour in critical situations.

Furthermore, the interaction between humans and algorithmic decision-making systems is shaped by artificial misinformation, which can distort perceptions of fairness, responsibility, and risk assessment (Shin 2024). Just as misinformation can alter how people interpret AI-generated content, it may also impact how they judge AV decision-making processes. If the public perceives AVs as making arbitrary or morally unacceptable choices—whether due to genuine ethical concerns or misinformation—adoption and regulatory acceptance could face significant barriers.

Addressing these challenges, AU is designed to counteract misinformation and enhance public trust by ensuring that AV decision-making is both explainable and grounded in societal values. Its structure allows for transparent reasoning behind ethical choices, making it possible to clarify why an AV acted in a certain way in a given scenario. By incorporating a feedback loop that continuously integrates societal perspectives, AU remains adaptable to evolving ethical expectations and public concerns. This responsiveness helps mitigate the risks of misinformation by ensuring that the ethical principles guiding AV decisions are not only technically robust but also widely understood and accepted.

AU is based on mathematical utility functions called ethical goal functions (EGF), which are composed of *attributes* and *values* that can be adjusted and determined by the relevant society. The simplest approach of such an EGF is to have linear individual attribute utility functions and a linear sum of utility functions. An example of such a function can be written in this form:

$$U(x) = \sum_i w_i u_i(A_i(x))$$

With $A_i$ the attributes, e.g., the patient's harm, the action's fairness, and the actor's cost; $u_i$ being the utility

function of the attribute $A_i$, and $w_i$ the relative weight indicating the relevance of the attribute, as a function of world state $x$.

Once specified, these functions can be tested and adjusted based on feedback, making AU a dynamic process that is referred to as a 'socio-technological feedback loop' (Aliman et al. 2019). The goal of this paper is to operationalize the AU framework by defining and evaluating a method to identify a set of societally aligned attributes for AV decision-making.

This set of societally aligned attributes can eventually apply to all possible situations without presenting all possible scenarios to participants, as described in AU. The main benefit of the method is to be *explanation-based* and *updatable*. This means that the attributes and weights of the EGFs must be explainable so that it can be clearly understood why the AV came to this decision if acted upon. In addition, due to the feedback loop, the EGFs are updatable, being able to be adjusted if needed if societal values evolve over time. To this aim, the initial set of attributes that we have selected is grounded in literature, through the theory of dyadic morality (Schein and Gray 2018) and by using biomedical ethics (Beauchamp and Childress 2019) as a basis, as it is a well-established and accepted ethical framework (Coin and Dubljević, 2021; Givens 2013). This set is then put through a feedback loop in the shape of a paper-based questionnaire, in which individuals have the opportunity to express their opinions on it as well as to add additional attributes that may have been missed.

## 2 Method

Our method consists of (1) defining an initial set of attributes using AU and biomedical ethics adapted to AVs, and (2) updating the initial set by scenario-based attribute ranking and supplementation with missing attributes.

Updating the set of attributes is done with feedback from experiencers, who are individuals selected to represent societal values and perspectives, appointed by the moral authority—typically a governance body or legislative entity. These experiencers ensure that the attributes remain contextually relevant and aligned with evolving societal expectations. Through scenario-based evaluations, they rank the importance of existing attributes and propose additional ones where necessary. This feedback loop enables a dynamic refinement process, ensuring that the attributes accurately reflect collective ethical concerns. By integrating experiencer input, the framework bridges theoretical models and practical applications, creating an adaptive and socially responsive ethical system for AVs.

### 2.1 Attribute selection

As a starting point for our attribute selection, we focus on biomedical ethics. Biomedical ethics offers a well-established framework centered on minimizing harm and prioritizing human safety, aligning closely with the goals of AV ethics. Its emphasis on decision-making in life-or-death situations and high-pressure scenarios provides practical, transferable principles for defining key attributes in AV ethical models (Beauchamp and Childress 2019; Givens 2013). Biomedical ethics were first described in Principles of Biomedical Ethics (Beauchamp and Childress 2019) and is based on the application of certain moral principles to examine moral dilemmas. This framework provides a practical method of dealing with real-world ethical dilemmas (Hain and Saad 2016) and is now widely accepted by society (Coin and Dubljević, 2021; Givens 2013; Rus and Groselj 2021). Its structure allows cross-mapping scarce resource-related dilemma situations from the medical domain to the AV domain. For example, an AV having to choose between risking injuring a child or an adult can be seen as similar to a doctor having to either save the life of a child or an adult by allocating scarce medication.

The first step is to divide the attributes into different levels, that represent the larger category to which they apply (see Fig. 1). Level 1 represents, following the theory of dyadic morality, the entity to which the attribute is associated (patient, action, or agent) (Schein and Gray 2018). Level 2 represents the fundamental trade-off between Utility, Cost, and Fairness, as described in biomedical ethics adapted to AVs (Beauchamp and Childress 2019).

*Utility* represents the Patient put at risk and is based on utilitarian theories prescribing principles and rules that maximize utility or welfare, by preventing or removing harm and promoting good. In biomedical ethics, it comprises of medical need and social utility. *Medical need* represents the urgency to which treatment is needed, and therefore the harm that will be endured by the individual without treatment. In the context of AVs, *Harm* represents all kinds of damage that can be dealt to an individual. It comprises therefore of *physical damage*, *psychological harm*, and *perceived vulnerability*. S*ocial Utility* can be seen as maximizing the overall utility of society by using social status to determine if certain populations have a better utility than others. For example, it can be argued that young adults should be privileged over seniors as they will benefit society by working for several years, or that health workers should be given priority (Beauchamp and Childress 2019). For AVs as for biomedical ethics, it could be interpreted as an incentive to protect a particular part of the population, for example, health workers.

In biomedical ethics, opposite the harm endured by the patient without treatment, are the *Costs* involved in the treatment itself. This includes the financial cost of the treatment

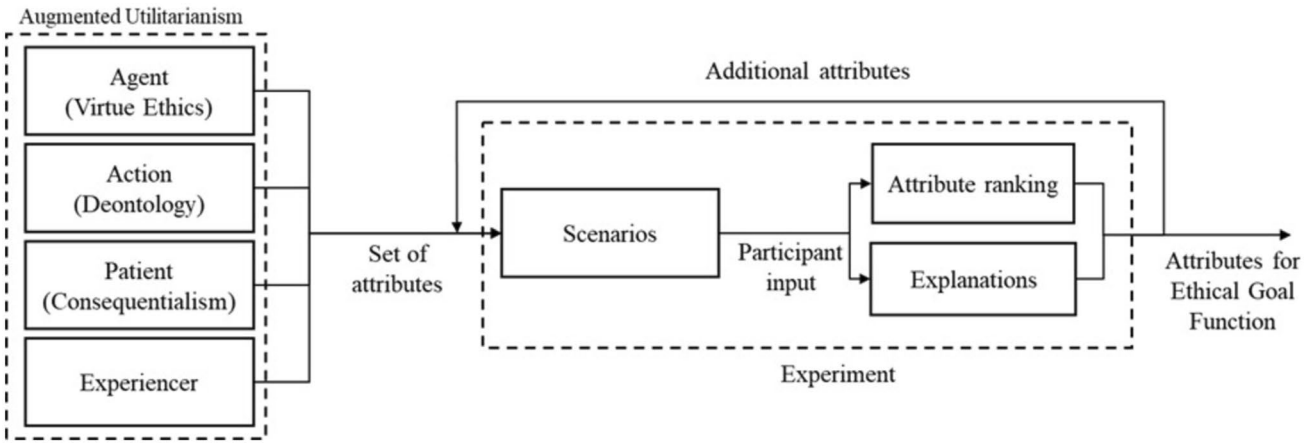


**Fig. 1** Illustration of the experimental method

as well as the risks associated with the treatment, which are often weighted against the benefits of getting the treatment. As such, in the context of AVs, safety is often weighted against the inherent goals of the car, which are to arrive at the destination intact and in a reasonable time. We then define as *Costs* the following attributes: *Timeliness* (Time of Arrival) and *Financial Cost* (self-preservation), the costs involved in repairing the car if it is damaged.

Finally, *Fairness* represents justice in the attribution of treatment, or liability. It is subdivided into *Liability* and *Fair Opportunity*. In biomedical ethics, it can be argued that patients can forfeit their rights to healthcare if their ill health results from their own actions (not being vaccinated, smoking, etc.). Liability of the patient, or *Moral Responsibility*, represents the required risk shifting when a patient is responsible by their actions for creating a risky situation. In the context of AVs, it can apply to someone crossing the road without looking or a red light violation. *Liability of the Action* represents the compliance of the action of the AV to safe rules of conduct, usually considered as being the law of the road. *Fair Opportunity* asserts that individuals should not receive social benefits based on undeserved advantageous properties and should not be denied social benefits based on underserved disadvantageous properties, because these individuals are not responsible for these properties. One *Fair Opportunity* rule associated with Patients is the rule of *fair innings*. It states that everyone has an equal chance to live a certain amount of time, after which they are no longer entitled to receive social care, including healthcare. In terms of Action, one mechanism of Fair Opportunity is, if no major disparities exist in utility for patients, to utilize a lottery system.

To summarize, our attributes are classified as follows (Table 1):

Level 5 attributes represent different possible instances of Level 4 attribute categories. In the present experiment, we report relevance scores at Level 4, as it explains decision-making, whereas Level 5 attributes can relate to different Level 4 attributes and are therefore not sufficient for explanation-based decision-making (for example, the Level 5 instance "Age" can relate to fair innings or perceived vulnerability).

Some attributes, such as social utility, fair innings or lottery, have been proposed for biomedical ethics but can be controversial and not widely accepted. As a means of exhaustivity, we have included them in our framework, which will allow them to be tested against AV scenarios, where they may or may not be considered relevant.

## 2.2 Experiment design

A sample size of 100 participants was chosen following variation of the results of a pilot experiment not reported here, to hold the role of experiencer. Participants were recruited on campus through posters, word of mouth, or were directly approached, and compensated 10€ for participation. The sample includes 46 females, 50 males, and four non-binary/prefer not to say. Seventy three participants were between 18 and 24 years old and 27 were between 25 and 34 years old.

For each Level 4 attribute, one or multiple instances (level 5) were selected. To cover all Level 4 attributes, 11 instances of attributes were tested: family status, gender, child, senior, disabilities, profession, the legality of the action, lottery, moral responsibility of the pedestrian, self-preservation, and time delay.

In the existing AV ethics studies, the scenarios encountered are very often critical (life or death situations) (Awad et al. 2018; de Melo et al. 2021; Faulhaber et al. 2019; Kallioinen et al. 2019; Sütfeld et al. 2017). However, these situations are exceptions and do not reflect the usual usage of the AV. Some attributes, such as self-preservation of the car or legality of the action of the AV, could be ranked differently

**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 (car) | Cost | Timeliness | Time of Arrival | / |
| | | Financial cost | Self-preservation | / |

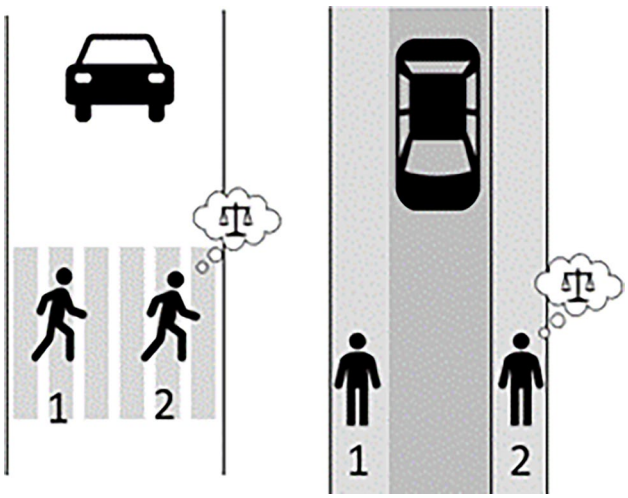


**Fig. 2** Critical and non-critical scenarios for the Social Status attribute, using the Profession Level 5 instance (one pedestrian is a lawyer)

for these situations. As an additional contribution, we, therefore, separated the scenarios into critical and non-critical scenarios, where critical scenarios are scenarios in which the probability of harm is high (life or death situations), and non-critical scenarios are scenarios in which the probability of harm is low (usual driving situations). For each instance, two scenarios, one critical and one non-critical, were created (see Fig. 2).

For each scenario, an explanation is given to the participants, such as "An empty AV is driving on a road. It comes across a crosswalk where 2 pedestrians are crossing, one of them is a cashier in a supermarket (1) and the other one is a lawyer (2). The 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". Participants are then asked to score the relevance of each of the Level 4 attributes on a Likert scale, from 1. Not at all, to 5. Completely, for each of the scenarios. All of the attributes were asked in each scenario to mitigate specific interaction effects.

For instances that relate to different attributes, for example, Family Status relates to psychological damage and social status, both attributes are then taken into account. It would not involve a trade-off between these two attributes because they are not compared to each other.

To verify that the scenarios described were pertinent, participants were asked if they thought that the scenario was relevant in identifying the attributes relevant to the population's decision-making for automated driving. They were also asked if they believed the scenario was critical or not. As we did not clearly state a definition of critical, the participants could self-define the word.

As the experiment is explanation-based, after each question, the participants were asked to give explanations about their decision-making. This helped us 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.

As the set of attributes has to be updatable to adapt to society's evolving values, participants had multiple opportunities to elicit new potential attributes. First of all, 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. Finally, they had another opportunity to add any other remarks that they believed could be relevant.

## 3 Results

For each scenario, the score of every attribute was computed using its average. The overall score of each attribute is then obtained by taking its maximum score on its relevant scenarios. For example, the moral responsibility attribute has only one associated scenario. Therefore, its overall score is the one obtained in that scenario. However, the social utility attribute can relate to several different scenarios (profession, gender, family status). In this case, the highest score across scenarios is kept, as our goal is to identify if an attribute is ever relevant.

The evaluation criteria for the method include determining its sensitivity (i.e., ensuring that attributes receive varying relevance scores rather than uniform ratings), assessing whether participants find the scenarios relevant for eliciting attributes for AV decision-making, evaluating whether participants believe the proposed attributes fully define AV decision-making, and, if not, whether they can contribute additional attributes to enhance the framework.

### 3.1 Attribute scores

The significance of the differences in results between non-critical and critical scenarios was assessed using a Mann–Whitney *U*-test, chosen due to the non-normal distribution of attribute scores. This non-parametric test evaluates whether there is a statistically significant difference in the distribution of scores between the two scenario types. The null hypothesis (H0) states that there is no significant difference in the average attribute scores between non-critical and critical scenarios. To ensure robustness, the distribution of the data was examined before selecting the test, and effect sizes were also considered to complement the *p* value analysis (Table 2).

We will not be discussing the choice to keep an attribute or to remove it completely as (1) the goal of the experiment was to evaluate the proposed methodology and not to draw conclusions on specific attributes, (2) even if an attribute has a low score, it could still be relevant, (3) the decision to take into account a specific attribute or not is ultimately the decision of the legislator, and we do not have the legitimacy to decide on this and (4) we used a specific and potentially biased sample of students and university staff.

**Table 2** Attribute scores for non-critical and critical scenarios

| | | Physical damage | Psychological damage | Perceived vulnerability | Moral responsibility | Fair innings | Legality av | Lottery | Self-preservation | Arrival time | Social utility |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Critical scenarios | Count | 96 | 100 | 94 | 97 | 96 | 96 | 96 | 92 | 100 | 97 |
| | Mean | 4.61 | 3.69 | 3.55 | 3.11 | 3.08 | 2.42 | 2.39 | 2.01 | 1.99 | 1.86 |
| | std | 0.77 | 1.01 | 1.32 | 1.32 | 1.53 | 1.30 | 1.61 | 1.22 | 1.23 | 1.22 |
| Non-critical scenarios | Count | 95 | 96 | 96 | 95 | 95 | 98 | 96 | 93 | 97 | 99 |
| | Mean | 3.91 | 2.81 | 3.27 | 2.89 | 2.59 | 3.07 | 2.03 | 3.80 | 2.37 | 1.63 |
| | std | 1.40 | 1.35 | 1.36 | 1.35 | 1.59 | 1.37 | 1.48 | 1.21 | 1.29 | 1.05 |
| | *p* value | **0.000** | **0.000** | 0.141 | 0.282 | 0.028 | **0.001** | 0.071 | **0.000** | 0.029 | 0.219 |

Significant *p*-values highlighted in bold

If we first look at critical scenarios, which are the ones that are most often studied in literature, we can notice that *physical damage* is the highest-ranked attribute ($M_C = 4.61 \pm 0.77$). This was expected, as this is one of the most studied attributes, notably in experiments studying utilitarianist theories (Awad et al. 2018; Bergmann et al. 2018; Bonnefon et al. 2016; Faulhaber et al. 2019; McManus and Rutchick 2019). Our results show a similar tendency for reducing total damage, following utilitarian ethics.

The second highest ranking attribute *is psychological harm* ($M_C = 3.69 \pm 1.01$). This attribute has never been explicitly studied in itself before, however, instances of it were used, such as the difference between adults and children (Bergmann et al. 2018; Faulhaber et al. 2019; Kallioinen et al. 2019; Li et al. 2019). In our experiment, according to the explanations that participants gave, family status seems to be the most important factor for psychological damage: "A man with a family will probably provoke more psychological damage to a higher number of people than a man without family". However, family status has never been studied before as an attribute for AVs. Moreover, when previous experiments studied the differences between children and adults, it was not clear what Level 4 attribute was being investigated, as they were not explanation-based experiments. Therefore, no comparison is possible with other studies, and this factor should be investigated further.

The third highest ranking attribute is *perceived vulnerability* ($M_C = 3.55 \pm 1.32$), here exemplified by situations of disabilities. The explanation that is mostly given is that people with a higher perceived vulnerability, notably in terms of a disability, are less able to move out of the way and possibly avoid the car and should therefore be protected. Bergmann and colleagues show similar results in their experiment (Bergmann et al. 2018), in which they look at the case of someone who is kneeling and therefore appears more vulnerable. In that case, 62% of participants chose to save the kneeling person. Our results are similar to the ones from Bergmann et al. and extend them to other types of perceived vulnerabilities.

*Moral responsibility* ($M_C = 3.11 \pm 1.32$) has been studied before, notably in the Moral Machine experiment (Awad et al. 2018) and in the experiment from Bergmann et al. under "involvement in traffic" (Bergmann et al. 2018). Similar to previous experiments, participants seem to consider it a relevant factor when casualties are involved. Moral responsibility has also been studied as a philosophical concept, notably by Kauppinen, who states that "an agent who is capable of doing so is morally required to shift a risk of harm from a non-liable party to a liable party and not to shift the risk from a liable to a non-liable party" (Kauppinen 2020).

Age has been studied as an attribute in numerous experiments (Awad et al. 2018; Bergmann et al. 2018; Diederich et al. 2011; Faulhaber et al. 2019; Johansson-Stenman and Martinsson 2008; Kallioinen et al. 2019), in which participants have a clear tendency to protect children. However, different Level 4 attributes can relate to age, such as *fair innings* ($M_C = 3.08 \pm 1.53$), but also perceived vulnerability and psychological damage. Our experiment provides more clarity on this by explicating all the attributes that may come into play in scenarios involving children and adults. Our results suggest that fair innings are one of the reasons why people tend to protect children and can be considered an independent attribute.

The *legality of the action of the AV* seems to be especially relevant in non-critical situations ($M_{NC} = 3.07 \pm 1.37$), but when lives are at stake, it seems to be of less importance ($M_C = 2.42 \pm 1.30$). The action of the AV has been studied in the Moral Machine experiment (Awad et al. 2018), where it scores very low for critical situations, similar to our experiment. There seems to be a usual preference for saving lives rather than obeying the law. However, it scores significantly higher in non-critical situations ($p < 0.001$), meaning that respecting the law still seems important if the probability of harm is generally low.

*Lottery* ($M_C = 2.39 \pm 1.61$) has not been explicitly studied in previous studies but corresponds to a "random choice" option. Here, we have made it an explicit option, so that participants can clearly show that they do not want the AV to make a deliberate decision. A lottery system is generally advocated when using personal features is prohibited and when all other parameters are similar. However, participants do not seem generally favorable to a lottery system. In real life, situations where all other attributes would be equal are virtually impossible, meaning that a lottery system may never be required.

*Self-preservation* ($M_C = 2.01 \pm 1.22$) is the attribute where the difference between critical and non-critical situations is the most significant. Understandably, participants generally felt that protecting lives was a lot more important than protecting the car. However, in non-critical situations, self-preservation was generally considered to be a relevant attribute ($M_{NC} = 3.80 \pm 1.21$, $p < 0.001$). In literature, the only experiment looking at self-preservation is from Faulhaber et al. (Faulhaber et al. 2019). However, only critical situations involving the driver's life and a pedestrian's life have been studied, which makes it a question of perspective more than of self-preservation of the AV (Faulhaber et al. 2019). In this experiment, we solely look at the material value of the car.

*Arrival time* ($M_C = 1.99 \pm 1.23$), is one of the lowest scored attributes. Participants generally argued that safety and reducing potential harm should always be a priority over arrival time. However, the goal of the car is to go from point A to point B in a time-efficient manner, otherwise, people could walk, cycle, or travel in slower modes everywhere.

Arrival time should therefore be used as the goal of the car and as a trade-off with safety.

In our experiment, *social status* ranks last both in critical ($M_C = 1.86 \pm 1.22$) and non-critical ($M_{NC} = 1.63 \pm 1.05$) situations. This includes scenarios involving profession and gender. Participants overwhelmingly believe that social status should not be used by AVs to make ethical decisions. Social status has been extensively studied, in the Moral Machine experiment using the instances of profession, fitness, and gender (Awad et al. 2018), but also in the experiment of Sütfeld et al. (Sütfeld et al. 2019) using gender, and in the experiment of Wilson and Theodorou (Wilson and Theodorou 2019). Our results are similar to the ones from Sütfeld and Wilson and Theodorou's experiments, where social status did not seem to have an impact on the participant's choices. However, in the Moral Machine experiment, social status has a higher score, ranking as the 5th most important attribute. The difference can be explained by the fact that the participants in our experiments are students, who tend to have a more progressive mindset, as well as the cultural differences between the northern American population and the European population. In addition, the MIT Moral Machine experiment is set up in a way where participants only have a choice between two outcomes, which might push them to select a "least worse" option and would not necessarily reflect their true feelings.

Overall, participants assigned distinct scores to the attributes, demonstrating that the method meets the sensitivity criterion.

After each scenario, participants were asked whether they found the scenario relevant for identifying attributes for AV decision-making. The results indicate that the scenarios were generally perceived as relevant, with 80% of participants responding "Yes" or "Maybe" and only 20% responding "No" across all scenarios.

### 3.2 Additional attributes

At the end of the questionnaire, participants were asked if they believed that the given attributes completely defined AV decision-making. The results were very disparate, with an almost 50/50 split (51/49) between participants, who were then asked what they thought was missing from the set of attributes. The most common answer was to consider the surrounding environment, e.g., if it is in the middle of a city or on an empty road in the countryside, and the consequences of a crash on it. This was not previously explicitly described as an attribute, as there is no equivalent in biomedical ethics, and prompts us to add another subcategory to the Level 3 Harm category called 'environmental harm'. This includes all damage done to non-humans, including property or nature damage, as well as animals, and their potential financial consequences (for example, destroying a statue would lead to substantial financial reparations).

Another proposed attribute is energy efficiency. As for the previous attribute, this has no equivalent in biomedical ethics and is specific to the mobility sector. Environmental considerations are becoming more and more crucial as we need to considerably reduce our carbon emissions. Even if the AV is electric, the electricity it uses can come from fossil fuels and have a considerable environmental impact. Energy efficiency is therefore a necessary addition to our attributes, especially for non-critical situations. As it concerns the car in itself, it will be placed under Actor > Cost.

This leads us to the following updated attribute table (Table 3):

### 3.3 Critical vs. non-critical

When comparing the scores of attributes for non-critical and critical scenarios, four attributes obtain significantly different scores (see Fig. 2): physical damage ($M_{NC} = 3.91 \pm 1.40$, $M_C = 4.61 \pm 0.77$, $p < 0.001$), psychological damage ($M_{NC} = 2.81 \pm 1.35$, $M_C = 3.69 \pm 1.01$, $p < 0.001$), self-preservation ($M_{NC} = 3.80 \pm 1.21$, $M_C = 2.01 \pm 1.22$, $p < 0.001$), and legality of the AV ($M_{NC} = 3.07 \pm 1.37$, $M_C = 2.42 \pm 1.30$, $p < 0.001$). These values suggest that in life-or-death situations, most participants, following a utilitarian theory, prioritize saving human lives above material considerations such as self-preservation and respect for the law. Having significantly different results means that the weights of these attributes might change according to the situation.

In 2020, the COVID-19 pandemic not only showed the world that a global health crisis was still possible in the modern days but also redefined medical resource allocation and triage. The traditional pillars of biomedical ethics (beneficence, non-maleficence, justice, and autonomy), while widely accepted in non-critical situations, have been deemed not sufficient in conditions of scarce resources (Emanuel et al. 2020) and new triage frameworks have been proposed. The framework proposed by Emanuel et al. promotes allocation according to four values: 1. Giving priority to the worst off, 2. Maximizing benefits yielded by scarce resources, 3. Treating people equally, and 4. Promoting and rewarding instrumental value. This framework can be seen as a refinement of the one from Beauchamp and Childress, as it builds on it by emphasizing specific attributes (beneficence, justice, social utility). Similarly, we can define a specific function for critical situations that will accentuate the weights of certain attributes.

**Table 3** Revised attribute classification table

| 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 |
| | | | | Gender |
| | | | Environmental harm | Nature damage |
| | | | | Animal harm |
| | | | | Property damage |
| | | 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 | / |
| Actor (car) | Cost | Timeliness | Time of Arrival | / |
| | | Financial cost | Self-preservation | / |
| | | Environmental cost | Energy Efficiency | / |

## 4 Discussion

### 4.1 Results interpretation

This study explored the ethical decision-making attributes prioritized in automated vehicle (AV) scenarios, comparing non-critical and critical situations. The results indicate that participants weigh certain attributes differently depending on the context, suggesting that ethical preferences are not static but highly situational. These findings align with previous research on moral decision-making in AVs, which highlights the variability in human moral intuitions depending on perceived risk levels (Awad et al. 2018; Cecchini et al. 2023).

According to the evaluation criteria, our method proves effective in eliciting relevant attributes for AV decision-making. While only about half of the participants felt that the proposed attributes provided a comprehensive framework, the method allowed them to contribute additional attributes, enhancing the overall framework. The results also show that the method is sensitive, as participants assigned varying relevance scores to the different attributes, indicating that they were able to differentiate between more and less important factors. Furthermore, the majority of participants found the scenarios relevant for identifying attributes, suggesting that the method successfully engages participants in the process of attribute elicitation. These findings demonstrate that our method is effective in capturing key ethical considerations for AV decision-making.

One key observation is that in critical scenarios, participants appear to prioritize attributes related to immediate harm reduction, while in non-critical scenarios, considerations such as fairness and adherence to rules seem more prominent. This distinction suggests that ethical frameworks for AV decision-making may need to be adaptable, ensuring that AVs can respond dynamically to varying levels of risk. In addition, the results reinforce the argument that ethical decision-making in AVs should not rely on a fixed, universal hierarchy of attributes but instead incorporate mechanisms that allow real-time adjustments based on situational factors.

While the findings suggest a shift in attribute prioritization across scenarios, several alternative explanations should be considered. First, participants may have interpreted "critical" situations as requiring immediate action, leading them to focus on attributes that emphasize direct harm minimization rather than broader ethical considerations. In addition, cognitive biases—such as loss aversion—may have influenced participants' responses, making them more sensitive to potential harm in critical scenarios. Another possibility is that the way scenarios were framed influenced attribute

prioritization; for instance, emphasizing uncertainty in decision-making may have led to a greater preference for conservative risk-averse choices. Future studies could further investigate these possibilities by varying the framing of scenarios or incorporating real-time decision-making simulations.

### 4.2 Broader societal implications

The ethical decision-making framework proposed in this study not only provides a structured approach to AV behaviour but also has significant implications for public perception, regulatory strategies, and trust in AV technology. As AVs become more integrated into everyday transportation, their decision-making processes will be scrutinized by both the public and policymakers, influencing their acceptance and adoption.

Public trust in AVs is closely tied to perceptions of their ethical behaviour. If AVs are seen as making unpredictable or morally questionable decisions, public skepticism may hinder widespread adoption. Research in AI ethics suggests that transparent and explainable ethical frameworks can help bridge this trust gap (Hagendorff 2020). By clearly defining how AVs prioritize ethical attributes in different scenarios, our framework could serve as a foundation for improving public understanding of AV decision-making. However, a key challenge remains: even if AVs follow a rational and ethically justifiable framework, individual users may still disagree with specific decisions due to personal moral intuitions and cultural differences. This highlights the need for public engagement and education to align societal expectations with AV ethics.

Misinformation and public misconceptions about AV safety and ethical behaviour could further complicate adoption. Media portrayals of AV incidents often focus on rare but dramatic failures, reinforcing public fears about AV reliability (Cave et al. 2019). In addition, misunderstandings about how AVs make ethical decisions may lead to unrealistic expectations—such as assuming AVs can perfectly emulate human moral reasoning. Addressing these misconceptions through transparent communication, public outreach, and collaboration between AV developers and policymakers will be crucial in shaping a well-informed public discourse.

Ethical decision-making in AVs is not just a technical challenge but a regulatory one. Policymakers must determine how to balance safety, fairness, and liability when defining AV behaviour. Some jurisdictions, such as the EU, have already begun discussing ethical guidelines for AVs, emphasizing principles like transparency and accountability (European Commission et al. 2019). Our framework can contribute to these discussions by offering a structured method for identifying and prioritizing ethical attributes.

By situating our framework within these broader societal discussions, we highlight its relevance beyond technical implementation, emphasizing the need for interdisciplinary collaboration between AI researchers, policymakers, and the public to ensure responsible AV deployment.

### 4.3 Limitations and future directions

This study has some limitations that should be acknowledged. First, while the use of a Mann–Whitney U-test ensured robustness in assessing statistical differences, the sample size and demographic diversity could impact the generalizability of the results. Further research should include a larger and more diverse participant pool to account for cultural and contextual differences in ethical decision-making. In addition, while AU provides a structured way to analyze moral attributes, there remains the challenge of translating these findings into practical AV decision-making algorithms. Future work should explore how these attributes can be operationalized in real-world driving scenarios, considering both ethical principles and technical feasibility.

Another limitation of this experiment is that it is based on an online questionnaire and is therefore not immersive. AU is a framework based not only on psychology and philosophy but also on neurosciences and aims at taking into account the emotional response produced by the AV's decision-making. Several studies have shown that participants' subjective, behavioural, and psychological responses in VR environments map their behaviour and experience in real-world settings (Rovira et al. 2009; Skulmowski et al. 2014). The natural next step to this experiment will therefore be to reproduce it in a VR setting.

Given these findings, future research should aim to refine the methodology by incorporating dynamic testing environments, such as interactive simulations or virtual reality experiments, to observe real-time ethical decision-making. Finally, integrating AU into policy discussions on AV ethics could help bridge the gap between theoretical models and practical implementation in automated systems.

By systematically addressing these factors, future studies can strengthen the foundation for ethical AV decision-making frameworks, ensuring that they are not only theoretically robust but also practically viable and socially acceptable.

## 5 Conclusion

With this study, we present a scientifically grounded method for defining societally aligned attributes for AV decision-making. Building on AU and biomedical ethics, we use a scenario-based experiment as a feedback loop to align AV decision-making with societal values.

Our results show that ethical decision-making in AVs depends on dynamic prioritization of attributes, with harm minimization and fairness emerging as key factors, while others, like autonomy, are more context-dependent. The statistical analysis confirms significant differences in attribute prioritization between critical and non-critical scenarios.

Beyond methodology, this study highlights the broader impact of ethical frameworks on public trust, regulation, and AV adoption. Future work should expand scenario diversity, integrate stakeholder perspectives, and explore how AVs communicate their ethical reasoning to ensure responsible deployment.

**Author contributions** All authors made substantial contributions to the conception and design of the work. C.G. drafted the work, realized the data acquisition, analysis and interpretation. All authors reviewed the manuscript and revised it critically for important intellectual content.

**Funding** This work was supported by Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek.

**Data availability** Experimental data is available here: https://doi.org/https://doi.org/10.24416/UU01-89YGU3.

## Declarations

**Conflict of interest** The authors declare no competing interests.



## References

Aliman N-M, Kester L (2019) Requisite variety in ethical utility functions for AI value alignment

Aliman NM, Kester L, Werkhoven P, Yampolskiy R (2019) Orthogonality-based disentanglement of responsibilities for ethical intelligent systems. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 11654 LNAI. https://doi.org/10.1007/978-3-030-27005-6_3

Aliman N-M, Kester L (2022) Crafting a flexible heuristic moral metamodel for meaningful AI control in pluralistic societies. In: Wernaart B (ed) Moral design and technology. Wageningen Academic, pp 63–80

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, Rahwan I (2018) The moral machine experiment. Nature 563(7729):59–64. https://doi.org/10.1038/s41586-018-0637-6

Beauchamp TL, Childress JF (2019) Principles of biomedical ethics, 8th edn. Oxford University Press, Oxford

Bergmann LT, Schlicht L, Meixner C, König P, Pipa G, Boshammer S, Stephan A (2018) Autonomous vehicles require socio-political acceptance—an empirical and philosophical perspective on the problem of moral decision making. Front Behav Neurosci. https://doi.org/10.3389/fnbeh.2018.00031

Bonnefon JF, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science. https://doi.org/10.1126/science.aaf2654

Cave S, Coughlan K, Dihal K (2019) "Scary robots" examining public responses to AI. In: AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp 331–337

Cecchini D, Dubljević V (2025) Moral complexity in traffic: advancing the ADC model for automated driving systems. Sci Eng Ethics 31(1):5. https://doi.org/10.1007/S11948-025-00528-1/FIGURES/2

Cecchini D, Brantley S, Dubljević V (2023) Moral judgment in realistic traffic scenarios: moving beyond the trolley paradigm for ethics of autonomous vehicles. AI Soc. https://doi.org/10.1007/s00146-023-01813-y

Coin A, Dubljević V (2021) Carebots for eldercare: technology, ethics, and implications. Trust Hum Robot Interact. https://doi.org/10.1016/B978-0-12-819472-0.00024-1

de Melo CM, Marsella S, Gratch J (2021) Risk of injury in moral dilemmas with autonomous vehicles. Front Robot A I:7. https://doi.org/10.3389/frobt.2020.572529

Diederich A, Winkelhage J, Wirsik N (2011) Age as a criterion for setting priorities in health care? A survey of the german public view. PLoS ONE. https://doi.org/10.1371/journal.pone.0023930

Emanuel EJ, Upshur R, Thome B, Parker M, Glickman A, Zhang C, Boyle C, Smith M, Phillips JP (2020) fair allocation of scarce medical resources in the time of Covid-19. N Engl J Med. https://doi.org/10.1056/NEJMsb2005114

European Commission, Directorate-General for Communications Networks, & Content and Technology (2019) Ethics guidelines for trustworthy AI. https://doi.org/10.2759/346720

Faulhaber AK, Dittmer A, Blind F, Wächter MA, Timm S, Sütfeld LR, Stephan A, Pipa G, König P (2019) Human decisions in moral dilemmas are largely described by utilitarianism: virtual car driving study provides guidelines for autonomous driving vehicles. Sci Eng Ethics. https://doi.org/10.1007/s11948-018-0020-x

Givens J (2013) Primum non nocere (first do no harm): can the principles of medical ethics be applied to finance? Ethics in Finance, Robin Cosgrove Prize Global Edition 2012–2013.

Gros C, Kester L, Martens M, Werkhoven P (2024) Addressing ethical challenges in automated vehicles: bridging the gap with hybrid AI and augmented utilitarianism. AI Ethics. https://doi.org/10.1007/s43681-024-00592-6

Hagendorff T (2020) The ethics of ai ethics: an evaluation of guidelines. Mind Mach 30(1):99–120. https://doi.org/10.1007/S11023-020-09517-8/TABLES/1

Hain R, Saad T (2016) Foundations of practical ethics. Medicine 44(10):578–582. https://doi.org/10.1016/J.MPMED.2016.07.008

Johansson-Stenman O, Martinsson P (2008) Are some lives more valuable? An ethical preferences approach. J Health Econ. https://doi.org/10.1016/j.jhealeco.2007.10.001

Kallioinen N, Pershina M, Zeiser J, Nosrat Nezami F, Pipa G, Stephan A, König P (2019) Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives. Front Psychol. https://doi.org/10.3389/fpsyg.2019.02415

Kauppinen A (2020) who should bear the risk when self-driving vehicles crash? J Appl Philos. https://doi.org/10.1111/japp.12490

Li S, Zhang J, Li P, Wang Y, Wang Q (2019) Influencing factors of driving decision-making under the moral dilemma. IEEE Access

7:104132–104142. https://doi.org/10.1109/ACCESS.2019.2932043
McManus RM, Rutchick AM (2019) Autonomous vehicles and the attribution of moral responsibility. Social Psychol Person Sci. https://doi.org/10.1177/1948550618755875
On-Road Automated Driving (ORAD) Committee (2021) J3016_202104: taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. SAE International, London
Pflanzer M, Traylor Z, Lyons JB, Dubljević V, Nam CS (2022) Ethics in human–AI teaming: principles and perspectives. AI Ethics 3(3):917–935. https://doi.org/10.1007/S43681-022-00214-Z
Regulation (EU) 2019/2144 of the European Parliament and of the Council. (2019). https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/vademecum_2018.pdf
Rovira A, Swapp D, Spanlang B, Slater M (2009) The use of virtual reality in the study of people's responses to violent incidents. Front Behav Neurosci. https://doi.org/10.3389/neuro.08.059.2009
Rus M, Groselj U (2021) Ethics of vaccination in childhood-a framework based on the four principles of biomedical ethics. Vaccines 9(2):1–16. https://doi.org/10.3390/vaccines9020113
Schein C, Gray K (2018) The Theory Of Dyadic Morality: Reinventing Moral Judgment By Redefining Harm. Person Social Psychol Rev. https://doi.org/10.1177/1088868317698288
Shin D (2024) Artificial misinformation. Artif Misinf. https://doi.org/10.1007/978-3-031-52569-8
Shin D, Koerber A, Lim JS (2024) Impact of misinformation from generative AI on user information processing: how people understand misinformation from generative AI. New Media Soc. https://doi.org/10.1177/14614448241234040
Skulmowski A, Bunge A, Kaspar K, Pipa G (2014) Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study. Front Behav Neurosci. https://doi.org/10.3389/fnbeh.2014.00426
Sütfeld LR, 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 pressure. Front Behav Neurosci. https://doi.org/10.3389/fnbeh.2017.00122
Sütfeld LR, Ehinger BV, König P, Pipa G (2019) How does the method change what we measure? Comparing virtual reality and text-based surveys for the assessment of moral decisions in traffic dilemmas. PLoS ONE. https://doi.org/10.1371/journal.pone.0223108
Wilson H, Theodorou A (2019) Slam the brakes: perceptions of moral decisions in driving dilemmas. In: CEUR workshop proceedings, 2419.