UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

DEPARTMENT OF GENERAL PSYCHOLOGY

Ph.D. COURSE IN: PSYCHOLOGICAL SCIENCES

SERIES XXXV

THESIS TITLE

MORAL JUDGMENT AND AUTONOMOUS VEHICLES: A COGNITIVE, EMOTIONAL, AND SOCIAL INVESTIGATION

Coordinator: Prof. GIOVANNI GALFANO
Supervisor: Prof. ANDREA SPOTO
Co-Supervisor: Prof. SIMONE CUTINI

Ph.D. student: GIOVANNI BRUNO

TABLE OF CONTENTS

Chapter 4

Study 2: Framing self-sacrifice in the investigation of moral judgment and moral emotions in human- and autonomous-driving dilemmas

Chapter 5

Study 3: The role of time constraints and prosocial orientation in the AV dilemma

Chapter 6

Study 4: The role of information availability and perspective-taking: moral judgment behind the Veil of Ignorance

Chapter 7

Appendix

Introduction

In the last decade, the growing interest in the development of autonomous vehicles (AVs) has had a crucial influence in revamping the investigation of moral judgments. The interest in this technology escalated quickly in the vision of the upcoming revolution in transportation systems and in the driving activity itself (Othman, 2022). Indeed, the indubitable and gradual implementation of the autonomous driving mode will proportionally reduce the pivotal role of human beings as active actors of the traffic environment, revolutionizing the figure of the driver into a passive passenger of a self-driving vehicle. This transformation will lead to a radical transformation of the perception and the interpretation of road-related events, as the individual will no longer be in charge of the vehicle's maneuvers, devolving the driving decisions to the artificial intelligence system. Most of the time, the delegation process will be advantageous and proceeds smoothly, relieving a series of driving and attentional tasks, and allowing to rest or perform non-driving-related activities in the car (Pfleging and Schmidt, 2015). At times, however, AVs will have to face complex traffic situations, subjecting passengers to the mere acceptance of the vehicle's decision, which might not correspond to their will or to their hypothetical embraced behaviors if at the wheel of the vehicle in those specific moments. In the most critical scenarios, the AV may have to opt for a driving maneuver that may compromise its passengers' protection, in order to not undermine the safety of other characters in the traffic environment (e.g., pedestrians, cyclists). When it comes to this, the emerged conflict in the evaluation of the most appropriate driving behaviors typically lies in the moral and ethical framework, delving into the long-standing discussion between contrasting moral doctrines (e.g., Bentham, 1781/1996; Kant, 1785). Traditionally, moral judgments towards AV behavior facing complex traffic scenario is investigated through the help of the sacrificial trolley problem (Foot, 1978). In this application, searching for the best possible collective outcome is coherent with the endorsement of the utilitarian moral code. Nonetheless, is not uncommon that supporting this interpretation can lead to the acceptance of a self-detrimental act, as the other side of the utilitarian coin. Bonnefon and colleagues (2016) highlighted for the first time the so-called "social dilemma of self-driving cars", adapting the scenario to an AV facing a trolley-like moral problem. From that moment, a new research line on the morality of autonomous transportation arose, providing a productive and wider overview on individual and cultural characteristics of moral judgments in this field (Awad, 2018a), and deepening the moral perception of this groundbreaking technology in manifold experimental applications (e.g., Huang et al., 2019; Kallionen et al., 2019; Meder et al., 2019). Despite the great interest in the topic, the investigation of morality toward AVs is still in its early stages, and a series of experimental questions still stands on the application of this context of investigation to the traditional structure of moral dilemmas. Considering the impact that applied moral investigations in AI (Artificial Intelligence) ethics have on the social discussion in preparation to the automation revolution, the present dissertation has the main aim to continue exploring the characteristics of moral judgment towards AVs, focusing on the use of the text-based moral dilemma as the main research tool in this field.

Structure of the thesis

The dissertation is organized as follows. Chapter 1 describes the autonomous transportation technology and the concerning ethical/social discussion about its pros and cons. Chapter 2 introduces the theoretical background of morality as well as features and characterizations of moral dilemmas. Additionally, the state of the art of the experimental relationship between AVs and the trolley problem is here presented, throughout the experimental applications developed in literature. Subsequently, the research plan will be presented, outlining and discussing the development and the results obtained from four experimental studies. In a first section, the first two studies aimed to experimentally reconsider a series of untested assumption in the investigation of autonomous driving moral behavior through the tool of the moral dilemma. Study One (Chapter 3) tested the applicability of the traditional trolley problem structure in the context of non-autonomous driving scenarios, in comparison with a validated set of sacrificial dilemmas (Lotto et al. 2014). Thereafter, human-driving and autonomous-driving scenarios were compared in Study Two (Chapter 4) in terms of moral judgment and the elicited moral emotions, further stressing the self-sacrifice framing as a textural difference between driving and non-driving moral dilemmas. In a second section, the AV dilemma has been specifically investigated. A number of features concerning the dilemma structure and the moral decision-making process were considered, aiming for shed new light on how the structure of the text-based AV dilemma may shape moral judgment. Study Three (Chapter 5) tested the relationship between time and cooperativeness in the context of moral judgment (e.g., Goeschl and Lohse, 2018; Rand et al., 2012; 2014), deepening the role of time availability and prosocial orientation in shaping the agreement towards the AV utilitarian moral code. Finally, Study Four (Chapter 6) focused on how the quantity of information provided - and the perspective assumed by the moral agent in the AV dilemma (e.g., AV passenger Vs. pedestrian) – may influence the interpretation of the scenario, affecting the endorsement of a particular moral behavior. In conclusion, Chapter 7 offers a general discussion of the relevant results obtained throughout the whole project, highlighting and elaborating on the new insights arisen on the application of autonomous transportation dilemmas to the trolley problem structure, and on the overall moral and social perception of this new technology. Datasets and R scripts for each of the four studies are retrievable in the Open Science Framework (OSF) project folder (https://bit.ly/3cksq6Q); experimental stimuli and additional tables are reported in the Appendix section, and a glossary of the most employed terms and labels is provided in the correspondent section.

The theoretical framework

Chapter 1

1.1 The autonomous driving revolution

Since the beginning of the transportation systems revolution, the quest for automation has always been the North Star of the many companies in the mobility and automotive industry. A series of futuristic attempts have been conducted in the last century (e.g., The Milwaukee Sentinel, 1926), testing several innovative solutions to improve the autonomy of driving operations (see Anderson et al., 2014; Stanchev and Geske, 2015). The first example of self-operative vehicle was developed in 1980 at Bundeswehr University Munich, Germany, where Ernst Dickmanns' team successfully altered a Mercedes-Benz van to perform autonomous driving on an existing highway. The performance of the van was then optimized in the upcoming years, allowing for the recognition of certain obstacles and autonomous lane changes (see Davidson and Spinoulas, 2015; Weber, 2014). Dickmanns' results were groundbreaking, boosting the development of a series of projects on autonomous driving with the beginning of the new century. One of the most important milestones was posed by the U.S. Defense Advanced Research Projects Agency (DARPA), that held the annual DARPA Grand Challenges between 2003 and 2007. In these occasions, research teams working on autonomous driving systems competed for an economic prize in the completion of selected road courses. If no vehicle was capable to compete the 240km off-road race in 2004 (Weber, 2014), five out of twenty-three competitors made the finish line in 2005 (see Buehler et al., 2007), as well as six out of eleven teams in the 2007 DARPA new Urban Challenge (see Buehler et al., 2009). These events spurred the research and development of the majority of the companies for the improvement of vehicle autonomy, both from the automotive establishment (e.g., Mercedes-Benz, General Motors, Volkswagen Group, Ford Motor) and from new manufacturers (e.g., Google, Waymo, Uber, Tesla). To date, the most important companies invest billions of dollars in their Research and

Development sector, venturing to launch full autonomous vehicles in the mass market by 2025 (e.g., Korosec, 2022; Olinga, 2022). Despite claims and promises, the advent of the upcoming autonomous driving era still has to deal with a series of technical, economical, legal and societal roadblocks.

1.1.1 The autonomous vehicle

Autonomous vehicles (AVs) are defined as vehicles in which *"operation occurs without direct driver input to control the steering, acceleration, and braking and are designed so that the driver is not expected to constantly monitor the roadway while operating in self-driving mode*" (NHTSA, 2013). This technology is also known as 'self-driving car' or 'driverless car'. The passage from traditional human-driving vehicles to autonomous is a gradual transformation, relatively to the level of involvement requested to the human being during the driving activity. A worldwide recognized classification of the autonomous transition has been defined by the Society of Automotive Engineers (SAE) and the US National Highway Traffic Safety Administration (NHTSA), which differs only in terms of labels. SAE (2021, Figure 1) distinguished six levels of automation depending on the on-board driver assistance system (Davidson and Spinoulas, 2015; Martínez-Díaz and Soriguera, 2018). This classification moves between No-automation (Level 0), Assistive driving systems (Level 1 and 2) and Automated driving systems (Level 3, 4 and 5). Giving this description, is important to emphasize that, since today, no vehicles sold in the international automotive market is implemented with a Level 4 or Level 5 automated driving system. To date, Germany has become the first country in the world to create a legal basis for the use of Level 3 and – in the future - Level 4 driving aids, allowing Mercedes-Benz' Level 3 Drive Pilot system to be actually implemented on purchasable vehicles (MacKenzie, 2022; Mercedes-Benz, 2022; Pingol, 2021).

- **Level 0 (No-automation)**: the driver (i.e., the human-being) is fully responsible for all the driving activities. The system (i.e., the vehicle) has a supportive role, providing momentary driving assistance or emergency safety interventions (e.g., collision warning or automatic emergency breaking).

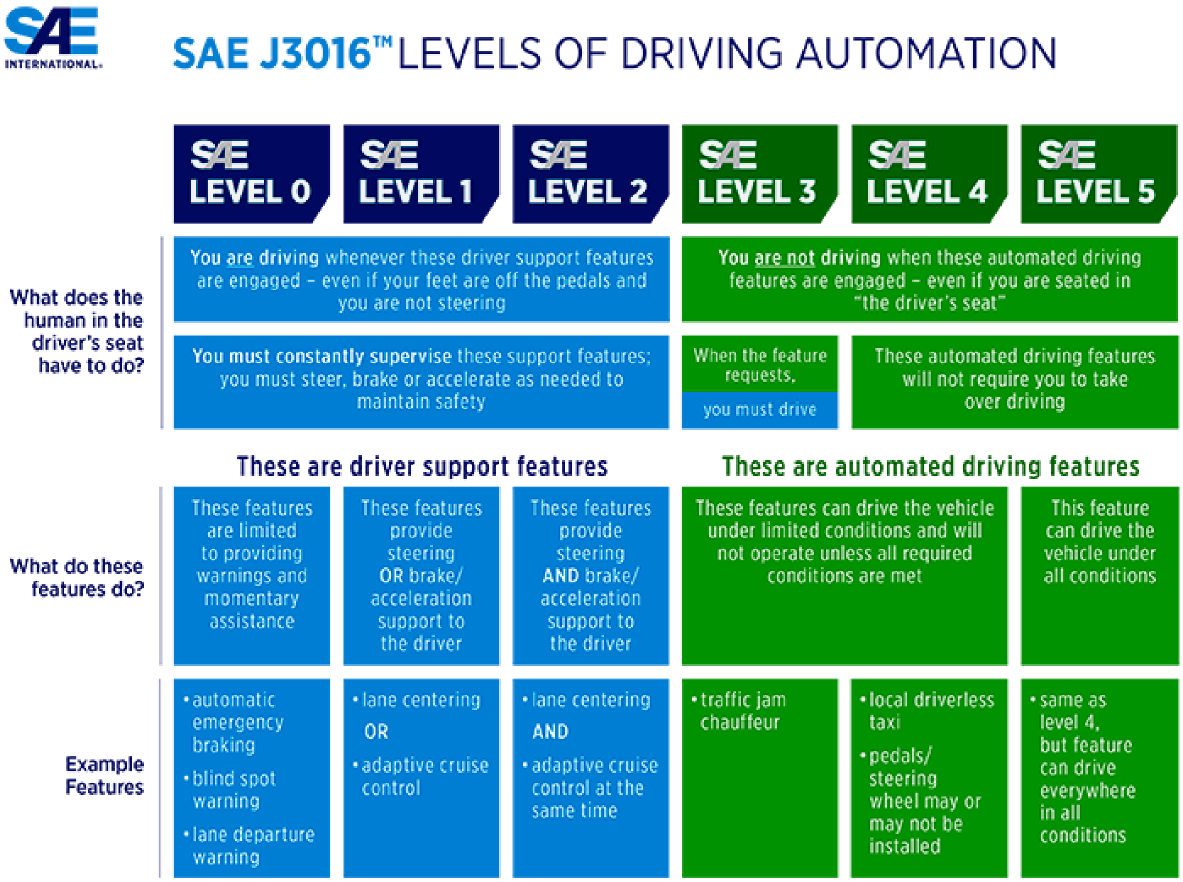
- **Level 1 (Driver Assistance):** the driver is fully responsible for all the driving activities. The system is capable to provide continuous assistance with either acceleration/breaking OR steering (e.g., adaptive cruise control), but the driver must be prepared to take control at any time.

- **Level 2 (Partial Driving Automation):** the driver is fully responsible for all the driving activities. The system is capable to provide continuous assistance with both acceleration/breaking AND steering (e.g., advanced driving assistance systems, ADAS), but the driver is required to supervise the technology and take over at any time.

- **Level 3 (Conditional Driving Automation):** when engaged, the system can perform all aspects of driving activities under limited circumstances. Under this level, the human-being/driver can be engaged in other non-driving related activities. However, he/she has to be present, alert and able to take control of the vehicle when needed (e.g., system failures) and when requested by the system.

- **Level 4 (High Automation):** when engaged, the system can perform all aspects of driving activities within limited-service areas (e.g., geographic boundaries) and certain conditions (e.g., limited performance under severe weather conditions). Under this level, the human-being is not needed to maneuver the vehicle in any circumstance (the vehicle may not have steering wheel and pedals), as the system is programmed to handle potential failures independently from any human support.

- **Level 5 (Full Driving Automation):** when engaged, the system is fully responsible for all the driving activities, in all roadways and under all conditions. Under this level, the human-being act only as passenger and is not needed to maneuver the vehicle in any circumstances (the vehicle does not have steering wheel and pedals), as the system it is programmed to handle potential failures independently from any human support.



SAE INTERNATIONAL

SAE J3016™ LEVELS OF DRIVING AUTOMATION

	SAE LEVEL 0	SAE LEVEL 1	SAE LEVEL 2	SAE LEVEL 3	SAE LEVEL 4	SAE LEVEL 5
What does the human in the driver's seat have to do?	**You are driving** whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering			**You are not driving** when these automated driving features are engaged – even if you are seated in "the driver's seat"		
	You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety			When the feature requests, you must drive	These automated driving features will not require you to take over driving	
	These are driver support features			These are automated driving features		
What do these features do?	These features are limited to providing warnings and momentary assistance	These features provide steering OR brake/ acceleration support to the driver	These features provide steering AND brake/ acceleration support to the driver	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met		This feature can drive the vehicle under all conditions
Example Features	• automatic emergency braking • blind spot warning • lane departure warning	• lane centering OR • adaptive cruise control	• lane centering AND • adaptive cruise control at the same time	• traffic jam chauffeur	• local driverless taxi • pedals/ steering wheel may or may not be installed	• same as level 4, but feature can drive everywhere in all conditions

Figure 1: The six levels of automation described by the Society of Automotive Engineers (SAE).

1.1.2 Potential benefits

Safety

The implementation of autonomous transportation has the potential to improve road safety by significantly reducing crashes. In the last decade (2011-2021), Italy observed a reduction in the number of road deaths (-26.3%), consistently with the trend observed overall in the European Union in the reduction of road deaths (EU = -31-3%), but assuming a higher starting point in terms of annual casualties (European Transport Safety Council, 2022). Indeed, in 2019 (before the COVID-19 pandemic), 172.183 road accidents occurred on the Italian roadways, resulting in 3.173 casualties and 17.600 injuries (ISTAT, 2020), the second highest absolute number after France (3.239). On the total number of road accidents, in the 90.3% of the times the fault is to be attributed to drivers' misconducts, and among the most frequent misconduct we find distraction, failure to observe precedence rules and high speed (overall the 38.2% of cases). Overall and in the same year, the percentage of driver errors as primary factor of road crashes in US is consistent with the Italian data, but with a higher number of people killed in motor vehicle traffic crashes (36.096 in 2019; National Highway Traffic Safety Administration, 2008, 2020). In 2019, 37.6% of fatal crashes in USA involved different combinations of distraction, fatigue/drowsiness, and alcohol-impaired driving. Assuming that, it is easy to imagine how the reduced human intervention during autonomous driving will increase road safety, lowering the number of traffic accidents (Koopman and Wagner; 2017; Litman, 2022). The reliability of autonomous systems in safely facing nearly every situation is still challenging but daily tested by manufacturers (e.g., Campbell et al., 2010; Singh and Saini, 2021). Several studies indicated safety as the most important factor affecting the adoption of AVs (e.g., Bansal et al., 2016; Kyriakidis et al., 2015; Schoettle and Sivak, 2014), and the fatal crash of Tesla's autopilot revealed some uncertainty on trust towards autonomous machines (Banks et al., 2018).

Efficiency

AVs may be able to provide not only more safety but also improved efficiency. Indeed, autonomous transportation should be of great benefit in urbanized and high-traffic areas, particularly exposed to traffic congestion and, consequently, fuel consumption (Igliński and Babiak, 2017). Additionally, AVs would be strictly law-abiding, driving in accordance with speed limits and minimizing the needs to abrupt and futile changes in velocity and brakes, which are known to be intensive fuel-demanding processes. For example, adaptive cruise control (ACC) and vehicle-to-vehicle (V2V) systems could lighten traffic flows, minimizing acceleration and braking (Atiyeh, 2012) and functionally increasing highway capacities (Shladover et al., 2012). Nonetheless, these impacts are still uncertain (e.g., Litman, 2022; Rodier, 2018). Some benefits require driverless car "platooning" (i.e., *flocking*, Heaslip et al., 2020), as a method for decreasing the distances between vehicles, increasing the capacity of roads, and minimizing aerodynamic drag through automated highway systems (Kesseris et al., 2007; Zabat et al., 1995). Platooning and eco-driving (i.e., *hypermiling,* the optimization of driving skills; Barth and Boriboonsomsin, 2009) can give the most significant support to the reduction of greenhouse gas (GHG) by 35%, but this result is strongly dependent from the level of AVs traffic penetration rate (Massar et al., 2021). The higher the AVs penetration rate, the stronger the reduction of GHG emissions in the long-term (Kopelias et al., 2020; Liu et al., 2019). Overall, the potential of AVs to remarkably tackle fuel pollution through the optimization of traffic management will depend on how AVs – and related infrastructures – will be designed and their market penetration regulated (Guériau et al., 2016).

Mobility

In order to be equitable, new transportation policies should favor different individuals and categories with respect to different abilities and needs (Litman, 2022). The revolution that AVs will bring to the whole transportation system will also have the advantage to provide easier access to improved mobility for the elderly, disabled, or those too young to drive. This is possible mainly in the last extensions of autonomous transportation, with no option of taking over the driving activity, allowing for independent personal mobility for categories with special needs (e.g., Anderson et al., 2014; Fagnant and Kockelman, 2015). This new feature would make AVs more attractive, especially in situations of highly-urbanized areas and parking-space shortages, where the introduction of shared AVs (SAVs) has the ability to replace privately owned vehicles on a 1:10 ratio (Fagnant and Kockelman, 2018). The potential advantage of shared AVs is pivotal since an increase in private mobility can lead to a proportional boost in Vehicle-Miles Traveled (VMT), leading to higher traffic congestion and emissions. As previously observed, the effective impact of this downside is strongly dependent on a careful and context-based implementation of traffic-management strategies (Anderson et al., 2014; Atiyeh, 2012; Litman, 2013). Interestingly, despite the elders being generally considered early adopters of autonomous transportation (Harper et al., 2016), previous studies show contrasting results regarding their positive (Sun et al., 2020; Williams et al., 2020) or pessimistic perception of AVs (e.g., Abraham et al., 2017; Piao et al., 2016).

Finally, driving activity has been recognized as a source of stress and multiple psychophysiological condition, which are critical factors in car crashes likelihood (Beanland et al., 2013; Useche et al, 2017; Vivoli et al., 2006). AVs may have the potential to decrease mental workload and driving-related stress, but mostly in highly automated vehicles (Level 4 or 5; De Winter et al., 2014), since regular disengagement from driving (e.g., Merat et al., 2014) and the involvement on secondary tasks (Eriksson and Stanton, 2017; Radlmayr et al., 2014) have a detrimental role on takeover time and driving transition.

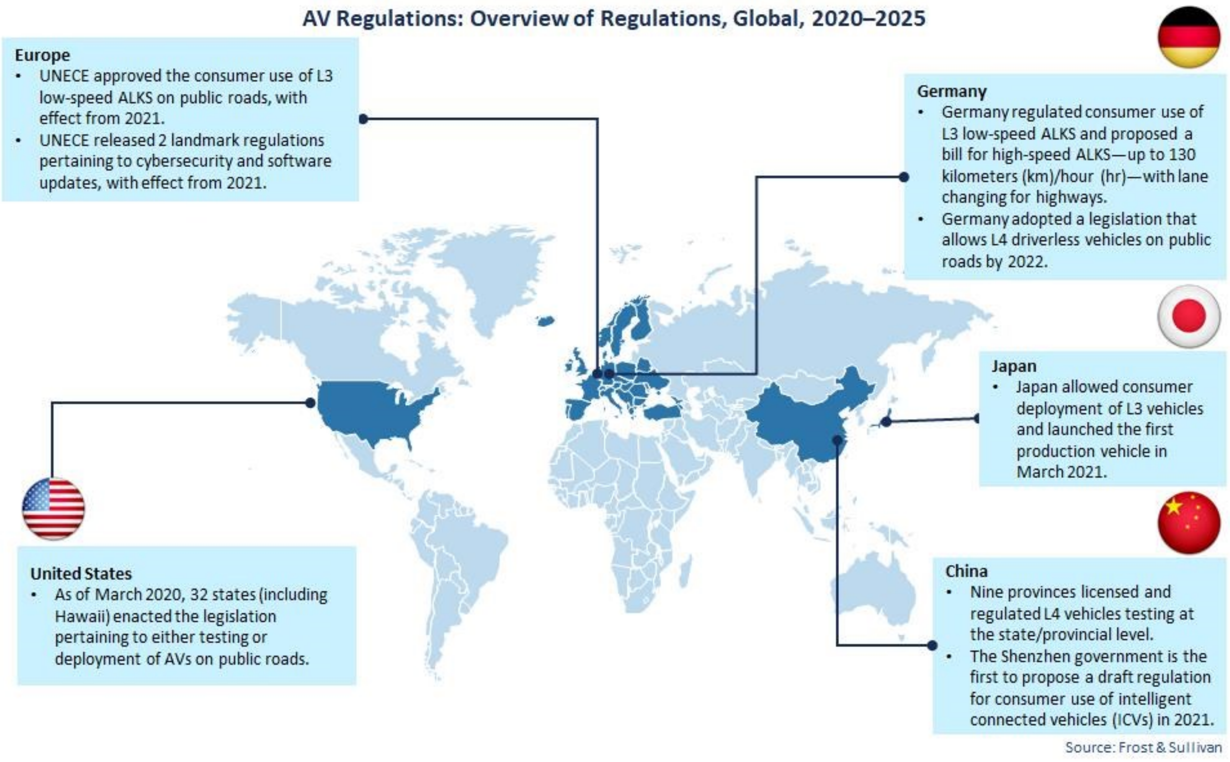
1.1.3 Barriers to implementation

Costs

When it comes to the critical aspects of autonomous revolution, the economic impact appears one of the trickiest obstacles (Fagnant and Kockelman, 2015). For the self-driving cars to be attracting to the large-scale market, costs will need to be tolerable so to allow for a massive distribution. Nonetheless, aiming for safety, these vehicles will have to include a series of fundamental technologies (e.g., Light Detection and Ranging, LIDAR) that are extremely expensive (Shchetko, 2014), leading to the definition of a market price outside the boundaries of the major part of its potential buyers. (Dellenback, 2013). Additionally, the availability of functional traffic management strategies (e.g., platooning) will request for a crucial reconsideration of roadways and transport planning (Litman, 2022; Martínez-Díaz and Soriguera, 2018), operations that will be extremely expensive for local and national administrations.

Regulation, legislation, liability

The introduction of AVs in the international mass market - and subsequently in the roadways system - has to deal with significant policy changes and a fundamental supportive legislation. Since today, a lot has been achieved but a series of open questions still stand. Assuming the United States of America as example, The AV regulatory landscape is rapidly evolving, even if resulting in different guidelines and testing certifications per state (Fagnant and Kockelman, 2015). By 2020, 32 states out of 50 countersigned a legislation for AV testing on public road, especially in California and Arizona (Kumar, 2021). Further steps have been done up to Level 4 of automation also in other countries, mainly in Germany, Japan and China (see Figure 2). In 2013, Smith underlined how actual legislation in the USA may have a detrimental effect on the introduction of AVs. Ten years later, through the New Car Assessment Program (NCAP), the NHTSA is aiming to integrate AVs into the existing road safety standards, since the current occupant protection regulation is written on the basis of traditional non-autonomous mode of transportation. The NHTSA suggests reinforcing driver-assistance technologies, strengthen testing procedures, and considering the use of emerging on-vehicle technologies monitoring driver performances (Uhlemann, 2022).



AV Regulations: Overview of Regulations, Global, 2020–2025

Europe
- UNECE approved the consumer use of L3 low-speed ALKS on public roads, with effect from 2021.
- UNECE released 2 landmark regulations pertaining to cybersecurity and software updates, with effect from 2021.

Germany
- Germany regulated consumer use of L3 low-speed ALKS and proposed a bill for high-speed ALKS—up to 130 kilometers (km)/hour (hr)—with lane changing for highways.
- Germany adopted a legislation that allows L4 driverless vehicles on public roads by 2022.

Japan
- Japan allowed consumer deployment of L3 vehicles and launched the first production vehicle in March 2021.

United States
- As of March 2020, 32 states (including Hawaii) enacted the legislation pertaining to either testing or deployment of AVs on public roads.

China
- Nine provinces licensed and regulated L4 vehicles testing at the state/provincial level.
- The Shenzhen government is the first to propose a draft regulation for consumer use of intelligent connected vehicles (ICVs) in 2021.

Source: Frost & Sullivan

Figure 2: Overview of the present AV regulation on the main international countries. Retrieved by: https://www.frost.com/frost-perspectives/standardized-regulatory-framework-and-rapid-technological-advancement-set-to-propel-autonomous-vehicles-globally/

Additionally, the transition to fully AVs will completely ignore the human-being during the driving activities, also bringing to a radical shift in the attribution of blame and responsibility in the cases of car crashes (Geistfeld, 2017). At this time, we are witnesses of a gradual increase in the implementation of autonomous features, and soon we will be part of a roadway landscape in which autonomous and non-autonomous vehicle will coexist facing unsolved technical and regulatory issues (e.g., Awad et al., 2018b; Munster, 2017). This factor has to be considered both from the legal and the social perspective, since research suggests that AVs may be blamed more than human drivers in the evaluation of the same driving decision (Dietvorst et al., 2015; Malle et al., 2015). Persons travelling in AVs will no longer be in control of the vehicle (Collingwood, 2017), so the responsibility of critical events will shift onto the vehicle itself and the companies involved in its production (Marchant and Lindor, 2012). Since no unique legal framework still exists outlining how liability is shared between AV's owner and third parties (US government delegates most of the responsibility in determining liability to state governments; NHTSA, 2017), the problem still stands.

AV acceptance and general attitudes

In this layered and complex framework of potential pros and cons related to the adoption of autonomous transportation, the subjective perception and the general attitudes towards AVs have a fundamental role on the actual spread of this technology (Fagnant and Kockelman, 2015; Gkartzonikas and Gkritza, 2019). Some people still struggle in accepting to delegate their decisions to machines (e.g., Hoff and Bashir, 2015; Kaur and Ramperstad, 2018; Zhang et al., 2019), and the endorsement of autonomous transportation by its potential stakeholders (e.g., passengers, pedestrians, cyclists) must be considered assuming the potential role played by socio-demographic and individual characteristics, as age, gender, education, socio-economic position, or previous experience with AVs (Martínez-Díaz and Soriguera, 2018; Othman, 2022). Despite the expected utility of autonomous transportation for the elder population (Harper et al., 2016), studies focused on this topic showed an inverse relation between age and positive attitudes towards AVs adoption, more appealing for youngers than elders (Abraham et al., 2017; Hulse et al., 2018; Piao et al., 2016; Richardson and Davies, 2018; Schoettle and Sivak, 2014). In terms of gender, overall males appears to be more optimistic than women about AVs (Schoettle and Sivak, 2014), less concerned about fully automation (Schoettle and Sivak, 2015), more prone to use it in the future (Piao et al., 2016), and more confident about letting them take full control of the driving activities (Abraham et al., 2017). Kahan et al. (2007) observed that female reluctance does not arise from biological or social reasons, but instead from perceived threats sand cultural identities. Nonetheless, Panagiotopoulos and Dimitrakopoulous (2018) suggested that the observed gender gap towards AVs acceptance is lessening in time. Previous experience was observed to positively influence public acceptance of AVs, both in behavioral (Piao et al., 2016) and simulative studies (Wintersberger et al., 2016), as well as socio-economic background. In this sense, AVs acceptance has been detected as inversely related with income, suggesting how developed and high-income countries have a greater awareness of AV technology but are more pessimistic about their level of safety (Moody et al., 2020), privacy and data security (Bazilinskyy et al., 2015) when compared to developing countries with great road safety challenges. Interestingly, Haboucha et al. (2017) demonstrated that people underestimate the absolute market price of AVs in favor of an advantageous price difference between AVs and regular human-driving vehicles, reconsidering the appealing of AVs in the mobility market relatively to other purchasable options. Finally, AVs acceptance seems also shaped by attitudinal variables, such as environmental concern and pleasure to drive (Haboucha et al., 2017), or the assumed perspective when imagining the potential interaction with the technology (Hulse et al., 2018).

Chapter 2

2.1 Moral Judgment

The interest in the field of moral psychology has grown exponentially during the last 15 years, attesting to a flourishing era of this research topic (Malle, 2021; Waldmann et al., 2012). In this wide framework, a considerable number of studies investigated moral judgment, as the core concept of moral cognition (Figure 3). Currently, no consensus has been reached on the actual definition of the topic, suggesting the need to distinguish different kinds of moral judgment through a taxonomical approach (Sinnot-Armstrong, 2016). Bertram Malle (2021) produced a comprehensive review of this distinction declining four main classes of moral judgment on a hierarchical basis (evaluations, norm judgments, moral wrongness, and blame judgments), which offers a potential satisfying definition. Firstly, (i) 'evaluations' represent the most basic and fast human reaction (e.g., Leuthold et al., 2015), splitting moral stimuli into good and bad, or positive and negative judgments. At this level, the morally relevant information is not disclosed, since it needs time to be integrated into the moral process (Guglielmo, 2015). Despite being traditionally considered an affective reaction, evaluative moral priming can occur earlier than markers of emotional arousal (Cusimano et al., 2017; Gui et al., 2016). At a second level, Malle defined (b) 'norm judgments' as the moral instructions that justify intentional actions that still have to take place. Here, the main judgment probes are framed on permissions ("Is it morally permissible to…"), moral appropriateness ("Is it appropriate to…"), and prohibitions, which have been observed as non-equivalent in daily moral judgment (Holleman, 1999). Norm judgments are expressed to claim the moral standards that allow evaluations to be interpreted (Nichols and Mallon, 2006), and are traditionally investigated as a categorical evaluation through moral dilemmas (e.g., Greene et al., 2001, Moore et al., 2008; see section 1.2.3). Subsequently, (c) 'moral wrongness' judgment has the role to determine if a norm-violating behavior was performed without sufficient justification (Cameron et al., 2017; Giner-

Sorolla et al., 2018; Riordan et al., 1983). It flags intentional violations in a large subset of moral psychology studies, independently from who performs the action, oppositely to 'evaluations' ("Is it morally wrong to…"). Finally, (d) 'blame judgments' are produced when a performed act is evaluated as universally and contextually wrong, violating a clear moral norm (Alicke, 2000; Malle et al., 2014). At this point, the moral agent has processed all the potential sources of information, like agent's reasons and potential justification, allowing for clear-cut criticism (Cushman, 2008, Malle et al., 2014).

In the present dissertation, we will focus on moral judgment as categorical norm judgments (b) in the form of moral permissibility, through the adaptation of traditional moral dilemmas in a specific context of investigation.

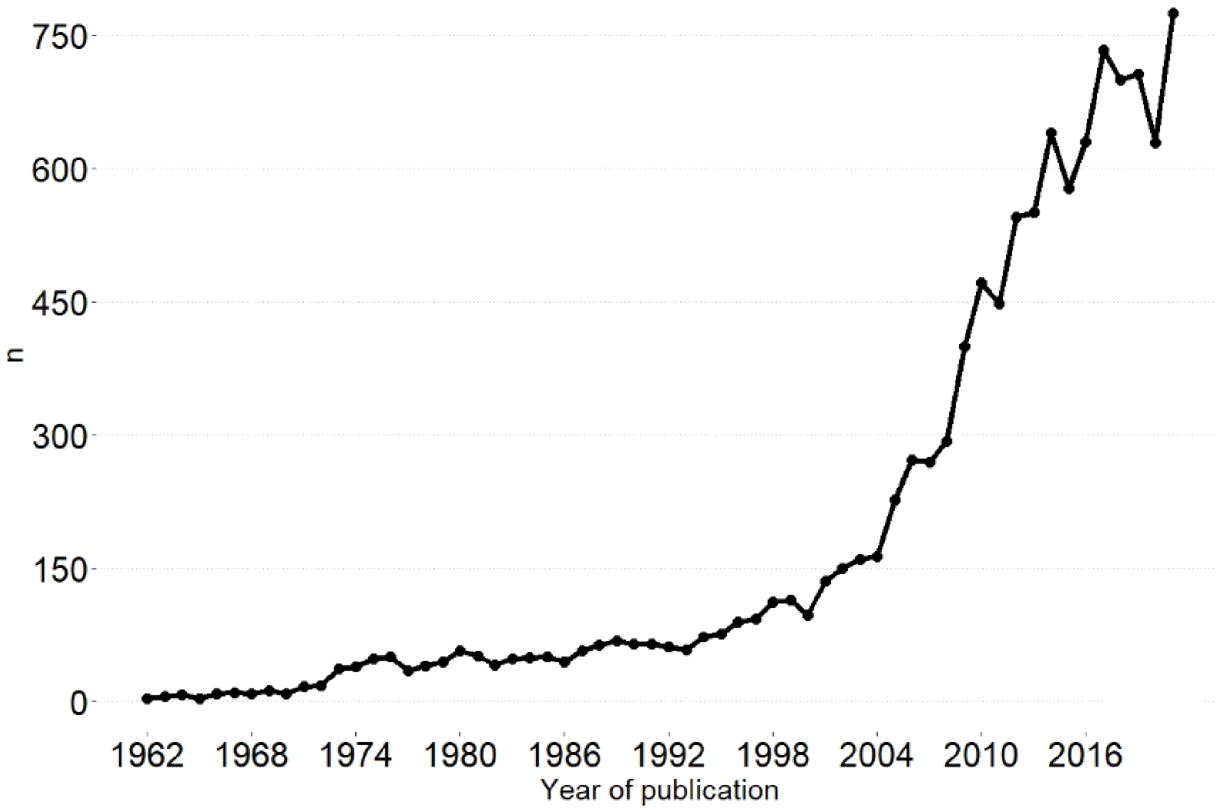


Figure 3: Number of 'Moral Judgment' keywords in published peer-reviewed articles, from 1962 to 2021 (retrieved by Scopus).

1.2.1 An overview of the main theories

The critical discussion on moral judgment fed on the genesis, the development, and the characterization of moral norms (see Machery and Mallon, 2010; Waldmann et al., 2012). Kohlberg (1981; Crain, 1985) assumed moral reasoning as the central component of moral development, neglecting the innate nature of this process. Over the years, his Rationalist Theory increased its questionability, as alleged to be simplistic (Cushman et al., 2006), biased toward men's moral reasoning (Gilligan, 1982), and Western-oriented (Simpson, 1974). A more emotional perspective of moral judgment was then given by Haidt's Social Intuitionist Model (2001), that distinguished moral reasoning, a conscious and stadial activity, from moral intuition, as an automatic and unconscious process with affective valence. After that the initial intuition has taken shape, reasoning has the role to provide subsequent rationalization to the judgment, also embedding the influence of social interactions. The main criticisms of Haidt's theory were related to the missing explanation on what mediates the relationship between intuition and reasoning (Harman et al., 2010), as well as on the primary role of intuition over reasoning (e.g., Paxton and Greene, 2010; Pizarro and Bloom, 2003). An important milestone was then laid by Haidt and Joseph (2004; 2007), in the direction of a moral domain-specific adaptation of innate dispositions to norms. Through the Moral Foundation Theory (MFT), they defined four and subsequently five moral domains (Harm/Care, Fairness/Reciprocity, Ingroup/Loyalty, Authority/Respect, and then Purity/Sanctity), which are innate and triggered by particular stimuli in response to specific adaptive challenges. Morality is then shaped and specified across cultures but maintains recurrent themes. The extension of the Relational Model Theory (Rai and Fiske, 2011) contrasted MFT, focusing on the role of relevant social relationships in shaping moral perspectives, on the basis of a more social-relational approach to moral cognition. (Waldmann et al., 2012). McHugh et al. (2022) agreed on the responsive nature of moral development, but shifted the focus from domain-specific to the acquisition of domain-general skills in training moral categorization (Moral Judgment As Categorization, MJAC). In this perspective, people are trained to learn the association of particular circumstances to moral wrongness/rightness, automatizing the process of such categorization in future relevant occasions (Barsalou, 2003).

Among the most important theories, the Dual Process Theory (DPT) by Greene and colleagues has surely been extremely influential and attractive from neuroscientific, psychological, and philosophical perspectives (Greene et al., 2001; 2004; 2008). DPT underlines the systematic competition between cognitive and emotional processes in the development of moral judgment. The authors claim that the nature of the moral problem (personal Vs. impersonal, see section 1.2.3) is capable to activate two separate systems (Greene et al., 2004): personal violations trigger socio-emotional and affect-based moral judgments, while impersonal violations may result in more controlled and cognitive-based decisions. These two mental processes act in a distinctive way: the automatic-emotional process is fast and intuitive, operating mainly at an unconscious level and highly influenced by emotional activation. Oppositely, the conscious-controlled process is slow and deliberative, producing a more analytical interpretation of stimuli, based on moral norms and situation-specific features, and less on affective states. Greene's theory can be interpreted as a domain-specific interpretation of the more general distinction between automatic and conscious though processes, well-described by Kahneman (2011) in the contraposition between System 1 (fast and intuitive) and System 2 (slow and deliberative).

This interpretation growth alongside with neuroimaging evidence (Greene et al., 2014; Nejati et al., 2021; Pretus et al., 2019; Riva et al., 2019), resulting in specific brain areas responding to personal dilemmas and associated to the automatic-emotional process (e.g., the ventromedial prefrontal cortex, the posterior cingulate cortex and the amygdala), as well as areas more sensible to impersonal dilemmas, activated in case of cognitive-controlled responses (e.g., the dorsolateral prefrontal cortex). Evidence of this dissociation has also emerged from the behavioral perspective, since longer decisional times (DT) were detected when a stronger cognitive control was requested for overcoming prepotent emotional responses in the evaluation of personal moral evaluation (Greene et al., 2001; 2004). In this regard, emotional 'hot' moral judgments are observed to be faster than their controlled 'cold' counterpart. Further specification of the DPT will be deepened coherently with the description of moral dilemmas in section 1.2.3.

1.2.2 The role of emotions

We have observed that there are different opinions on how cognition and emotion interact in the development of moral judgment. Overall, most of the theories agree on a complementary relationship between the two systems, but the direction of this relationship still remains a point of discussion (e.g., Crockett, 2013; Greene, 2016; Haidt, 2003a; Figure 4). Several researchers endorse the emotions-as-consequence paradigm, assuming that specific moral emotions are triggered by the cognitive appraisal of particular moral transgressions (Huebner et al., 2009; Roseman, 1996). On this basis, emotions like anger, disgust, and contempt ('other-referred', see Haidt, 2003b) are elicited by behaviors violating different kinds of moral norms (Rozin et al., 1999), with different level of intentionality and threat (Hutcherson and Gross, 2011; Russell and Giner-Sorolla; 2011). Using a philosophical definition, this can be defined as a 'pure Kantian' model of moral judgment (see Huebner et al., 2009), where affective reactions are generated reactively to rational appraisal (Kohlberg, 1969; 1981). Oppositely, when rationalization of moral judgment take place as a consequence of a primary emotive activation, it can be defined as a 'pure Humean' model, in accordance with Haidt's Intuitionist Model (2001). Recent neuropsychological data highlighted the possibility of the contemporary integration of emotion and deliberation in the making moral judgment ('hybrid' approaches, Damasio, 1994; Greene 2001; Greene 2004), as also suggested by Helion and Ochsner (2018). The authors endorse a bidirectional model between the two systems, with emotional processes that motivate different kind of cognitions, than resulting in different perceived moral emotions. Finally, it is also worth mentioning a number of models that consider emotions and conscious reasoning as a consequence of moral judgment, which is driven by a distinctive, analytical and independent moral faculty. In these 'pure Rawslian' models, emotions and reasoning can affect the distinctive behavior, but not the moral judgment, which is the first result of an a-priori moral analysis (e.g., Hauser, 2006; 'Universal moral grammar' by Mikhail, 2007).

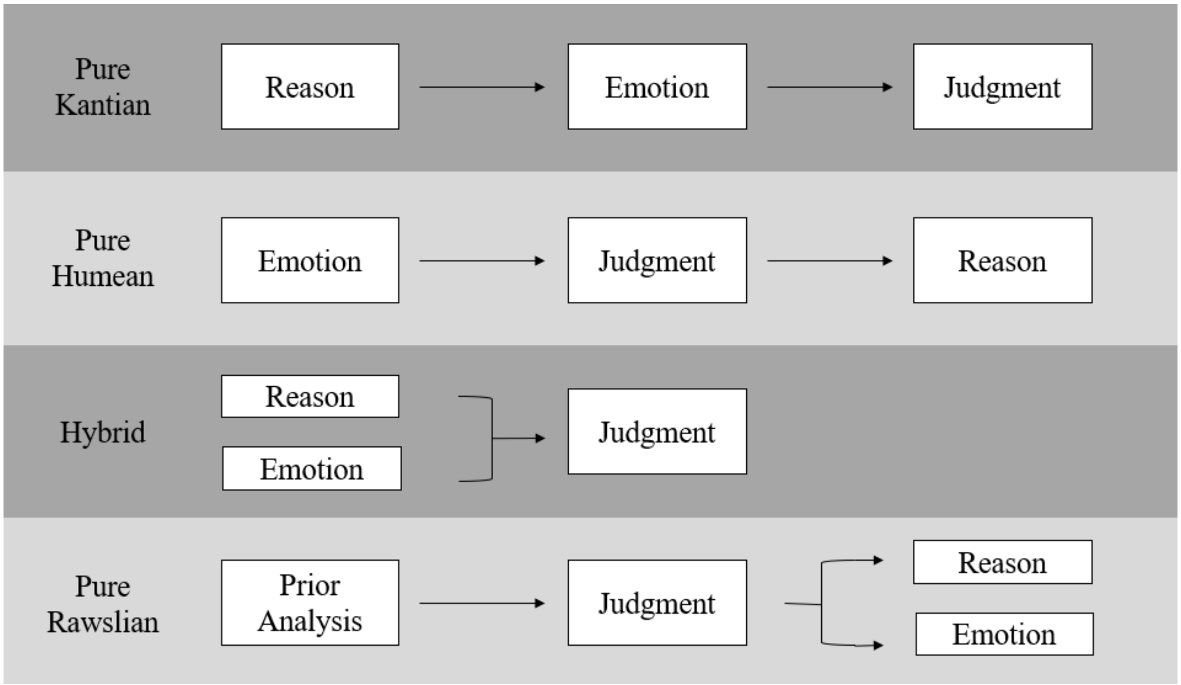


Figure 4: A schematic description of the four interpretations of moral judgment described above (inspired by Huebner et al., 2009)

The nineties have been the setting of a vivid debate on the categorization of moral emotion (e.g., Ekman, 1992; Ekman and Davidson, 1994; Russell 1991), and Haidt's moral family categorization resulted to obtain the highest credit among different interpretations. Haidt defined moral emotions as those emotions linked to the interest of collectivity or at least of one person other than the moral agent (2003b). They are typically interpreted on the basis of trigger events not directly referred to the self (e.g., transgressions, tragedies), and resulting action tendencies, because of the natural motivation – once the emotion is perceived – to execute a motivated response to the eliciting event. Haidt (2003b) endorsed the existence of two broad families, other-condemning (contempt, anger, and disgust) and self-conscious emotions (shame, embarrassment, and guilt), as well as two smaller categories, namely other-suffering (compassion) and other-praising emotions (gratitude and elevation). Knowing precursors and outcomes of the emotional experience is important to integrate the understanding of observed behavioral adherence to moral norms. The description of elicitors and consequent actions were carefully described by Haidt, and also reviewed by Tangney et al., (2007), with a focus on the distinction of particular couple of emotions (e.g., shame Vs. guilt) in several applied contexts. Finally, only in the last few years researches start providing normative values on the self-reported emotional experience during the process of moral judgment (e.g., Christensen et al., 2014; Lotto et al., 2014; Pastötter et al., 2013), despite valence and arousal are aptly considered to account for most of the variance in moral emotional judgment (e.g., Bradley and Lang, 1994).

1.2.3 The (sacrificial) moral dilemma

Currently, a widely used experimental tool for the investigation of moral judgment is the moral dilemma. In these situations, moral agents are required to choose between two (or more) conflicting courses of action, indirectly endorsing a particular moral code but with a foreseen faith: no matter what they select, a downside is expected (McConnell, 2002). In the sacrificial version of the dilemma (i.e., ‘sacrificial dilemmas’, Bartels and Pizarro, 2011), the main character has to face a conflict between a number of unsatisfying actions with outcomes involving the loss of at least one human life. Traditionally the option at disposal of the moral agent are two, but few examples of multiple-choice dilemmas exist in the literature (e.g., *the moral trilemma*; Thomson, 2008). To date, a large number of studies focus on sacrificial moral dilemmas (Bauman et al., 2014), and the reason for its popularity is justified by a series of advantages: a systematic exploration of a wide number of variables, the high internal consistency ensured by the material, and the possibility to elicit almost infinite moral conflicts in several contexts

(Christensen and Gomila, 2012). Despite the limited use of this tool in older studies (e.g., Kohlberg, 1964), moral dilemmas gained more attention contextually to the investigation of the neuropsychological correlates of moral judgment (Greene et al., 2001). The inspiration came from the philosophical exploration of ethics principles, where moral dilemmas are intended as ’thought experiment’, as intuition exercises for revealing potential inconsistency in individual moral norm. The paradigmatic case of sacrificial moral dilemma, considered as the new standard to test moral permissibility, is the trolley problem, where the utilitarian normative ethical theory is traditionally put to test by deontologism (Kahane, 2015). Utilitarianism - as a form of Consequentialism - claims that people should always aim to maximize overall welfare, denying that moral rightness can be evaluated by anything other than consequences (Bentham, 1781/1996; Mill, 1861/2004). The moral action is then the action that produces the greatest amount of good for the greatest number of people. Chasing utilitarianism means to assume an impartial and neutral approach: the only goal is to maximize overall benefit. Oppositely, deontologism is a duty-based normative theory, that assess morality on the basis of what we ought to do (deontic theory), more than what we should do. The right and moral behavior is the one that follows prescribed moral norms (i.e., Kant’s categorical imperatives) that is absolutely immoral to violate, despite the consequences of the derived actions (Kant, 1785). Applying the trolley problem paradigm, the utilitarian normative theory is traditionally tested against deontologism, or in general against a non-utilitarian approach, applying the theory to a concrete case-scenario (Di Nucci, 2013). The first version of the trolley problem (i.e., the Trolley or Switch dilemma) was offered by Philippa Foot (1967), claiming:

> *SWITCH: You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman. If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on*

your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman. Is it appropriate for you to hit the switch in order to avoid the deaths of the five workmen?

Here, the utilitarian norm requires the moral agent to pull the switch, causing the death of the single workmen on the secondary railway. This is the action that minimize the overall harm, assuming an 'economic' approach evaluating the harmful impact of each alternative. A renewed version of this scenario was composed by Thomson (1985) and investigated by Greene et al. (2001) in comparison with the original Switch problem (Figure 5). This second version is known as Footbridge or Push dilemma, and reconsidered the mean to fulfill the utilitarian goal:

PUSH: A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. You are on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to you on this footbridge is a stranger who happens to be very large. The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workmen will be saved. Is it appropriate for you to push the stranger on to the tracks in order to save the five workmen?

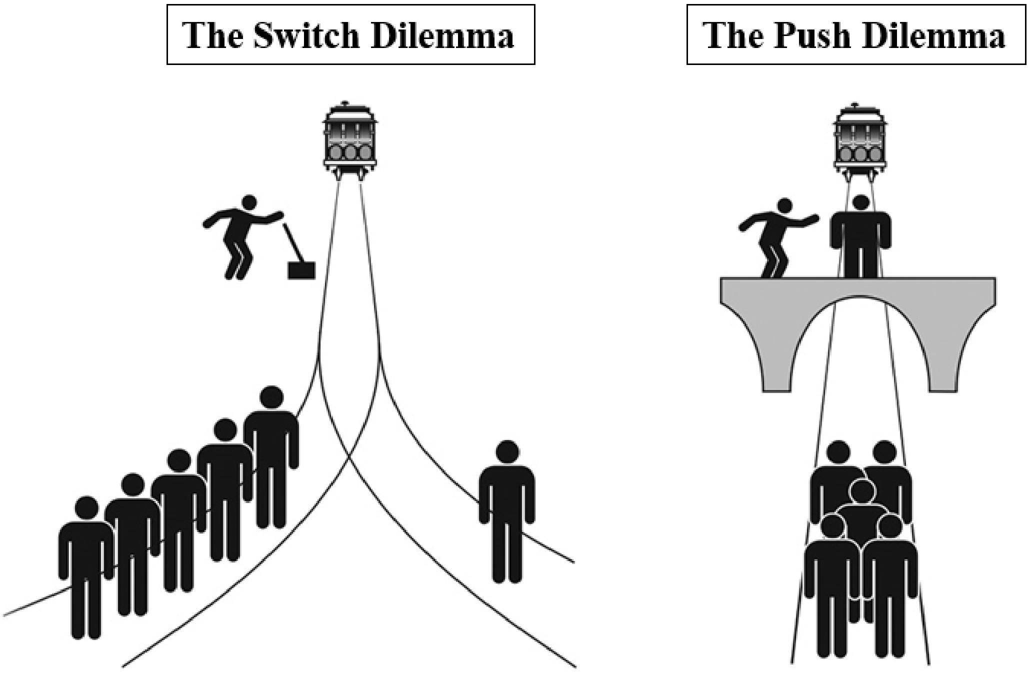


Figure 5: The Switch (on the left) and the Push (on the right) version of the trolley problem.

Despite no morally relevant differences being retrievable between the Switch and the Push dilemma, Greene et al., (2001) demonstrated that the thought of actively pushing a person off the bridge reduces the attractiveness of the utilitarian resolution (Cushman et al., 2006; Greene et al., 2009). Consistently, the clear trend in favor of the utilitarian approach observed in the Switch case was inverted in the Push dilemma, endorsing the possibility that the cause of this difference may lay on the higher emotional salience of the Push case (i.e., personal dilemma) when compared to the Switch counterpart (i.e., impersonal dilemma). Greene distinguished personal from impersonal dilemmas on the basis of the relationship between the *offender* and the *victim(s),* as the possibility to cause physical harm to a specific person or group of persons, not resulting in redirecting an existing threat onto a third party. Further insights about this contrast may be obtained by taking into consideration the role of intentionality in the process. The role of intention has been widely deepened in the investigation of moral judgment, as the conscientious act to realize a negative outcome (Sosa et al., 2021). A large amount of evidence confirms that people tend to judge more harshly intentional acts than accidental/unintended harmful behaviors (e.g., Borg et al., 2006; Cushman et al., 2006; Greene et al., 2001; Kleiman-Weiner et al., 2015; Legnado and Channon, 2008). The higher emotional activation and moral wrongness perceived in the Push dilemma can then be interpreted on the basis of the intentionality of the moral action (i.e., pushing the large stranger off the bridge), with the help of two similar but different moral doctrines inspired by Aquinas (1274/1952). The first interpretation is based on the Doctrine of Doing and Allowing (DDA, Quinn, 1989), which assumes that people are more sensitive to the consequences of action than of omission or inaction (Waldmann and Dieterich, 2007). In other words, committing harm takes more to justify than just allowing it (Woollard, 2012). Here, the moral difference stands on the active Vs. passive involvement of the moral agent, well exemplified by the different perceptions of active Vs. inactive euthanasia (Christensen and Gomila, 2012). The peculiarity of the Push dilemma lies in the direct contact that the moral agent needs to have with the victim, carrying out a proactive action to accomplish the utilitarian faith. According to the DDA, this occurrence is harder to rationalize since the focus is on the means instead of on the end. Nonetheless, currently the Doctrine of the Double Effect (DDE; Foot, 1967) has gained great credit in the interpretation of moral judgment (e.g., Hauser et al., 2007; Lotto et al., 2014; Zelazo et al., 1996). This theory assumes the moral inadmissibility of voluntarily causing harm, but allows for the possibility of causing unintentional harm, if and only if is intended as a foreseen side-effect of an action aimed to carry out the best possible outcome. On this basis, the sacrifice of the single worker is morally permissible as a foreseeable but undesired consequence of diverting the run of the runaway trolley, for the achievement of the best possible option (i.e., saving five workmen), while pushing the man off the bridge remains an immoral act because it intended to cause harm for reaching the same greater goal. The interpretation of this different moral judgment on the basis of intentionality and DDE leads to a new taxonomy of sacrificial moral dilemmas, slightly different than Greene et al., (2001): impersonal or Switch-like dilemmas become incidental, as personal or Push-like are renamed as instrumental.

The large amount of research that used sacrificial moral dilemmas as experimental tool for the investigation of moral judgment has allowed defining a series of personal and situational factors influencing the endorsement of the utilitarian approach (Bruers and Braeckman, 2014; Klenk, 2021). Among personal factors, evidence shows increased utilitarian judgments under cognitive and emotional impairment or alcohol dependence (Khemiri et al., 2012), as well as among males (e.g., Fumagalli et al., 2010) and in individuals with psychopathy traits (e.g., Patil, 2015). Among situational factors, instead, we can detect an influential role of cognitive control (Conway and Gawronski, 2013) and time constraints in expressing moral judgment (Suter and Hertwig, 2011; Tinghög et al., 2016; Rosas and Aguilar-Pardo, 2020). Additionally, and coherently with the DPT, more cognitive control (allowed by the use of a foreign language in the presentation of the dilemmas, see Corey et al., 2017; Muda et al., 2018) and less negative affect (e.g., Strohminger et al., 2011) are able to positively influence the endorsement of utilitarian resolutions to sacrificial dilemmas.

2.2 The AV dilemma

In the last few years, the consistent growth in the deployment of sacrificial moral dilemmas in several experimental applications has taken advantage of the renewed interest in moral perception and social acceptance of autonomous transportation. The application of this tool in the investigation of driving-related situations is undoubtedly straightforward, replacing the runaway trolley with an AV and of the two train tracks with two implementable driving maneuvers relevant to two opposite moral codes. Assuming the manipulation of several important features (e.g., number of characters involved, relationship with the characters involved, potential traffic rules violation), the AV version of the trolley problem depicts a prototypical traffic-set situation (Bonnefon et al., 2016). Here, an AV is driving *x* passenger(s) from point *a* to point *b,* when a number of *y* pedestrians ($x < y$) unexpectedly cross the road intersecting its travel path. If the AV continues straight, it will unavoidably run over the pedestrians, causing their death. The only way to avoid that is for the AV to steer suddenly, crashing against a heavy object (e.g., the guardrail) and sacrificing its own passenger(s). Morally, the passive non-steering solution is framed as nonutilitarian (i.e., deontological) and self-protective, while the resolutive steering maneuver is interpreted as utilitarian and self-sacrificial. The first application of the sacrificial dilemma in the context of autonomous driving has been implemented by Bonnefon, Shariff, and Rahwan (2016), which brought to light the so-called *dilemma of self-driving cars*. In all the six studies presented, the authors manipulated a series of relevant information about the moral context concerning the AV (e.g., personal involvement, number of characters involved, the role of the characters involved, relationship with the characters involved, the presence of forcing AV behaviors; Figure 6). Specifically, in one of these studies participants were asked to indicate the likelihood to buy an AV programmed to minimize casualties (i.e., utilitarian AV), or to prioritize its passenger before anyone else, even if the consequence would be killing a higher number of people. In their study, the AV dilemma was presented as:

> *AV DILEMMA: You and [a family member / a coworker] are in the car traveling at the speed limit down a main road on a bridge. Suddenly, 20 pedestrians appear ahead, in the direct path of the car. The car could be programmed to: swerve off to the side of road, where it will impact a barrier and plunge into the river, killing you and your [family member / coworker] but leaving the pedestrians unharmed; or stay on your current path, where it will kill the 20 pedestrians, but you and your [family member/ coworker] will be unharmed.*

Here, participant firstly expressed their moral preference on a 0-100 slider anchored to the two proposed courses of action (swerve or stay on track), and then rated their willingness to buy an AV programmed to follow the same actions. Results showed that even though participants were agreeing on the higher morality of the utilitarian and self-sacrificial options, they would prefer to buy a self-protective AV for themselves. This result has been eligibly interpreted as a social dilemma, since, if both self-protective and community-protective (i.e., utilitarian) AVs were available in the mass market, very few people would opt to share the utilitarian algorithm, even though this decision would be different from what they consider to be the moral choice. Bonnefon et al. (2016) also highlighted the positive relationship between the number of characters involved and the appeal of the utilitarian behavior, as well as the detrimental role of high-level regulation in the enforcement of utilitarian AVs.

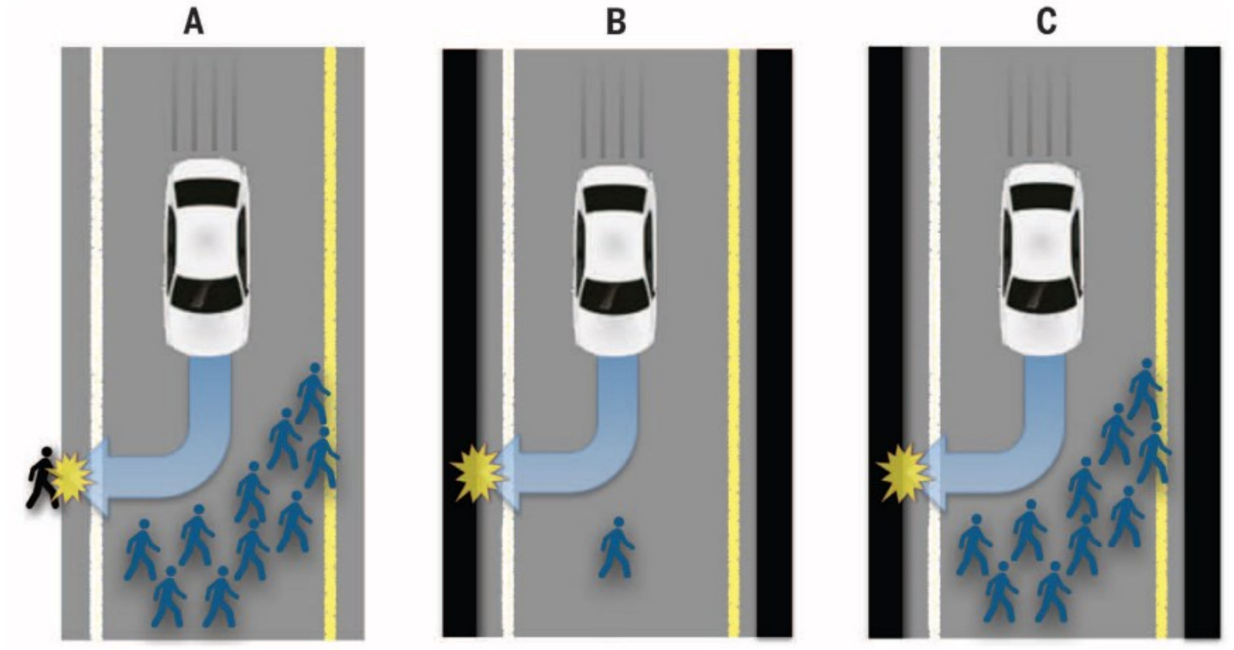


Figure 6: Three potential traffic situations of unavoidable harm involving an AV, used as a demonstrative example by Bonnefon et al., (2016). In each of these situations, the vehicle has to decide if (A) killing a high number of pedestrians crossing the road or a third blameless bystander, (B) killing a single pedestrian crossing the road or its own passenger, and (C) killing a high number of pedestrians crossing the road or its own passenger.

From this point on, several studies deepened the moral perception of this new technology through sacrificial dilemmas, focusing on several features that may act in shaping moral judgment towards AVs. Surely, the massive 'moral machine' project has collected the largest amount of evidence on this topic, documenting global preferences, cultural clusters (Awad et al. 2018a), and universal qualitative patterns of preferences towards AVs (Awad et al., 2020) through the presentation of a series of moral dilemmas. Awad et al. (2018a) recognized three main building blocks for machine ethics, as the global preferences for sparing humans (Vs. animals), sparing more lives and sparing young lives (Vs. old). The respect of traffic rules was also observed as important in this sense, a feature coherently judged as important also by Li et al. (2019). Consistently, they observed how cultural more than geographical proximity allows to define a number of moral clusters toward AVs, which are characterized by comparable care for a series of moral preferences evaluating this technology. This result was lately interpreted also in consideration of relational mobility, as the easiness with which people can create new social relationship (Thomson et al., 2018), a cultural-specific feature that appears to have a role on the likelihood to endorse sacrifice aiming for the greater good (Awad et al., 2020). Contextually, Zhao et al. (2019) deepened the complexity in design the perfect autonomous system, claiming for the higher equality that would arise from the selection of a random maneuver when facing the moral dilemma. Interpreted as a more moral deontological resolution, this study highlights the need for the development of safe AI systems and for a general consensus on the moral code that will lead AVs behavior in the future. Frank et al. (2019) confirmed the assumptions of DPT in shaping moral judgment, claiming that assuming intuitive and fast decisions, people tend to shift towards a more deontological approach. Importantly, it also suggests that the perspective assumed by the moral agent is not trivial in reaching a decision, since individuals tend to prefer self-protective behavior in vision of an imminent AV crash. This latter result was consistent with the evidence collected afterwards and from different approaches (Huang et al., 2019; Kallioinen et al., 2019; Martin et al., 2021a), confirming how perspective can influences moral judgment towards self-protective behaviors.

The utility of the trolley-like problem structure was not relegated only to the textual form but helped in investigating moral perception of AVs also in more immersive and ecological settings

(Navarrete et al., 2012; Skulmowski et al., 2014). Sütfeld et al., (2017) employed virtual reality (VR) to assess for the first-time moral behavior in a traffic scenario while on board of a simulated AV, speculating on the suitability of VR as a rightful assessment tool in this context. A number of additional studies follow (e.g., Faulhaber et al., 2019; Kallioinen et al., 2019; Li et al., 2019), agreeing on an overall propensity towards utilitarian resolutions. Besides the evident advantages, the widespread interest on trolley-like dilemmas in the definition of social expectations towards AVs resulted in a great discussion arising on the reliability of this experimental tool in this specific field of investigation. De Freitas et al. (2021, 2020) criticized the use of the driverless dilemma is just inappropriate in the investigation of multi-layered ethical challenges concerning autonomous transportation, and coherently Mukhopadhyay et al., (2020) claimed that trolley-like problems are too abstract to bring strong evidence on a real-life situation as the one depicted in the AV dilemma. Oppositely, Gill (2021) underlined the importance of ethical dilemmas considering the future diffusion of AV technology among its potential adopters, while Krügel and Uhl (2022) suggest that this technique may be highly informative to disentangle important open questions concerning autonomous transportation.

2.3 The present project

Concurrently with the daily optimization of autonomous transportation technology and the complex adaptation of societal and mobility habits – as well as the legal regulatory system -, the interest in the moral perception and social acceptance of driverless vehicles' behavior is consistently growing in the applied psychology research framework. We have had the opportunity to see in the sacrificial moral dilemma the most commonly implemented technique to this aim, as a flexible and adaptable experimental tool to deepen this distinctive framework. Several shreds of evidence have already been collected in the last few years (e.g., Bonnefon et al., 2016, Awad et al., 2018a; Gill, 2021), but a series of experimental queries remain open, both on the appropriateness of this tool in the field of traffic scenarios, and on how particular features on the dilemma may shape moral judgment, potentially jeopardizing the interpretation of results. In this context, the general objective of the present project is to continue deepening this research line, aiming to give a preliminary answer to some (un)asked applicative questions on the well-known AV dilemma. Specifically, the focus of the present project is on the textual version of sacrificial moral dilemmas, as the main experimental modality employed in the investigation of moral judgment (e.g., Cushman et al., 2006; Greene et al., 2001; Lotto et al., 2014; Moore 2008). Firstly, the reliability of the traditional sacrificial and incidental dilemma in the field of human-driving behavior has been tested in Study 1, validating the first dilemma set applied to the driving context. Then, the effect of autonomous and nonautonomous transportation on moral judgment has been compared in Study 2, controlling the scenarios for a fundamental structural difference (the self-sacrifice framing) that distinguishes traditional sacrificial dilemmas from its AV application. Once cleared potential differences in the moral evaluation of human- and autonomous-driving driving behaviors, we shifted our attention on a series of personal and situational factors discussed in the moral literature that may shape moral judgment towards AVs. In Study 3 we focused on the interaction between time availability and prosocial orientation (Suter and Hertwig, 2011; Tinghög et al., 2016; Rosas and Aguilar-Pardo, 2020), a well-debated topic in the field of social dynamics and cooperation (Goeschl and Lohse, 2018; Montealegre and Jimenez-Leal, 2019; Rand et al., 2014), to understand if individual prosocial orientation may have a role on the endorsement of the utilitarian moral code, with or without time constrains. Subsequently, the role of perspective-taking has been considered in shaping moral judgment towards AVs, comparing pedestrian Vs. passenger's perspectives. This investigation was computed manipulating the amount of available personal and contextual information through a new sequential behavioral paradigm for moral dilemmas. To this aim, Study 4 operationalizes the Raswlian Veil of Ignorance (VOI; Rawls, 1971/2009) in experimental form, benefiting from the trilemma version of the Switch trolley problem (Di Nucci, 2013; Thomson, 2008. For each study, specific introduction, hypothesis set, and discussion has been produced. We assume that the specific application of sacrificial dilemmas to the context of traffic events does not follow different decision-making processes and moral norms than traditional sacrificial dilemmas (e.g., Schein, 2020), believing that the evidence collected from the computed studies may then be aptly interpreted in the wider moral judgment framework for further applications. Nonetheless, the distinction between human- and autonomous-driving cannot be taken for granted, as a completely different experience that may affect the moral judgment towards the suitability of particular driving behaviors. Hopefully, the results obtained from this project will be useful to enrich the knowledge on the moral perception of AVs behavior and social expectancies toward this revolutionary technology, both on a normative and on a descriptive level.

Chapter 3

Study 1: the trolley problem in the field of human-driving scenarios

3.1 Rationale of the study

As highlighted in section 1.3, the occurred interest in the social acceptance and moral perception of autonomous transportation has boosted the employment and adaptation of sacrificial moral dilemmas in autonomous driving-related situations. Traditionally, the dilemma of Autonomous Vehicles (AVs) is build following the structure of an incidental dilemma on the basis of the DDE. This principle offers a taxonomy of moral dilemmas based on the intention that guides the moral agent. In incidental dilemmas, causing harm for pursuing a greater good is acceptable as a foreseen – but accidental – side effect. As we have seen in section 1.2, DDE interpret the Switch dilemma as incidental, assuming the sacrifice of the single worker as a predicted but unintended consequence. The decision to shape the AV dilemma as incidental becomes clearer when hypothesizing the correspondent dilemma in its instrumental form. Here, the sacrifice of one (or few) individuals is assumed as a conscious means to save a higher number of people, as the "large stranger" is the means to save five workers in the traditional Push dilemma. Apply the AV dilemma to the instrumental form appears a challenging task, as its structural requests collide with the development of a credible scenario. Indeed, no realistic driving circumstances request the driver (or the autonomous vehicle) to consciously sacrifice somebody to avoid the death of more people, while this is not strictly true when imaging an incidental version of a driving dilemma. Since today, the AV adaptation of the trolley problem has been the object of study for several experimental applications in the fields of moral psychology and applied ethics, also owing to the considerable flexibility of the tool. In this context, morality toward AVs have been deepened by adapting the trolley problem in virtual environments such as VR (e.g., Benvegnù et al., 2021; Kallionen et al., 2019; Sütfeld et al., 2017) and driving simulators (e.g., Frison et al., 2016; Samuel et al., 2020), but the main research method is still the textual form (e.g., Bonnefon et al., 2016; Sütfeld et al., 2019). The selection of the method is clearly related to the research needs, but Sütfeld et al., (2019) observed that the textual and VR moral dilemmas appear to measure similar constructs. Clearly, the text-based dilemma is easier to adapt to the driving context, considering the higher flexibility of the textual form when compared to a more complex virtual simulation. In all the observed applications, the dilemma stood in the comparison between an active utilitarian option (e.g., swerving off the road or breaking) and a passive non-utilitarian one, traditionally interpreted as deontological (e.g., driving straight). As described in section 1.3, results show good consistency between driving behavior-based studies, depicting a strong approval of the utilitarian resolution which also appears to be compatible with the proportion of moral preference observed in the general non-driving moral literature (Cushman et al. 2006; Greene et al., 2001; 2008; Lotto et al., 2014; Moore et al., 2008). Descriptively, this similarity may be consistent with a structure-based interpretation of moral dilemma in the development of moral judgment, which would reduce the importance of the specific context described in the depicted moral scenario (Schein, 2020). Nonetheless, this result is still not supported by experimental evidence, since until today no study compared a specific set of driving-based moral scenarios with the traditional non-driving dilemmas. In the context of autonomous transportation, Bonnefon et al., (2016) detected an incongruency between moral decision and moral evaluation in the context of autonomous transportation, since the application of utilitarian moral reasoning for dilemmas concerning others did not match the research for self-protection when the moral agent’s life was at stake in the dilemma (i.e., “the social dilemma of self-driving cars”). The self-involvement factor has been widely deepened in the literature, manipulating the self Vs. other benefit in sacrificial dilemmas and establishing the role of self-involvement in the endorsement of a particular moral code (Moore et al., 2008; Lotto et al., 2014). Additionally, the role of emotional processing in shaping moral judgment is well-known (Greene et al., 2001, 2004, 2008). The endorsement of the utilitarian resolution in moral dilemmas appears to be related to a lower emotional activation (arousal), since the activation of slow and deliberative processes have the potential to overwhelm fast and unconscious emotional response. This evidence has been confirmed by a limited number of studies on non-driving dilemmas which measured emotional activation through self-reported techniques, confirming a reduced arousal and no specific emotional valence in the endorsement of the utilitarian behavior (e.g., Lotto et al., 2014; Pletti et al., 2016; Sarlo et al., 2012).

To date, the interpretation of moral judgment towards AV's behavior has been obtained generalizing the evidence collected on sacrificial and incidental non-driving dilemmas (Cushman et al., 2006; Greene et al., 2001, 2008; Moore et al., 2008; Ugazio et al., 2012), or directly applying this experimental tool to the autonomous transportation framework (Awad et al., 2018a; Bonnefon et al., 2016; Graham et al., 2016). Nonetheless, relatively low attention has been paid to ensure the actual reliability and generalizability of sacrificial dilemmas to the context of driving behaviors before the introduction of any specification on the level of automation. A number of empirical data are today available in the literature on moral judgment, moral acceptability, emotional activation and decision times facing non-driving dilemmas, while no information has been collected in the specific field of human-driving behavior. With Study 1 we aimed to take a step back from the captivating application of moral dilemmas to autonomous driving, describing the applicability of this tool in the context of driving activity. For this reason, we built and tested the first set of sacrificial and incidental dilemmas in this specific context, comparing it with a validated set of sacrificial, incidental but non-driving moral scenarios in terms of (i) decision times, (ii) moral decision, (iii) moral evaluation, and (iv) emotional activation. Additionally, the self-involvement factor was manipulated, to detect potential differences in the endorsement of the utilitarian moral code when the life of the moral agent was at stake.

3.2 Hypothesis

A series of hypotheses were advanced in the present study, considering the thoughts described in section 1.2 and in the rationale of the present chapter:

- Consistently with a structure-based interpretation of moral dilemmas (Schein, 2020), no differences are expected between the human-driving and the non-driving dilemma sets for any of the considered dependent variables, namely: decision times, moral decision, moral evaluation, and emotional activation.

- Considering the evidence collected on the role of personal benefit in shaping moral judgment (e.g., Moore et al., 2008; Lotto et al., 2014; Sachdeva et al., 2015), a higher endorsement of self-protective outcomes is expected in dilemmas involving the moral agent as a potential victim of the scenario, as well as a higher and more unpleasant emotional activation.

- Coherently with the DPT (Greene et al., 2001), faster decision times, as well as a stronger (arousal) and negative (valence) emotional activation, are expected in the occurrence of nonutilitarian moral decision, independently from the dilemma set and consistently with previous findings (Lotto et al., 2014).

- Eventually, a dissociation between moral decision and moral evaluation is expected. A negative evaluation of the moral acceptance of the selected behavior (utilitarian or non-utilitarian) would be consistent with the moral incongruency observed by Bonnefon et al., (2016), despite the human-driving nature of the present set.

3.3 Method

3.3.1 Participants

An a-priori power analysis been computed on G-power statistical software (Faul and Erdfelder, 1992) before starting the data collection, assuming a medium effect size (Cohen's d = 0.25) and a correlation of 0.50 among repeated measures, with an alpha error probability of 0.05 and 0.90 power. The system suggested a minimum of 124 participants, and a total of 152 participants were recruited for the experiment. Females accounted for 49.34% of the final sample (75 females). Overall, the mean age was 25.7 (SD = 5.48, range = 18–57), and 69.08% of participants were enrolled in university courses (n = 105), with 42.76% matriculated in human sciences and cultural-related programs (e.g., psychology, sociology, philosophy; n = 65). Most participants (95.4%) had held driver licenses (n = 145), and almost the totality of them (99.3%, n = 151) drove a maximum of 15,000 km per year. Only 4.60% were involved in a car accident in the prior 12 months (n = 7). The study was administered through Qualtrics software (Qualtrics, Provo, UT) and approved by the local ethics committee (ID No.: 3514). Before participation, each participant gave formal written consent, which was voluntary and unremunerated.

3.3.2 Materials

Forty-two incidental moral dilemmas were adopted for the experiment, two nonsacrificial (i.e., filler scenarios) and forty sacrificial, equally balanced between human-driving and non-driving scenarios. The twenty traditional non-driving dilemmas, as well as the two nonsacrificial fillers, were retrieved from the validated set produced by Lotto et al., (2014, edited from Sarlo et al., 2012). This set provides a number of self-involvement (in which the moral agent could be one of the potential victims) and other-involvement incidental moral scenarios, applied to different settings and with different sacrificial ratios (i.e., lives saved Vs. lives sacrificed). In the present application, this material was reproduced according to its original version, but considering a constant 1:3 sacrifice ratio between saved and sacrificed characters, so to control for this potential factor (e.g., Bonnefon et al., 2016). Coherently, the remaining twenty stimuli were structured as sacrificial and incidental driving-type moral dilemmas. As mentioned in the rationale of the present study, we focused only on incidental driving dilemmas, since applying road-based scenarios to the instrumental dilemma structure seems extremely unrealistic. A low plausibility of moral dilemmas has been recognized as a detrimental factor in moral judgment (Bauman et al., 2014; Gold et al., 2014). Consistently with the validated set, the new driving dilemmas were composed of a hypothetical moral scenario, depicting the moral agent at the wheel of a traditional nonautonomous vehicle facing a critical traffic event which could have been approached following two possible maneuvers: proceed straight, resulting in the sacrifice of three individuals (*nonutilitarian* or *deontological* resolution), or suddenly steer off the road, admitting the sacrifice of a single person (*utilitarian* resolution). The new driving-type dilemma set was composed controlling for a number of potential confounders that could have affected the moral judgment, acting on the endorsement of a particular moral code. We took in consideration the following details: (a) no violation of road rules and general safety, to prevent the clear allocation of responsibilities, (b) no characterization of individuals involved in the dilemmas (e.g., gender, age, ethnicity, or relationship with the moral agent), (c) no leading language, to prevent the risk of *response bias* (Loftus and Palmer, 1974), (d) constant sacrifice ratio (1:3), (e) limitation to a unique type of moral dilemma (i.e., killing or letting die) and (f) check for number of letters and words in each scenario. The dilemma set employed in the present study is retrievable in Appendix, and further descriptive information on each scenario are available in the Appendix and, together with the supplementary material, at the following Open Science Framework (OSF) project link: https://bit.ly/3cksq6Q.

Since we aimed to investigate the effect of self-involvement in the endorsement of the utilitarian resolution, in ten scenarios per type (driving and non-driving) the moral agent was not involved as a potential casualty of the scenario (*other-involvement* dilemmas), while in *self-involvement* dilemmas the utilitarian resolution allows for the protection of the self and two other individuals, resulting in the sacrifice of a third-party individual. This is consistent with traditional validated self-involvement dilemmas (e.g., Greene et al., 2001; Lotto et al., 2014; Moore et al., 2008), in which the self-sacrificial resolution is always framed in the nonutilitarian option (i.e., *"I die, and many others die"*), allowing the moral agent to opt for a utilitarian and self-protective counterpart (e.g., the "Crying baby" dilemma, Greene et al., 2001). This point has been further stressed in Study 2 (Chapter 4). Table 1 present an example of each type of moral dilemma used in the study: traditional non-driving dilemmas with self-involvement (i.e., Traditional Self, TS), traditional non-driving dilemmas without self-involvement (i.e., Traditional Other, TO), human driving-type dilemmas with self-involvement (i.e., Driving-type Self, DS), and human driving-type dilemmas without self-involvement (i.e., Driving-type Other, DO). Additionally, also an example of non-sacrificial dilemma is reported (i.e., Filler Nonsacrificial, FN).

Table 1: Sample of traditional non-driving, driving-type and filler dilemma, controlled per Self Vs. Other risk involvement.

Dilemma	**Scenario**	**Resolutions**
Traditional Self-involvement (TS)	A large man is leading you and two other explorers out of an underground cave on the west coast of Scotland. Due to its size, the man gets stuck in the narrow opening right at the exit of the cave. In a short time, the high tide will overwhelm you and everyone will drown except the man, who has his head out of the cave.	**A.** Since the passage is blocked, you let the high tide overwhelm you and all three drowned. **B.** You use the explosives you brought with you to widen the cave opening. You know that the man stuck will die, but you and the other two explorers will be able to escape.
Traditional Other-involvement (TO)	You are a building worker who is maneuvering a crane on a building site. You have just started your day on the site, when you realize that the cable of the crane is about to break. Attached to the cable is an enormous steel beam which is directly above a crew of six who are working on the outside of a building in construction.	**A.** As there is no way to warn them, you let the cable break and the steel beam kills the six workers. **B**. You slightly move the crane arm to another area of the construction site. You know that two workers work there, who will be crushed by the steel beam and will die, but the six workers will be saved.
Driving-type Self-involvement (DS)	It is night, and you are driving your car. A violent storm has hit your city for a few hours; it is still raining, and the asphalt is slippery. You are approaching a traffic light when suddenly two cyclists cross the road right in front of you. On the right sidewalk, you see a pedestrian. You try to brake, but the tires slide on the asphalt.	**A**. You continue straight, running over the two cyclists, who will die. Your car will continue to slide on the asphalt, crashing against a nearby building, and you will die. **B**. You suddenly steer right. You know you will run over the pedestrian on the sidewalk, who will die, but your car will slow down in an open field and you, and the two cyclists will be unhurt.
Driving-type Other-involvement (DO)	You are driving your car on a panoramic road. A single cyclist is riding on the cycling path on your right, parallel to the roadway. As you drive along the road, you suddenly see three workers in the middle of the road removing a small obstacle. You are too close to them and do not have enough time to brake.	**A.** You continue straight, running over the three workers, who will die. **B.** You suddenly steer right. You know you will run over the single cyclist on the cycle path, who will die, but the three workers on the main road will be unhurt.
Filler Nonsacrificial (FN)	You were invited to a birthday by an acquaintance. You do not really want to go or spend a lot of money on the gift because of your superficial relationship with the person. You will find a branded sweater in excellent condition in a second-hand shop. When the package is opened, the birthday girl is embarrassed in front of such an important gift.	**A.** You smile satisfied and reassure the person by saying that an important occasion like a birthday deserved a very special gift. **B.** At the opening of the gift you immediately tell the person the truth about the sweater, saying that today's fashion trend is relaunching the use of branded second-hand clothes.

3.3.3 Experimental procedure

As well as the other experimental applications of the present project, data collection for Study 1 was conducted online. This decision has been taken considering the limited possibility to access laboratories during the COVID-19 pandemic, as well as to allow external participants to access the University of Padua structures. Several studies compared online and lab methods performing decisional tasks and experimental procedures, highlighting a series of disadvantages (e.g., small behavioral differences, less performance accuracy, higher dropout rate in multiple-session studies), but – overall - comparable and reliable results (Buso et al., 2021; Donderand et al., 2008) dependent upon a series of practical recommendations (e.g., attention checks, warnings, explicit instructions and increased transparency; Newman et al., 2021).

Study 1 was programmed and distributed via Qualtrics software. The program provided an anonymous link to the survey, which was then distributed via social networks and institutional communication channels following a snowball non probabilistic sampling technique (Goodman, 1961; Parker et al., 2019). The data collection was performed from July 14th to August 23rd, 2020. All the participants were required not to perform the survey through smartphones or tablets, but only using laptops, in order to avoid problems of data comparability between multiple devices (Krebs and Höhne, 2021). To test the survey structure and the intelligibility of the new driving dilemmas, a pilot study was conducted on 12 participants, presenting the whole 42-stimuli dilemma set as described in Figure 7. The pilot participants did not find any problem in the comprehension of the driving scenarios, but they converged on the idea that the experimental session was too time-consuming (more than 1 hour) and cognitively demanding. For this reason, the sample was divided into four “lists” (*n* = 38 per list) beforehand. Each list was composed of 18 dilemmas: eight traditional nondriving, eight driving-type, and two fillers. Considering the balance between self- and other-involvement moral dilemmas, each participant disposed of four dilemmas per type. For each category (TS, TO, DS, DO, see table 1), two dilemmas were "list-specific", thus presented only in one list, whereas the remaining two per category (including fillers) were common among the four lists (i.e., "anchors"). The anchors were selected considering the normative scores from Sarlo et al., (2012) and the pilot scores, and a preliminary analysis on moral decisions and decision times ensured no differences between the selected anchors per each category. This methodological approach allowed us to halve the completion time while ensuring adequate statistical power through a sufficient number of answers per each driving-type dilemma. A graphical representation of the dilemma set composition is retrievable at Figure 7.

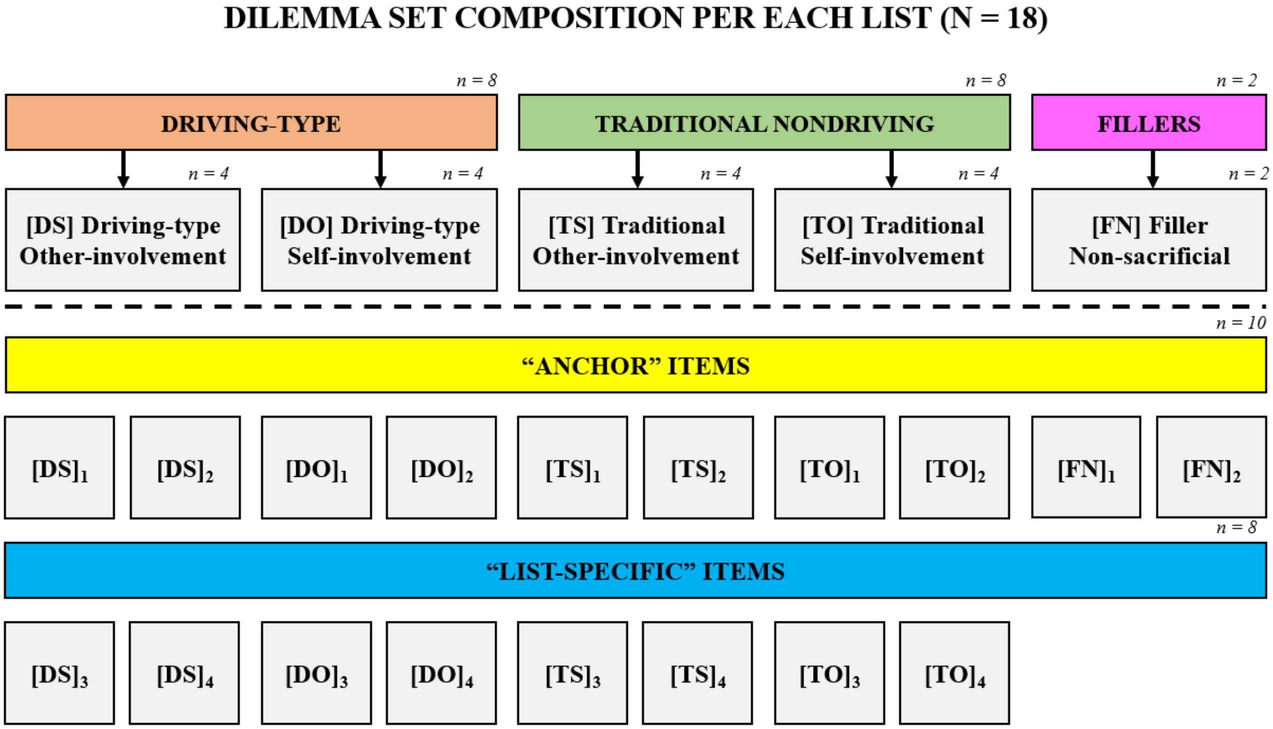


Figure 7: the composition of the 'list-specific' dilemma set.

The experimental procedure deployed in the present study was consistent with the validation study by Lotto et al. (2014), assuming the experimental material validated in this study as the reference point for the development of the driving-type dilemmas. A detailed explanation of the procedure is described in Palmiotti et al., (2020) and Sarlo et al. (2012). The mean completion of the experimental procedure was 26 mins (SD = 7.26 mins), and before the beginning of any experimental activity, the participants were requested to read and fill out an informed consent about their participation and data protection regulation. Subsequently, a series of sociodemographic information and driving habits were asked to the participant. Then, the experimental instructions were presented, describing the sequence of events that would characterize the presentation of each moral dilemma (see Figure 8). Each dilemma was first presented by its scenario as a text, in black type (font Arial, size 10) against white background. The participant had illimited time to read the scenario, before moving to the presentation of the two outcomes. Here, the nonutilitarian option (outcome A) was always presented first and was maintained the sole option on the screen for five seconds. Immediately after, the utilitarian option (outcome B) was added to the screen below the nonutilitarian one, which remained on the screen for further seven seconds. The presentation time of the two outcomes was proportionate to the length of two options, with the utilitarian one slightly longer then the nonutilitarian. A fixed presentation time was suggested to try to prevent participants from beginning the decision process while reading the two options (Sarlo et al., 2012). After this time, the option keys appeared on the screen, allowing the participant to indicate their preferred outcome correspondently to their moral decision. The decision time was recorded by Qualtrics as the time between the occurrence of the option keys and the selection of the preferred option.

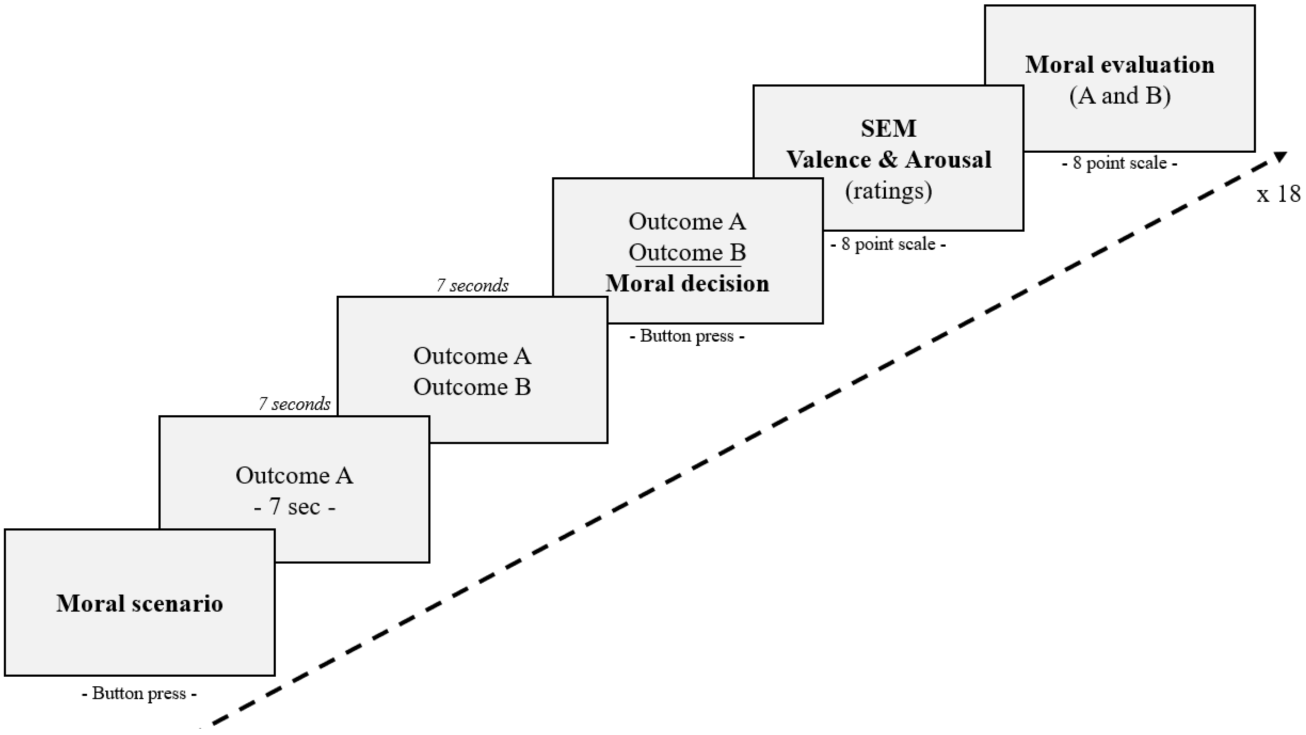


Figure 8: The experimental procedure of Study 1. The following sequence was repeated 18 times, one time per administered dilemma.

Once a decision was taken, the respondent was asked to rate their emotional state at the moment of the decision, so contemporary with the decision-making process. With this aim, the Self-Assessment Manikin (SAM) was administered (Bradley and Lang, 1994), as a non-verbal pictorial assessment technique directed towards the intensity (arousal) and the quality (valence) of the emotional activation. These two parameters were self-assessed with two 9-point graphic scales, for arousal (1 = calm, 8 = activation) and valence (1 = unpleasantness, 8 = pleasantness). Per each dilemma, the last requested activity was to evaluate the moral acceptability of the two proposed option, both the selected and the unselected one, on an 8-point scale (0 = completely unacceptable, 7 = completely acceptable). The described procedure was then repeated 18 times, until the completion of the experimental procedure.

3.3.4 Analysis

The statistical analysis was conducted in the R environment (version 4.1.1; R Core Team, 2021). First of all, we tested the consistency of the anchor dilemmas between the lists, assuming no differences between categories (TO, TS, DO, DS) and specific anchors among lists, in order to confirm the correctness to combine the four list groups in a single dataset. We focused on potential differences among the lists in terms of decision times, moral decision (utilitarian, nonutilitarian), and emotional activation (valence and arousal). Moving to the complete dataset (not considering separation by lists), six dependent variables were taken into consideration in six correspondent models: (M_1) decision time, (M_2) moral decision, (M_3) valence, (M_4) arousal, (M_5) moral evaluation of the nonutilitarian option, and (M_6) moral evaluation of the nonutilitarian option. Given the nature of the data, a generalized linear model (for moral decision) and five mixed effects linear models were fitted to the data using the R package *lme4* (Baters et al., 2015), setting the participant as random variables. The models presented in the main analysis (M_1 - M_6) resulted from six corresponding forward stepwise model comparisons, considering models with several different predictors, and choosing the best-fitting ones on the basis of the Akaike Weights comparison procedure (Wagenmakers and Farrell, 2004). Post hoc pairwise comparisons were considered when requested, using the R package *emmeans* (Lenth, 2020). Bonferroni's correction was set as an adjustment method. A 98% acceptance interval was considered in terms of decision times, which were transformed in their logarithmic form consistently with Lotto et al., (2014). The final dataset and further supplemental information are retrievable in the OSF project folder: https://bit.ly/3cksq6Q.

3.4 Results

Anchors and fillers

No significant interaction was observed between dilemma category and lists neither in terms of decision times (χ^2 9 = 4.26, *p* = 0.89), nor looking at the proportion of endorsement of a particular outcome (χ^2 9 = 3.68, *p* = 0.93). The same evidence resulted from the interaction between individual anchors and lists (decision times: χ^2 12 = 6.43, *p* = 0.89; moral decision: χ^2 12 = 10.42, *p* = 0.58), il line with our prediction on the consistency of anchors dilemmas between lists. Further confirmation was retrieved when investigating arousal (decision times: χ^2 12 = 6.43, *p* = 0.89; moral decision: χ^2 12 = 10.42, *p* = 0.58) and valence (decision times: χ^2 12 = 6.43, *p* = 0.89; moral decision: χ^2 12 = 10.42, *p* = 0.58). These results confirmed the homogeneity hypothesis about the four lists, based on non-different trends responding to the anchors dilemmas from the four investigated categories. Considering this, we combined the four lists in a single group, computing the next analysis on the complete dataset without list-specification. Additionally, four independent mixed models were fitted on the nonsacrificial dilemma set (i.e., fillers) to detect potential gender differences at this level (Lotto et al., 2014), considering decision times, moral decision, valence, arousal, and moral acceptability as dependent variables. No significant differences between women and men were observed in any of the computed models.

Complete dataset

A general representation of the predictors included in the six models, consistently with the effects sizes and the obtained estimates from the computed analysis, are retrievable at Table 2. Additionally, in support of the described inferential statistics, descriptive information is summarized in Table 3 and Table 4, divided by dilemma category and level of involvement of the moral agent.

Following the evidence obtained from the computed onward stepwise regression, the Decision Time mixed model (*m1*) was fitted including as fixed effects: Dilemma Category (Driving, Nondriving), Risk-Involvement (Self, Other), Moral Decision (Utilitarian, Nonutilitarian) and Experimental Order (from 1 to 18). Additionally, the interaction between Risk-Involvement and Dilemma Type was investigated.

Table 2: Beta estimates e p-values from M_1 to M_6.

	N (%)	**M_1 Decision time**	**M_2 Moral decision**	**M_3 Valence**	**M_4 Arousal**	**M_5 Moral evaluation (NUT)**	**M_6 Moral evaluation (UT)**
Moral decision							
Nonutilitarian (NUT)	15 %	-	-	-	-	-	-
Utilitarian (UT)	85 %	$-.37^{***}$	-	0.15^{*}	$-.17^{*}$	$-.76^{*}$	$.74^{***}$
Gender							
Female	75	-	-	-	-	-	
Male	76	-	-	$.37^{*}$	-0.81^{**}	.14	$.85^{***}$
NR	1	-	-	-	-	-	-
Age		-	-	$-.05^{*}$	-	-	-
Risk Involvement					-		
Self (S)		-	-	-	-	-	-
Other (O)		$-.35^{***}$	.43	.04	$-.13^{*}$	$-.23^{***}$	$.11^{*}$
Experimental Order		$-.08^{***}$	-	-	.00	-	-
Dilemma category							
Traditional (T)		-	-	-	-	-	-
Driving-Type (D)		$.78^{***}$	-1.24^{***}	-.01	.11	$.12^{***}$	.02
Dilemma Type							
DS – DO		$.35^{***}$	-.43	-.08	.15	$.23^{*}$	-.11
DS – TS		-.07	$.59^{***}$	-.07	-.04	$-.40^{*}$	.07
DS – TO		-.09	$.80^{**}$	-.06	.04	$.22^{*}$	-.14
DO – TS		$-.42^{***}$	1.03^{***}	.00	$-.20^{*}$	$-.63^{*}$	$.19^{*}$
DO – TO		$-.45^{***}$	1.24^{***}	.01	-.12	.01	-.02
TS – TO		.03	.21	.01	.08	$.62^{*}$	$-.21^{*}$
Emotional State							
Valence		-	-	-	-0.29^{***}	-	-
Arousal		-	-	$-.22^{***}$	-	-	-
Moral evaluation							
NUT Outcome		-	-	-	-	-	20^{***}
UT Outcome		-	-	-	-	$.22^{***}$	-
R^2 marg (R^2 adj) *		.15 (.36)	.05 (.30)	.13 (.55)	.10 (.67)	.11 (.59)	.12 (.65)

*Notes: *p < .05; **p < .01; ***p < .001; NUT = nonutilitarian, UT = utilitarian.*

Overall, results showed faster decision times in driving dilemmas ($\chi^2{}_1 = 43.10$, $p < .001$; Figure 8) then in nondriving ones, as well as in dilemmas not involving the moral agent as potential victim ($\chi^2{}_1 = 17.05$, $p < .001$). A further interaction effect was detected between these two factors ($\chi^2{}_1 = 24.41$, $p <$

.001), and post-hoc comparisons highlighted faster times in response to driving-type other-involvement dilemmas (DO), when compared with the same level of risk in traditional scenarios (DO – TO: t = -8.14, *p* < .001), and with its self-involvement driving-type counterpart (DO – DS: t = -6.41, *p* < .001). Slower decision times were detected when endorsing the nonutilitarian resolution ($\chi^2{}_1 = 38.62$, *p* < .001), and a gradual progressive in velocity was observed throughout the experimental session ($\chi^2{}_1 = 427.66$, *p* < .001), which is also evident in Figure 9.

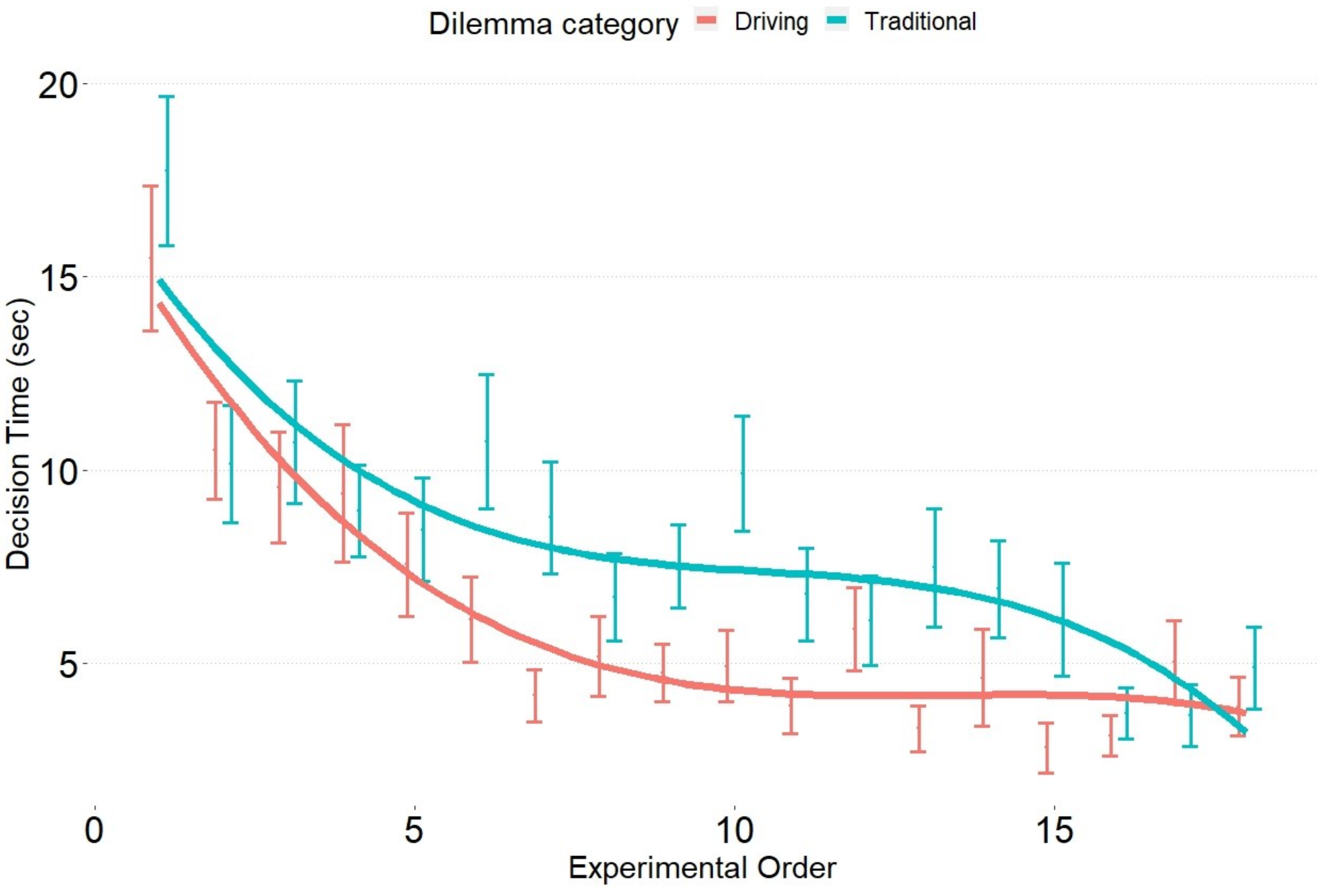


Figure 9: Smoothed curves with error bars representing means and standard errors for decision times (in seconds), divided by experimental order and dilemma category (traditional nondriving, driving).

The moral decision model (*m2*) was implemented as a generalized mixed effect linear model, setting the binomial family distribution as the reference point. The comparison of the models suggested the selection of the model with Dilemma Category and Risk-Involvement as fixed effects, as well as their interaction. Unexpectedly, no differences was observed between Self Vs. Other-risk involvement ($\chi^2{}_1$ = 0.05, p = .81). More utilitarian resolution were counted in the driving dilemma set (89%) when compared to the traditional nondriving set (80%, $\chi^2{}_1$ = 47.35 p < .001), which was consistent between levels of involvement ($\chi^2{}_1$ = 6.29 p = .012, Figure 10).

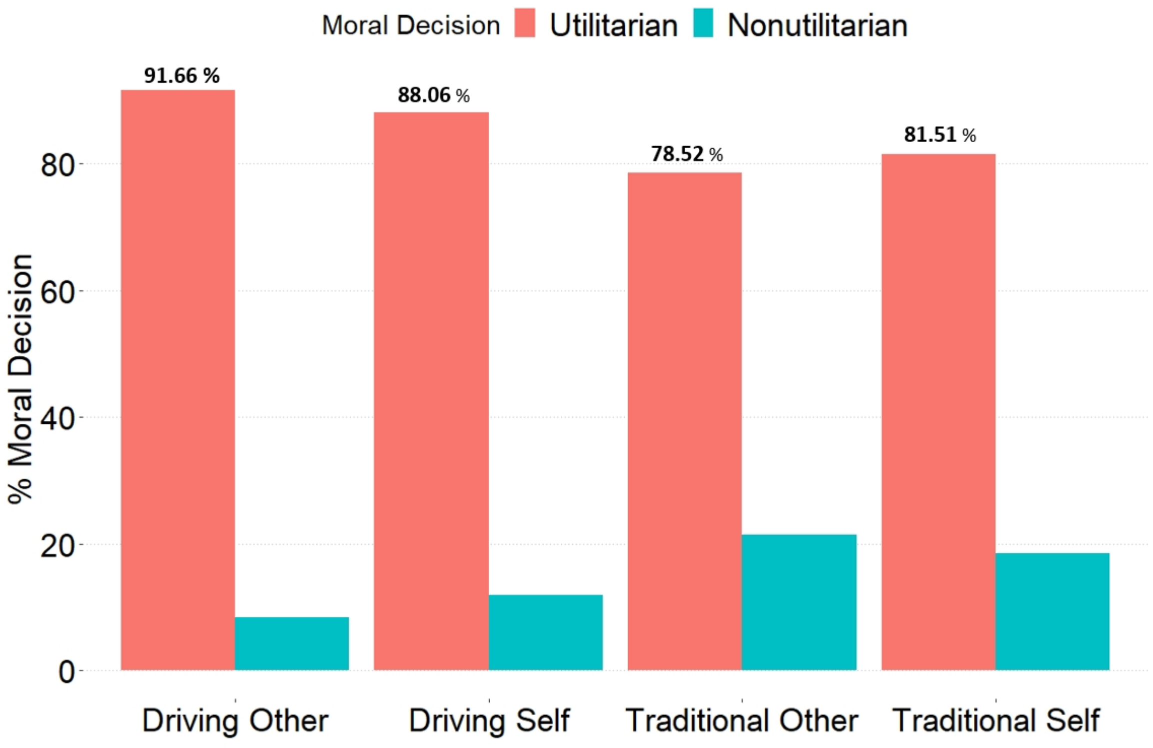


Figure 10: bar chart of moral decision percentage frequencies, divided by level of risk-involvement and dilemma category in columns, and by moral decision by color (utilitarian in red, nonutilitarian in light blue)

Subsequently, emotional activation was investigated, setting valence (*m3*) and arousal (*m4*) as dependent variables of two similar mixed models including as fixed effects Dilemma Category, Risk-Involvement (and their interaction), Moral Decision, Gender and the opposite emotional parameter considered (Arousal in *m3*, Valence in *m4*). No differences were observed between driving and traditional dilemma in neither of the two indices (*m3*: $\chi^2{}_1$ = 1.46, p = .22; *m4*: $\chi^2{}_1$ = 3.18, p = .0.07), and in the interaction with the risk-involvement factor (*m3*: $\chi^2_1 = 0.70$, $p = .40$; *m4*: $\chi^2_1 = 0.17$, $p = .0.67$). Albeit slightly, self-involvement dilemmas were perceived as more arousing than their counterpart (*m4*: $\chi^2_1 = 4.86$, $p = .0.03$; Table 3), while no differences were observed in terms of valence. Overall, the nonutilitarian behavior was confirmed as the more arousing (*m3*: $\chi^2_1 = 4.37$, $p = .03$; mean scores: 6.09 Vs. 5.74) and the one characterized by less positive emotional activation (*m4*: $\chi^2_1 = 4.61$, $p = .03$; mean scores: 2.53 Vs. 2.66). The inverse relationship between arousal and valence was fully confirmed (*m3*: $\chi^2_1 = 174.49$, $p < .001$; *m4*: $\chi^2_1 = 155.50$, $p < .001$), and women consistently reported higher levels of arousal (*m4*: $\chi^2_1 = 8.87$, $p = .002$) and lower levels of valence (*m3*: $\chi^2_1 = 4.30$, $p = .03$).

Table 3: Mean and Standard Deviation of the dependent variables considered, divided by Dilemma Type (nondriving, driving-type) and Risk Involvement (self, other involvement).

	Traditional dilemma	**Driving-type dilemma**	**Self involvement**	**Other involvement**
Decision time (sec)	8.34 (11.14)	6.06 (8.88)	7.41 (9.66)	6.99 (10.36)
Moral decision: utilitarian	80.12 %	89.07 %	85.60 %	83.59 %
Valence	2.69 (1.72)	2.68 (1.64)	2.65 (1.72)	2.73 (1.64)
Arousal	5.84 (2.26)	5.80 (2.25)	5.86 (2.27)	5.79 (2.23)
Morality: Nonutilitarian action	2.01 (1.98)	1.74 (1.81)	2.07 (1.99)	1.68 (1.81)
Morality: Utilitarian action	2.29 (2.05)	2.26 (1.98)	2.23 (1.99)	2.32 (2.03)

Table 4: Mean and Standard Deviation of the dependent variables considered, divided by the interaction between Dilemma Type and Risk Involvement.

	Traditional Self (TS)	**Traditional Other (TO)**	**Driving-type Self (DS)**	**Driving-type Other (DO)**
Decision Time (sec)	7.87 (10.21)	8.80 (12.01)	6.96 (9.11)	5.17 (8.64)
Decision Type: utilitarian	81.51 %	77.02 %	87.98 %	91.66 %
Valence	2.68 (1.80)	2.70 (1.63)	2.61 (1.64)	2.75 (1.65)
Arousal	5.88 (2.29)	5.80 (2.22)	5.83 (2.25)	5.77 (2.25)
Morality: Nonutilitarian action	2.25 (2.10)	1.76 (1.87)	1.88 (1.87)	1.60 (1.75)
Morality: Utilitarian action	2.27 (1.99)	2.32 (2.10)	2.20 (1.98)	2.32 (1.97)

Finally, moral evaluations toward the two proposed outcomes were deepened through two further mixed models (*m5*: nonutilitarian outcome acceptability*, m6:* utilitarian outcome acceptability), which considered the fixed effects described for models *m3* and *m4*, as well as the Moral Evaluation towards the opposite moral outcome. An effect of the type of dilemma was detected in the evaluation of the nonutilitarian outcome (*m5*: $\chi^2_1 = 17.23$, $p < .001$), with lower scores while judging this option in the driving framework (Table 3). No differences were observed in the evaluation of the utilitarian option. The nonutilitarian option was described as more moral acceptable when the moral agent was involved as a potential victim of the utilitarian option in both the types of dilemmas (DO – DS: t = -3.33, p = .005; TO – TS: t = -.962, p < .001), and especially in the traditional nondriving set (DS – TS: t = -6.08, p < .001; Figure 11). Expectably, the non-chosen moral option was described as less morally acceptable than the chosen one (*m5*: $\chi^2{}_1 = 110.66$, $p < .001$; *m6*: $\chi^2{}_1 = 105.58$, $p < .001$), and gender differences were revealed only at the utilitarian level, with men who described this option as more moral than women (*m6*: $\chi^2{}_1 = 14.01$, $p < .001$).

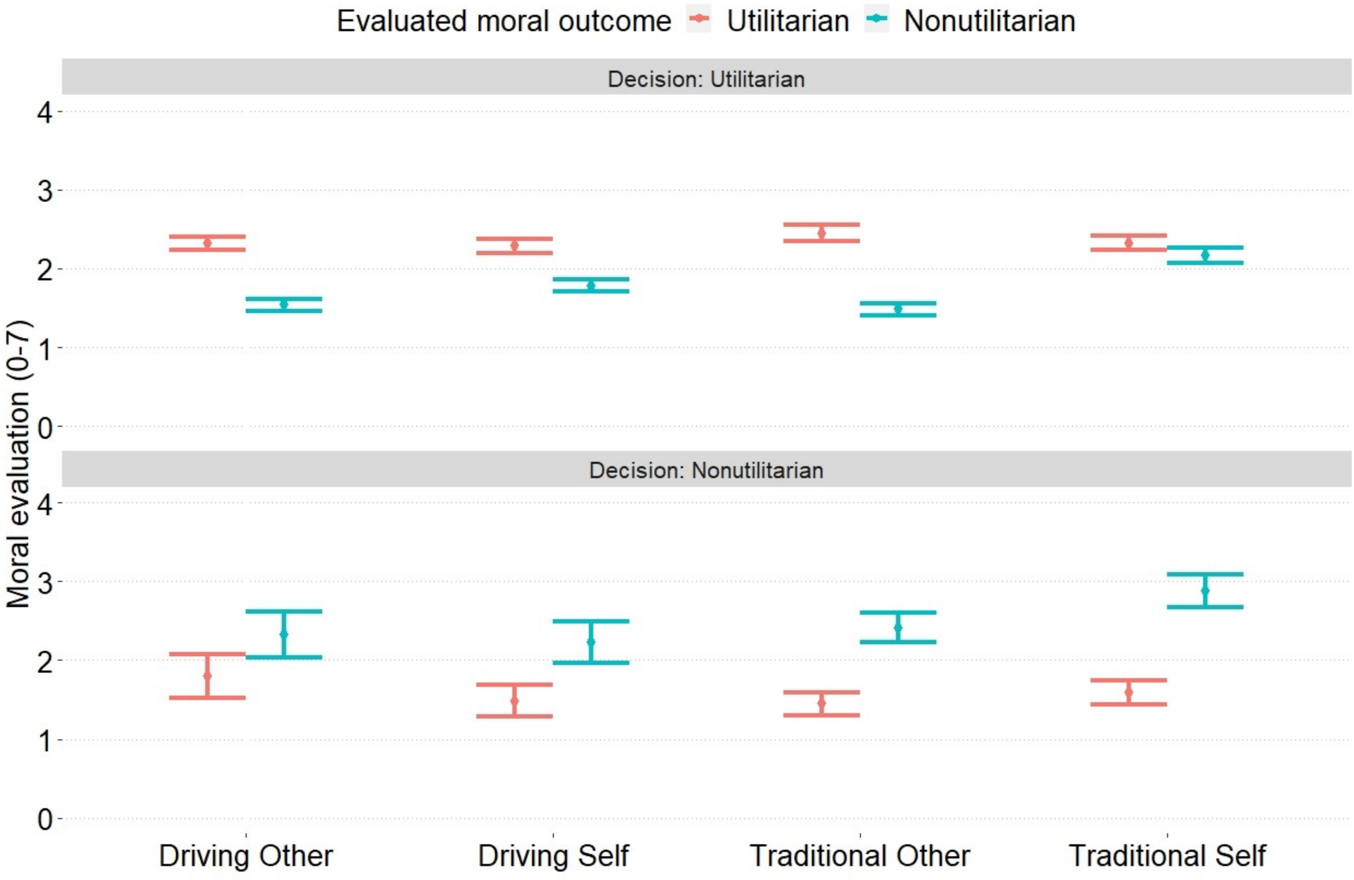


Figure 11: Error bars plot representing means and standard errors of moral evaluations (utilitarian in red, nonutilitarian in light blue), divided by moral decision in rows, and risk-involvement and dilemma category in columns.

3.5 Discussion

The present study was conceived to investigate for the first-time moral judgment in the specific context of driving activity, focusing on the most-common manual driving vehicles and using a readaptation of sacrificial and incidental moral dilemmas. To this aim, a new set of 20 driving-type sacrificial dilemmas was developed and then tested following a validated experimental procedure (Sarlo et al., 2012; Lotto et al., 2014) and selecting for an online distribution modality. The new set was then compared with an already validated set of traditional nondriving moral scenarios, selected by Lotto et al. (2014). The research was conducted controlling for the level of personal risk-involvement, in order to detect potential differences in the endorsement of a particular moral code when the moral agent life was at stake. Information were collected in terms of moral endorsement and moral acceptability, decision times and emotional activation (valence and arousal) at the time of the decision. The corresponding normative scores collected on the new driving set are reported in the supplementary material of the project.

Overall, results showed clear similarities between driving and nondriving dilemmas, resulting in a striking endorsement of the utilitarian resolution (above 80%) in both sets, comparable scores of medium emotional activation and unpleasantness, as well as low moral acceptability of the two proposed resolution, independently from the chosen moral decision. This evidence suggests the activation of non-different moral decision-making processes when facing sacrificial and incidental driving and nondriving dilemmas, potentially highlighting a trivial role of dilemma context in shaping moral judgment. Therefore, this result will lend credence to our first hypothesis, so to the pivotal role of dilemma-structure in the interpretation and resolution of sacrificial dilemmas, that goes beyond the contextualization of their storyline (Schein, 2020). A 'structure-based' interpretation of moral dilemmas would stand for a generalization of moral principles and decision processes across different contexts, which would be supported by non-different declinations of moral judgment and related emotional activation. Several studies discussed the potential decontextualized nature of moral psychology, claiming for a low sensitivity of moral judgment to contextual factors (Bauman et al., 2014; Bloom, 2011; Schein, 2020). Clearly, the development of moral judgment depends on several other scenario's characteristics, such as the number of casualties (e.g., Bonnefon et al.2016) or the relationship with the characters (e.g., Shaw et al., 2017; Simpson et al., 2016), but when controlling the moral scenario for these conditional factors, the structure may take the lead in the development of the moral judgment. This seems the case in the present research, where the two dilemma sets were equally structured as sacrificial and incidental (following the DDE), and no further contextual information was provided to the moral agent other than their role (driver or non-driver). Nonetheless, from a statistical perspective, the resolution of driving-type dilemmas resulted in significant faster decision times and a higher percentage of utilitarian endorsement, despite the level of personal risk. A consistent research line criticizes moral dilemmas with a low real-life plausibility, that often refer to unrealistic circumstances that may have a detrimental effect on moral judgment (Bauman et al., 2014; Gold et al., 2014; Watkins, 2020). A scarce plausibility of sacrificial dilemmas has also the effect to systematically increase the appealing of the deontological moral code, as the result of a potential distortion of the moral judgment process (Körner et al., 2019). From our perspective, drivers can imagine more easily risky driving situations as potential, albeit rare, daily challenges. This is especially true when compared with traditional dilemmas, which are typically framed into extreme and unrealistic circumstances (for instance, see the 'Soldier' and the 'Underground Cave' dilemmas in the Appendix). Considering this, we believe that when moral judgments are applied to lifelike situations, cognitive demands decrease (Conway and Gawronski, 2013; Schein, 2020; Sütfeld et al., 2019), enabling the individual moral inclination to emerge faster and sharper. This result appears resistant to gender differences, as also confirmed by previous studies (Navarrete et al., 2012; Palmiotti et al. 2020).

Interesting insights arise from the improved responsiveness to both driving and nondriving dilemmas during the experimental procedure. In the context of applied moral psychology, several studies have focused on the role of time pressure in shaping moral decision-making, resulting in contrasting evidence (Goeschl and Lohse, 2018; Jacquet and Cova, 2021; Suter and Hertwig, 2011; Tinghög et al., 2016; see Study 3), but few enlightened the continuous reduction of decisional time during the resolution of iterated dilemma studies. The number of experimental stimuli is a well acknowledged factor in the computation of the statistical power, in the aim of a reliable statistical computation of the collected data (e.g., Baker et al., 2020; Lerche et al., 2017). Nonetheless, answer to a textual moral dilemma is not a trivial task (Broeders et al., 2011) since it requests a relevant amount of time and cognitive effort. Having this in mind, presenting a disproportionate number of scenarios may have a detrimental effect. Drifting from a truthful definition of the individual moral inclination is a plausible risk, which may be caused by several factors, mainly related to a reduced commitment to the experimental task (e.g., cognitive fatigue, boredom, hurry to conclude the task). The evidence collected in the present research has the potential to stimulate a new discussion for the methodological optimization of the iterated approach to moral investigation.

Interestingly, no differences were observed in terms of moral judgment - end emotional activation - between self and other-involvement dilemmas, irrespectively from the dilemma category. A large number of studies have underlined the important role of individual risk in shaping moral decisions (e.g., Huebner and Hauser, 2011; Lotto et al., 2014; Petrinovich et al., 1993), focusing on self-protective actions and on the improved interest in the utilitarian moral code as the number of potential casualties grown (Bergmann et al., 2018; Bonnefon et al., 2016). Additionally, the direct involvement of the moral agent appears even more critical when looking at the driving activity: a concrete life-threat can be quite a conflictual factor in the moral decision-making process, both if in the shoes of the driver or of the vehicle's passenger. In our vision, it is possible that framing self-sacrifice only in the nonutilitarian option (consistently with Lotto et al., 2014) had the unpredicted effect to reduce this conflict, resulting in a reduced impact of self-involvement scenarios. This point deserves further investigation and has been tackled in the Study 2 of the present project.

The evidence collected from the two emotional activation indices, valence and arousal, seems in line with the structure-based hypothesis on the interpretation of moral dilemmas. As expected, the inverse relationship between arousal and valence was confirmed (higher emotional activation, low pleasantness), but no differences were observed between driving and nondriving dilemmas, both from a quantitative and from a qualitative point of view. Overall, the nonutilitarian behavior was confirmed as the more arousing option and with the lower (albeit slightly) valence, but surprisingly, the utilitarian behavior was endorsed faster than its counterpart, which seems in partial opposition with our hypothesis and with the DPT (Greene et al., 2001). In fact, we would have expected slower utilitarian resolution than nonutilitarian, coherently with the observed weaker and less negative emotional activation. Several of considerations may rise at this point: first, it is possible that the emotional activation at the time of judgment is less detectable through online experiments. The limited control over participants' behavior during this kind of activity, and the potential role of external factors, allows the possibility that the online modality are less sensitive to self-reported evaluation at a specific time, as at the moment of the decision (Hänggi, 2004). For this reason, a possibility is to investigate emotional correlates to moral judgment following a different approach, and assuming the bright sides and the limitation of the method. Second, the non-randomization of the two moral outcomes may have played a contrasting role in detecting emotional activation at the time of decision. The presentation of the two options in a fixed order was taken to be consistent with previous validation procedures (Palmiotti et al., 2020; Sarlo et al., 2012) and with the storylines validated by Lotto et al. (2014). Methodologically, this technique was assumed as an attempt to postpone – as much as possible – the reasoning process only at the presentation of the decision page. Furthermore, the structure of self-involvement dilemmas limited the possibility to counterbalance the two options between and within participants, since the full interpretability of the utilitarian option was allowed only if presenting the nonutilitarian option first. Study 2 will tackle both the issues, evaluating the referred intensity of a number of moral emotions after the moral decision, and readapting the experimental stimuli in a way that allows the randomization of their alternatives.

Finally, the dissociation between moral decision and moral evaluation was confirmed by the data, consistently with the incongruency already detected by Bonnefon et al. (2016) in the specific AV case. A small advantage of the assumed decision was yet observed in the evaluation of acceptability of the proposed moral resolutions (especially in the nonutilitarian case, Figure 10), but overall, a low acceptability of the two moral codes was detected, independently from the assumed decision.

In conclusion, Study 1 allowed us to test the applicability of the traditional sacrificial dilemma in the field of transportation, especially in the well-known context of traditional human driving with no implemented automation. New evidence was brought in favor of the structure-based interpretation of moral dilemmas, greenlighting the possibility to apply sacrificial and incidental dilemmas to the context of driving activity. Nonetheless, a clear advantage of lifelike situations was observed in eliciting the utilitarian resolution quickly and sharply. The driving activity seems a clear example of this phenomenon, as on-road storylines are easier to contextualize than traditional unrealistic dilemmas. A series of theorical and methodological questions still stand on the structure of textual sacrificial dilemmas and their presentation in an online form. Study 2 tried to answer these questions, comparing moral judgment in opposite levels of driving automation.

Chapter 4

Study 2: Framing self-sacrifice in the investigation of moral judgment and moral emotions in human- and autonomous-driving dilemmas

4.1 Rationale of the study

Study 1 gave us the chance to test the applicability of sacrificial and incidental dilemmas in the context of driving events, in comparison to a validated set of traditional dilemmas. This study allowed us to detect differences and similarities between more realistic human-driving dilemmas and typical un-customized scenarios, which we considered to be the baseline for the development of further advanced investigations in the field of autonomous transportation. At this point, our interests shifted specifically on the mobility framework, taking advantage from the evidence collected in Study 1 for the development of a new investigation in the field of moral judgment.

In Section 2.2 we have highlighted the new interest in the moral evaluation of driving behavior, especially when applied to autonomous transportation. A relevant number of research focus on investigating moral evaluation of AVs' behaviors, mainly shaping and adapting the traditional trolley problem to the autonomous driving context. The structure of the AV dilemmas – initially proposed by Bonnefon et al., (2016) - follows the typical structure of the trolley problem, replacing the runaway train with the AV and the two tracks with two implementable driving maneuvers, which depict two opposite moral codes. This is very similar to what happens in human-driving moral dilemmas, but in this case the driving maneuvers are made directly by the moral agent acting on the wheel. Here we have the main difference between human and autonomous driving moral dilemmas: where the moral judgment and the actions made by the individual (as the driver) are the main cause of the consequences brought on by a human-driving vehicle, the chain of events caused by the predictable decisions of the AV are in any way influenced by the individual’s beliefs and contingent evaluations (as the passenger). Clearly - and ideally, since no completely autonomous vehicles are actually available in the mass market -, purchasing an AV also means accepting the possibility that its behavior may not be in line with what we would do in a particular driving situation, such as a potential road crash. In this sense, there might be the possibility to experience unintended events leading to undesired autonomous behaviors. The role of intention has been widely investigated in the moral literature, detecting a lower likelihood of causing intentional than unintentional harm, consistently with the observed contrast between the “impersonal” *Switch* and the “personal” *Push* dilemmas (e.g., Borg et al., 2006; Cushman et al., 2006; Moore et al., 2008). In this sense, also the expectation towards artificial intelligence (AI)’s behaviors may play an important role on the moral perception of driving events. Albeit the utilitarian moral code is perceived as the default moral norm both for humans and AIs (Kallionen et al., 2019; Li et al., 2016), autonomous agents are expected to strictly behave in this way, despite less responsibility would be assigned to them in case of failure (Malle et al., 2015). Specifically, AVs are perceived as less blameworthy than human drivers (Pizarro et al., 2003), and a reduction of direct responsibility is consistently ascribed also to humans when onboard of an AV (Gill, 2021). This evidence support the possibility of observing different moral decision processes towards human and AVs, but no studies focused on this comparison yet.

Furthermore, Study 1 allowed us to focus on a specific characteristic of sacrificial dilemmas. We know that dichotomous moral dilemmas allow the respondent to choose between two moral options: utilitarian and nonutilitarian (or deontological). In the specific case of self-involvement dilemmas, traditional dilemmas (Greene et al., 2001; 2004; Lotto et al., 2014; Moore et al., 2008) always jeopardized the life of the moral agent in the nonutilitarian outcome. The following dilemma is an example retrieved from Moore’s dilemma set (2008):

THE BURNING BUILDING: You and five other people are trapped in a burning building. There is only one emergency exit through which all of you could escape to safety, but it is blocked by burning debris. You notice another person in the hallway leading to the exit who has been injured but is about to crawl to safety through a small hole at the bottom of the exit door. You and the five people behind you do not have time to climb through the small hole.

The hallway's emergency system puts out fire by eliminating oxygen from the hall and you can activate the system by pressing a nearby button. The fire will go out, but the injured person will suffocate and die. However, if you do not do this, you and the five people behind you will die.

In order to save the majority and themself, in this dilemma the moral agent has to take a decision that will cause the death of a single individual. In other words, the endorsement of the utilitarian behavior matches the self-protective option (i.e., "*I live, and many survive*"). This appears as a fundamental difference between traditional self-involvement dilemmas and AV dilemmas. In fact, when the moral agent is depicted as the passenger of an AV, its sacrifice occurs only upon the selection of the utilitarian option (i.e., "swerve off to the side of road, where it will impact a barrier and plunge into the river, killing you and your [family member / coworker] but leaving the pedestrians unharmed", Bonnefon et al., 2016). Here, the endorsement of the utilitarian behavior collides with the self-protective option, resulting in the acceptance of self-sacrifice in order to achieve the greater collective goal (i.e., *"I die, but many survive"*). Additionally, in order to achieve the – utilitarian - self-protective outcome in traditional 'non-AV' dilemmas, the moral agent has to endorse a proactive behavior (i.e., pressing the nearby button to activate the emergency system), which deviates from the natural course of the events. Oppositely, the endorsement of the – nonutilitarian - self-protective option in the AV dilemma request the moral agent to endorse a passive behavior (i.e., accepting that the AV will maintain its current path).

Altogether, Study 1 showed non different emotional activation pattern between nondriving and driving dilemmas, reproducing – albeit partially - the expected distinctions between utilitarian and nonutilitarian endorsement at the moment of the decision (Greene et al., 2001; Lotto et al., 2014). Study 1 investigated the general activation of unspecified emotions at the time to the moral decision-making process, in line with the hybrid approach of contemporary integration between emotion and deliberation in the making of the decisional process. In Study 2 we opted to investigate moral judgment from a 'pure Kantian' perspective (Huebner et al., 2009, see section 1.2.2), assuming affective reactions generated as a consequence of the reasoning process. Indeed, in evaluating a particular moral issue, moral judgment is influenced by moral emotions, which assume a mediation role between norms and the produced moral decision (Tangney et al., 2007). Kroll and Egan (2004) describe how moral emotion typically arise in response to daily events that push people to behave (or avoid behaving) in a morally ascribable way, following collective or personal interests (Greenbaum et al., 2020; Haidt, 2003b). They can affect the moral agent before the actuation of the moral decision, during the evaluation of potential alternatives, remaining detectable after the decision itself (Tangney et al., 2007).

Haidt (2003b) described moral emotions relying on two main categories: other-condemning and self-conscious moral emotions. Other-condemning emotions (i.e., other-referred emotions, such as contempt, anger, and disgust) are negative feelings directed towards other individuals or groups, in response to their moral violations. Oppositely, self-conscious emotions (i.e., self-referred emotions, such as shame, embarrassment, and guilt) are directed to the self, in case of violations of personal moral standards. In the following study, we will focus on two other-condemning and two self-conscious emotions: anger, disgust, shame and guilt. Descriptively, anger is considered the most prototypical other-condemning emotion of the category, motivating concrete actions to restore the disrupted moral order (Haidt, 2003b; Hutcherson and Gross, 2011), while disgust is a repulsive reaction that arise in contrast to murky moral conducts (Rozin et al., 1999; Schnall et al., 2008). Shame, for its part, is a public-oriented self-conscious emotion, focused on a global negative evaluation of the self in reaction to particular moral event. On the other hand, guilt is a private-oriented self-conscious emotion, that arises as a negative self-condemnation specifically referred to a specific behavior (Lewis, 1971). Assuming the important difference in the sense agency and responsibility between human (moral agent as the driver) and autonomous driving vehicles (moral agent as a passenger), the distinction between these two categories of emotion can be a fruitful integration in the investigation or emotional correlates to moral judgment (Malle et al., 2014; McManus and Rutchick, 2019).

In conclusion, Study 2 aimed to compare human and autonomous driving vehicles in terms of moral judgment and moral emotions, assuming potential differences between these two opposite modes of transportation. Additionally, the self-sacrifice framing of the moral dilemma was considered, as a fundamental structural differences of this experimental tool between traditional nondriving and autonomous driving scenarios.

4.2 Hypothesis

A series of hypotheses were advanced in the present study, considering the evidence collected in Study 1 and the thoughts described in the rationale of the present chapter:

- No differences are expected between human and autonomous driving moral dilemmas in terms of moral decision and decision times, consistently with the structure-based interpretation of moral dilemmas (Schein, 2020). Nonetheless, in case of differences, a higher utilitarian endorsement and faster decision times would be admissible in the human-driving case, considering results from Study 1 on the relatively great plausibility of daily scenarios (Körner et al., 2019).
- In both the driving categories, a lower endorsement of the utilitarian outcome - as well as a correspondent worst moral evaluation - is expected when the utilitarian option leads to

the moral agent's self-sacrifice (i.e., *"I die, but many survive"*), when compared to the self-protective utilitarian resolution (*"I live, and many survive"*).

- A significant effect of the level of automation on moral emotions' perception, with higher intensity of other-referred emotions responding to AV dilemmas and higher intensity of self-referred moral emotions facing human-driving dilemmas.
- A significant effect of the self-sacrifice framing on moral emotions' perception, with higher intensity of other-referred emotions after the endorsement of self-protective outcomes, and higher intensity of self-referred emotions after the endorsement of self-sacrificial outcomes.

4.3 Method

4.3.1 Participants

An a-priori power analysis has been computed on G-power statistical software (Faul and Erdfelder, 1992) before the computation of any statistical analysis, assuming a small effect size (Cohen's d = 0.10), and a correlation of 0.50 among repeated measures, with an alpha error probability of 0.05 with 0.90 power. The system suggested a minimum of 140 participants, and a total of 183 participants were recruited for the experiment. Females accounted for 51.36% of the final sample (94 women). Overall, the mean age was 27.82% (SD = 10.55, range = 18–66), and 65.22% of participants were enrolled in university courses (n = 120), with 50.55% (*n* = 92) matriculated in human sciences and cultural-related programs (e.g., psychology, sociology, philosophy). Most participants (90.21%) had held driver licenses (n = 166), and most of them (75.54%, n = 139) drove a maximum of 15,000 km per year. Almost half of the sample (48.91%, n = 90) was involved in a car accident in their lives, and only 4.35% (n = 8) were involved in one of them in the prior 12 months. 87.5% (n = 161) of the sample had already heard about the AV technology. Additionally, in order to assess positive and negative affect at the time of the participation, the Positive and Negative Affect Schedule scale was administered (Terracciano et al., 2014; Watson et al., 1988). A mean positive-affect score of 31 (*SD* = 7.20) and a mean negative-affect score of 22.80 (*SD* = 8.03) was detected in the sample, with no differences between men and women. The study was administered through Qualtrics software (Qualtrics, Provo, UT) and approved by the local ethics committee (ID No.: 3514). Before their participation, each participant gave formal written consent, which was voluntary and unremunerated.

4.3.2 Materials

Twelve self-involvement sacrificial and incidental moral dilemmas were developed for this study, following the structure and the indications of Study 1. As incidental dilemmas, the resulted sacrifice was interpreted as a foreseen but unintended consequence for the achievement of the greater goal. The dilemma set was divided in six human-driving and six autonomous driving dilemmas. In the first case, the moral agent was depicted as the driver of a nonautonomous vehicle (SAE's level 0), with their hands on the wheel and totally in charge of the vehicle's maneuvers in response to the critical situation. In the AV case, the moral agent was the passenger of a completely autonomous vehicle (SAE's level 5), with no power to change the AV's decisions on how to react to the events depicted in the moral dilemmas. In both cases, the moral agent had to figure themself facing – together with another passenger on board - a particular non-critical traffic event, such as overtaking a slower vehicle (see Appendix for the full dilemma set). In each storytelling, an unpredicted critical problem was presented, forcing the driver/AV to react with a driving maneuver that was coherent or against the utilitarian moral code. The critical events were related to three kind of obstacles that may be found while driving: an unpredicted pedestrian crossing, a problem of visibility, and an obstacle on the road. These three topics shaped the dilemmas in blocks of four units, in order to let the dilemma, set be focused on concrete driving situations. Before the data analysis, a preliminary statistical check confirmed the generalizability of the dilemmas in a single set, in disregard of their thematic application. Endorsing the utilitarian moral code clearly stood for protecting the majority of the characters involved, but in some scenarios this possibility did not guaranteed self-protection. Indeed, in six dilemmas (three per level of automation) the self-protective option was framed in the nonutilitarian outcome: if the moral agent aimed to endorse the utilitarian option, they had also to admit their own self-sacrifice (i.e., Utilitarian sacrifice framing, *"I die, but many survive"*). Oppositely, in the remaining six dilemmas, the moral agent was allowed to endorse the utilitarian option also protecting themselves (i.e., Nonutilitarian sacrifice framing, *"I live, and many survive"*), coherently with the traditional form of the sacrificial dilemmas (e.g., Moore et al., 2008). In Study 2, in order to allow the randomization of the two moral alternatives in each self-involvement dilemma, all the critical information about the moral situation were solely presented in the moral scenario, and not in the two outcomes. Finally, the sacrificial dilemmas employed in this study undergone to the same checks developed in Study 1 (see section 3.3.2), maintaining in the present case a 2:4 ratio between lives saved and lives sacrificed, considering the presence of the additional passenger in each dilemma. Table 5 illustrates an example for each dilemma (Human driving with Utilitarian Sacrifice, Human-US; Human driving with Nonutilitarian Sacrifice, Human-NS; AV driving with Utilitarian Sacrifice, AV-US; AV driving with Nonutilitarian Sacrifice, AV-NS). The complete dilemma set is available in Appendix and in the supplementary material (https://bit.ly/3cksq6Q).

Table 5: Sample of human-driving and autonomous driving dilemmas (text translated from Italian).

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

*Notes: * The two options were randomized across dilemmas*

4.3.3 Experimental Procedure

Consistently with Study 1, Study 2 was programmed and distributed via Qualtrics software. The program provided an anonymous link to the survey, which was then distributed via social networks and institutional communication channels following a snowball non probabilistic sampling technique (Goodman, 1961; Parker et al., 2019). The data collection was performed from October 15th, 2021, to December 30ty, 2021. All participants were required not to perform the survey through smartphones or tablets, but only using laptops, in order to avoid problems of data comparability between multiple devices (Krebs and Höhne, 2021).

To test the comprehensibility of the experimental task, a pilot study was conducted on 10 participants, that confirmed the feasibility of the experimental design and the intelligibility of the experimental material. The experimental procedure was highly consistent with Study 1 and is graphically described in Figure 12.

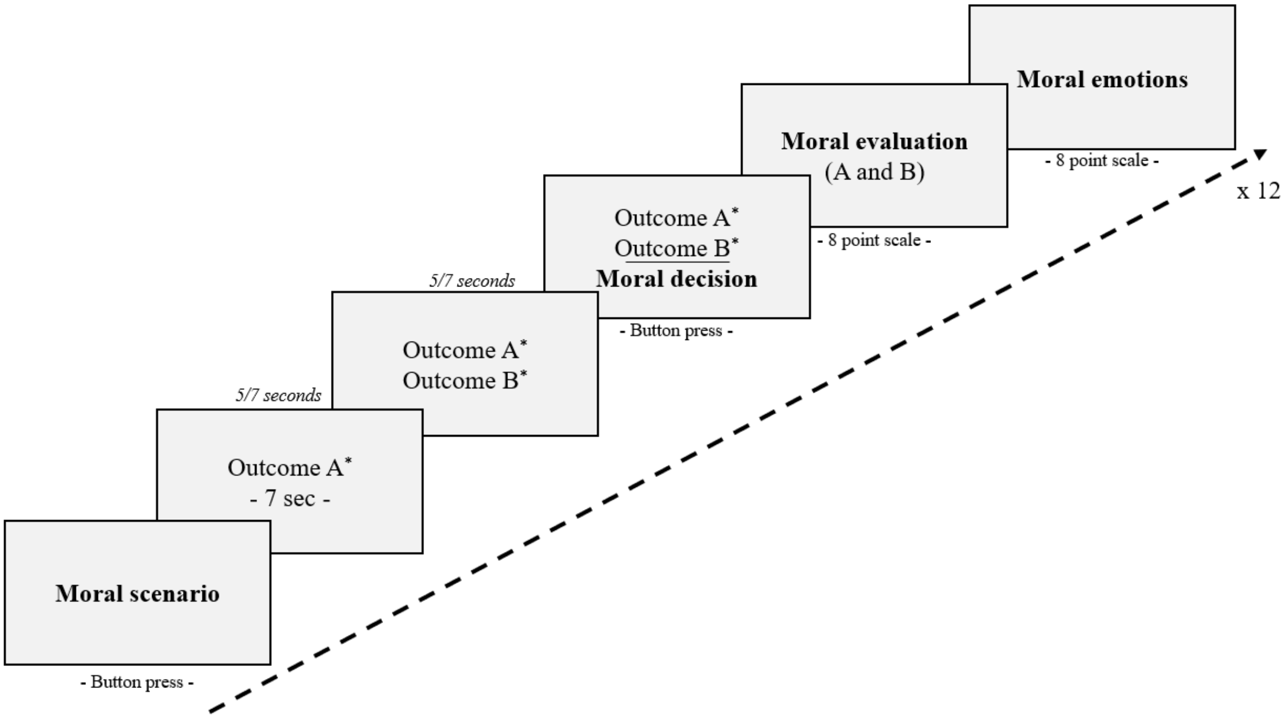


Figure 12: The experimental procedure of Study 2. The following sequence was repeated 12 times, one time per administered dilemma.

The mean completion time of the experimental procedure was 26.40 mins (SD = 15.57 mins). Before the beginning of any experimental activity, the participants were requested to read and fill out an informed consent about their participation and data protection regulation. Subsequently, the Positive and Negative Affect Schedule scale (PANAS) was administered, followed by a detailed explanation of the dilemma presentation phase (Figure 11). Twelve dilemmas (6 human and 6 autonomous driving) were randomly presented to each participant in a textual form. The sole scenario (without the two alternatives) was presented at the center of the screen, in black type (font Arial, size 10) against white background. At that point, the participant had illimited time to read the scenario, before moving to the presentation of the two outcomes (utilitarian and nonutilitarian), which were presented in a random order, differently from Study 1. Despite the randomization, the two moral outcomes were again presented on the screen in two different moments: the first option was presented and maintained on the screen 5 or 7 seconds, depending on its length. Then, the second option appeared below the previous one, with the same presentation time logic (see section 3.3.3 for further explanation or Sarlo et al., 2012; Lotto et al., 2014). Subsequently, the two option keys appeared on the screen below the options, and the participant was requested to indicate their preference (moral decision) coherently with the described moral code. Following the dilemma phase, the moral acceptability of the two proposed moral outcomes was collected on an 8-point scale (0 = completely unacceptable, 7 = completely acceptable), and then the same Likert scale (0 = no intensity, 7 = maximum intensity) was administered for assessing – after the moral decision - the perceived intensity of the four investigated moral emotions (other-referred: anger and disgust; self-referred: shame and guilt). The survey ended with a series of sociodemographic questions and driving habits.

4.3.4 Analysis

The statistical analysis was conducted in the R environment (version 4.1.1; R Core Team, 2021). As a preliminary step, we tested potential differences between the three dilemma's topics (unpredicted pedestrian crossing, visibility problem, road obstacle), in terms of moral decision (utilitarian, nonutilitarian) and the referred emotional reaction (anger, disgust, shame, and guilt). After this methodological assurance, eight dependent variables were considered for eight correspondent models: (M_1) decision time, (M_2) moral decision, (M_3) moral evaluation of the nonutilitarian option, (M_4) moral evaluation of the utilitarian option, and the four moral emotions (M_5 anger, M_6 disgust, M_7 shame, M_8 guilt). Given the nature of the data, seven mixed effects linear models and a single generalized linear model (M_2 moral decision) were fitted to the data through the R package lme4 (Baters et al., 2015), setting the participant as random variables. The models presented in the main analysis ($M_1 – M_8$) are the result of eight corresponding forward stepwise model comparisons, which considered models with a number of different predictors. The chosen one was selected based on the Akaike Weights comparison procedure (Wagenmakers and Farrell, 2004). Post hoc pairwise comparisons were considered when requested, using the R package *emmeans* (Lenth, 2020). Bonferroni's correction was set as an adjustment method. A 98% acceptance interval was considered in terms of decision times, which were transformed in their logarithmic form consistently with Lotto et al., (2014). The final dataset and further supplemental information are retrievable in the OSF project folder: https://bit.ly/3cksq6Q.

4.4 Results

As a preliminary check, no significant differences were observed between dilemma's topics, nor in terms of moral decision ($\chi^2{}_2 = .83$, $p = .66$), neither in terms of moral evaluation (nonutilitarian: $\chi^2{}_1 = .91$, $p = .63$; utilitarian: $\chi^2{}_1 = 1.20$, $p = .54$). Based on these results, the dilemma's topic was not considered as a potential factor in data analysis. Table 7 summarizes the predictors included in the selected models (M$_1$- M$_8$), coherently with the obtained estimates and the corresponding effect sizes.

Table 7: Beta estimates e p-values from M$_1$ to M$_6$.

	N	M$_1$	M$_2$	M$_3$	M$_4$	M$_5$	M$_6$	M$_7$	M$_8$
	(%)	Decision	Moral	NUT	UT	Anger	Disgust	Shame	Guilt
		time	Decision	evaluation	evaluation				
Moral decision									
Nonutilitarian	24 %	-	-	-	-	-	-	-	-
Utilitarian	76 %	-2.90**	-	-	0.93***	-.32	.07	.35***	.56***
Gender									
Female	94	-	-	-	-	-		-	-
Male	88	-	-.65**	-	.72	.14	.85***	-1.11***	-1.39***
Sacrifice Framing									
Ut-Sac (US)		-	-	-	-	-	-	-	-
Nut-Sac (NS)		-3.63***	-1.25***	-	-0.93***	.32	-.05	-.58***	-.65***
Exp Order		-1.05***	-	-	-.05	.04***	.02***	.04***	.01
Dilemma category									
Human (H)		-	-	-	-	-	-	-	-
AV		.66	-.10**	.12	-0.11	.10	.15**	-.26***	-.44***
Moral evaluation									
NUT option		-	-	-	-	-	-	-	-
UT option		-	-	-	-.01	-	-	-	-
R^2 marg		.06	.09	.01	.10	.05	.02	.09	.11
(R^2adj)*		(.19)	(.33)	(.02)	(.62)	(.79)	(.77)	(.72)	(.74)

*Notes: *p < .05; **p < .01; ***p < .001; NUT = nonutilitarian, UT = utilitarian; UT-Sac = Utilitarian Sacrifice framing, NUT-Sac = Nonutilitarian Sacrifice framing*

Looking at the decision times (M$_1$), the model regression suggested the linear mixed model which includes the Dilemma Category (human driving, AV), the Moral Decision (utilitarian, nonutilitarian), the Sacrifice Framing (utilitarian sacrifice, nonutilitarian sacrifice) and the Experimental Order (1-12) as fixed effects, as well as the interaction between Dilemma Category and Moral Decision. No difference was observed between human and autonomous driving ($\chi^2{}_1$ = .14, p = .71), whilst, consistently with Study 1, higher decision times were observed in case of nonutilitarian decision ($\chi^2{}_1$ = 8.03, p = .004). No interaction was observed between dilemma category and moral decision ($\chi^2{}_1$ = .57, p = .45), while participants requested more time for responding to nonutilitarian-sacrifice framed dilemmas ($\chi^2{}_1 = 31.24$, $p < .001$). This evidence is represented graphically in Figure 13, which also depicts the significant reduction of decision time throughout the experimental session ($\chi^2{}_{11} = 130.33$, $p < .001$).

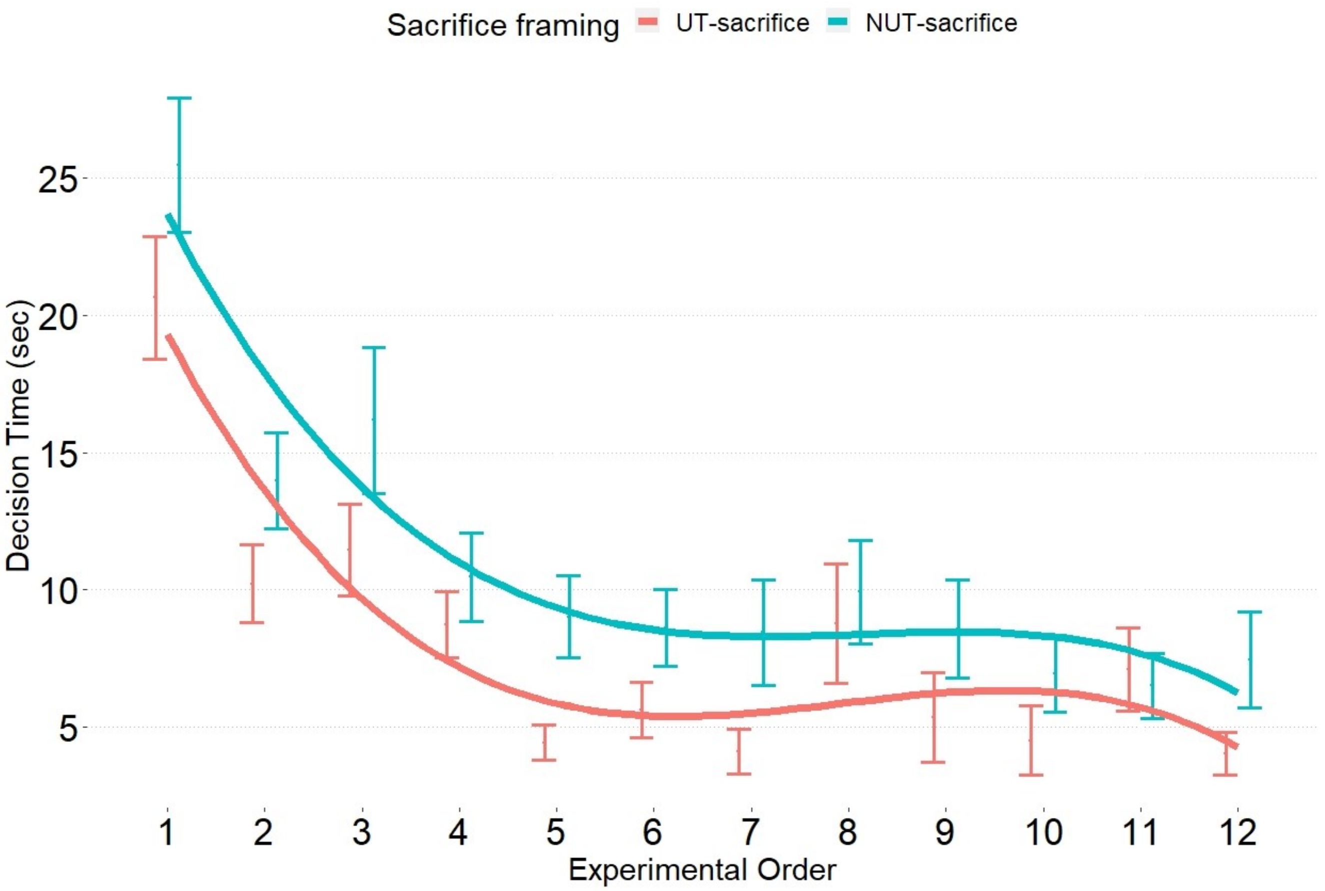


Figure 13: Smoothed curves with error bars representing means and standard errors for decision times (in seconds), divided by experimental order and sacrifice framing (UT-sacrifice: sacrifice framed in the utilitarian outcome, NUT-sacrifice: sacrifice framed in the nonutilitarian option).

The binomial distribution was set for implementing the generalized mixed model M_2 on moral decision. The model comparison suggested the model with Dilemma Category, Sacrifice Framing (and their interaction), and Gender as predictors. A small but significantly higher percentage of utilitarian decision was observed in response to human-driving dilemmas (79%), when compared to AV dilemmas

(74%, $\chi^2{}_1 = 6.78$, $p = .009$, see Figure 14). Predictably, a higher endorsement of the utilitarian resolution was also observed in nonutilitarian-sacrifice framed dilemmas ($\chi^2{}_1 = 88.97$, $p < .001$). No interaction between these two factors was detected ($\chi^2{}_1 = 2.17$, $p = .14$), and women appeared 'more utilitarian' then men ($\chi^2{}_1 = 10.74$, $p = .001$).

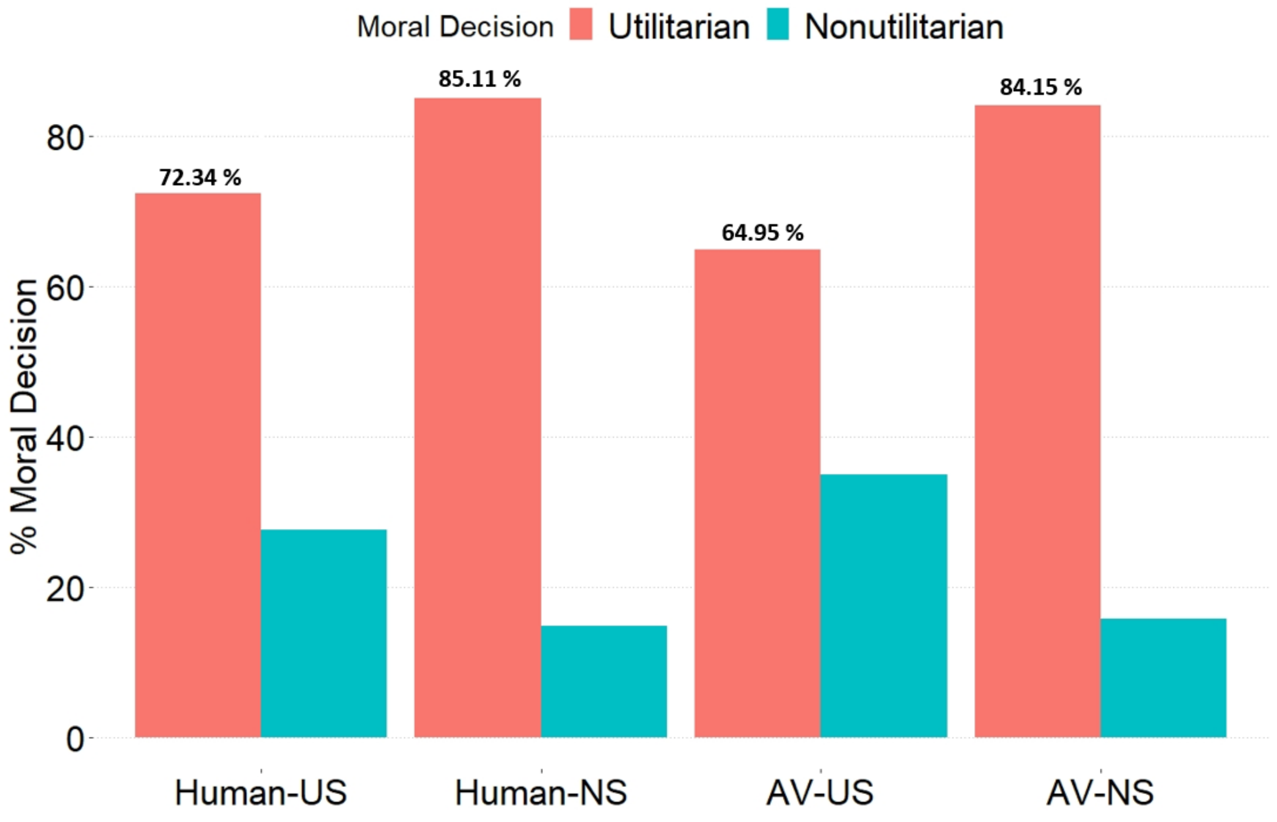


Figure 14: bar chart of moral decision percentage frequencies, divided by type of sacrifice framing (US: sacrifice framed in the utilitarian outcome, NS: sacrifice framed in the nonutilitarian option), and dilemma category in columns, and by moral decision by color (utilitarian in red, nonutilitarian in light blue)

Subsequently, the focus moved on to the evaluation of moral acceptability of the two proposed options (M_3: nonutilitarian, M_4: nonutilitarian). Model comparison procedure suggested to fit M_3 only considering the Dilemma Category as fixed effect, while M_4 was fitted assuming Dilemma Category, Moral Decision (and their interaction), Sacrifice Framing, Gender, Experimental Order and the evaluated morality of the nonutilitarian option as fixed effects. Overall, both the options were evaluated as poorly moral, consistently with study 1 (Table 7 and Table 8). No significant effect of the dilemma category was detected in the evaluation of the nonutilitarian (M_3: $\chi^2{}_1 = 3.00$, $p = .08$) and the utilitarian outcome (M_4: $\chi^2{}_1 = 2.82$, $p = .09$). As expected, the endorsement of the utilitarian maneuver corresponded with a higher evaluation of its moral acceptancy (M_4: $\chi^2{}_1 = 162.19$, $p < .001$). Interestingly, when the utilitarian option assumed the self-sacrificial act, the acceptability of this option was higher than the correspondent self-protective behavior (M_4: $\chi^2{}_1 = 269.09$, $p < .001$, Figure 15). The acceptability of the utilitarian option was negatively affected by the experimental time (M_4: $\chi^2{}_{11} = 43.63$, $p < .001$), and gender differences observed in Study 1 was confirmed, with lower evaluation of morality of the utilitarian option reported by women (M_4: $\chi^2{}_1 = 10.46$, $p = .001$).

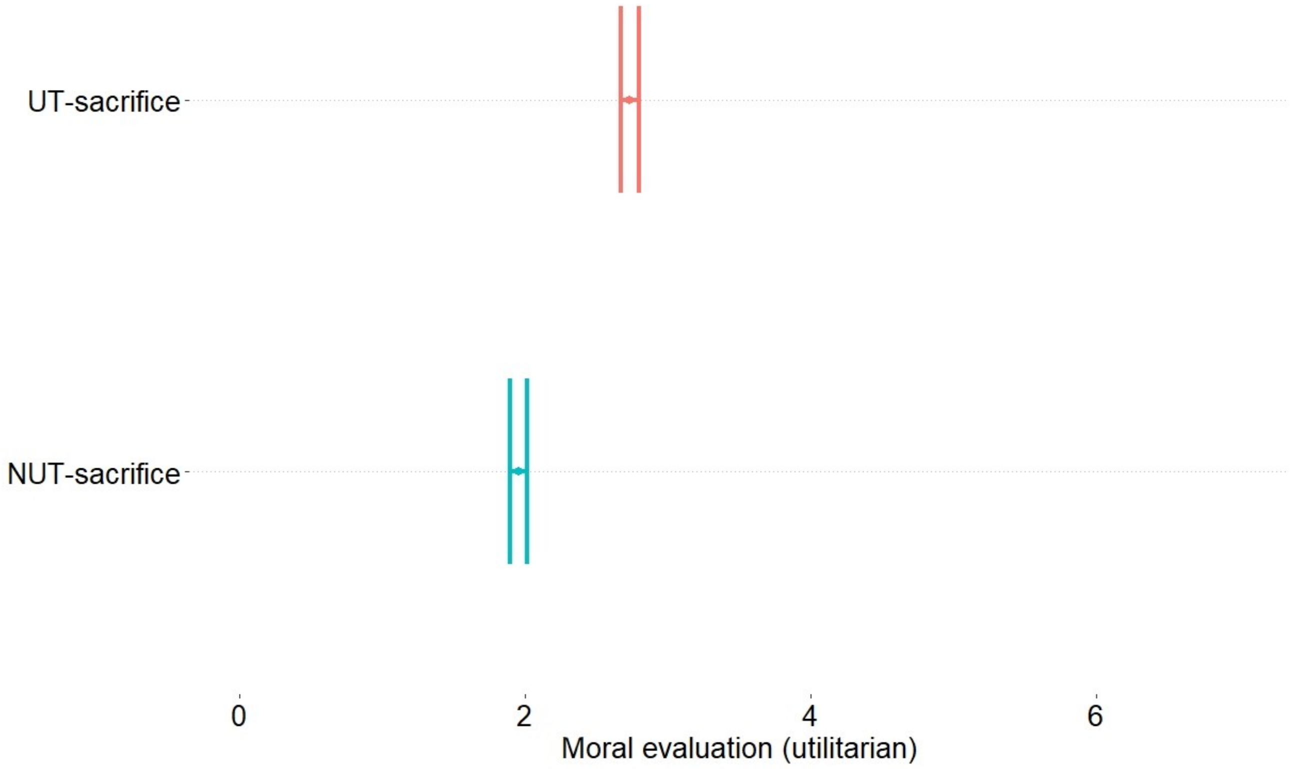


Figure 15: Error bars plot representing means and standard errors of the moral acceptability of the utilitarian driving behavior, divided by sacrifice framing (UT: Utilitarian Sacrifice Framing, in red; NUT: Nonutilitarian Sacrifice Framing, in light blue).

At this point, the focus moved onto the self-referred intensity of the four moral emotions considered. Two linear mixed models were fitted setting the two other-referred moral emotions as dependent variables (M_5: anger, M_6: disgust). In both cases, the final models resulted from the model comparison procedure considered Dilemma Category, Moral Decision, Sacrifice Framing and Experimental Order as fixed effects, as well as Moral Decision in interaction with Dilemma Category and with Sacrifice Framing. No difference weas observed between driving styles in terms of anger (M_5: $\chi^2_1 = 1.09$, $p = .29$), albeit higher levels of disgust were assessed in the case of AV dilemmas (M_6: $\chi^2_1 = 7.31$, $p = .007$). The moral decision and the self-sacrifice framing did not have an effect on other-condemning moral emotions, but a significant interaction between sacrifice framing and moral decision was detected in terms of anger ($\chi^2_1 = 13.65$, $p < .001$), exhibiting a greater intensity of this moral emotion when the decision corresponded to self-sacrifice (Figure 16). Both anger and disgust intensity appeared to increase during the experimental procedure (M_5: $\chi^2_{11} = 37.20$, $p < .001$, M_6: $\chi^2_{11} = 11.66$, $p < .001$).

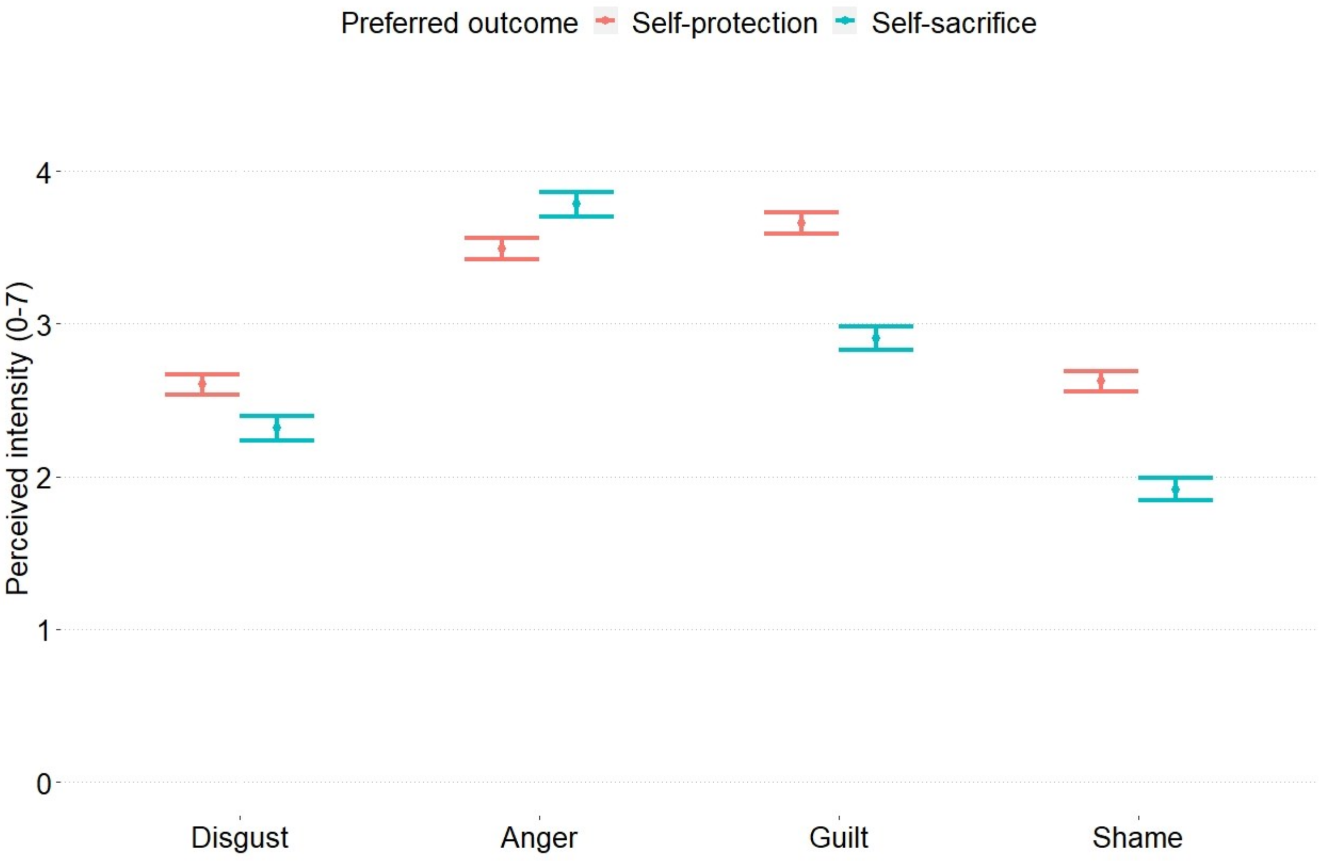


Figure 16: Error bars plot representing means and standard errors of the intensity of other- and self-referred emotions (Anger and Disgust, Shame and Guilt), divided by preferred outcome (Self-protection, Self-sacrifice), despite the sacrifice framing.

Finally, two additional linear mixed models were fitted for the investigation of the two self-conscious moral emotions, shame (M_7) and guilt (M_8). The fitted models took in consideration the same factors assumed in the other-referred models M_5 and M_6, plus the Gender effect. Shame and guilt appeared to follow the same trends: higher intensity of the two self-referred emotions were detected in human-driving scenarios (M_7: $\chi^2{}_1 = 23.45$, $p < .001$, M_8: $\chi^2{}_1 = 44.86$, $p < .001$), in case of nonutilitarian decisions (M_7: $\chi^2{}_1 = 52.98$, $p < .001$, M_8: $\chi^2{}_1 = 36.12$, $p < .001$), and in nonutilitarian-sacrifice framed dilemmas (M_7: $\chi^2{}_1 = 106.93$, $p < .001$, M_8: $\chi^2{}_1 = 113.77$, $p < .001$). Interestingly, the moral decision had a significant interaction with both levels of the sacrifice framing factor (USF, NSF), greater intensities of shame and guilt were observed when pursuing self-protection, when compared to the endorsement of the self-sacrificial outcome (M_7: $\chi^2{}_1 = 92.09$, $p < .001$, M_8: $\chi^2{}_1 = 104.83$, $p < .001$). Gender differences were also detected, with higher scores of shames and guilt for women (M_7: $\chi^2{}_1 = 15.89$, $p < .001$, M_8: $\chi^2{}_1 = 22.04$, $p < .001$).

Table 7: Mean and Standard Deviation of the dependent variables considered, divided by driving style (human, AV) and sacrifice framing (NS: Nonutilitarian Sacrifice Framing, US: Utilitarian Sacrifice Framing).

	Human-Driving	**AV Driving**	**US-framing**	**NS-framing**
Moral decision: utilitarian	78.71 %	74.59 %	68.65 %	84.63 %
Moral evaluation: utilitarian	2.41 (2.01)	2.28 (2.00)	2.74 (2.04)	1.96 (1.90)
Moral evaluation: Nonutilitarian	1.32 (1.62)	1.45 (1.79)	1.43 (1.74)	1.35 (1.66)
Shame (self-referred)	2.43 (2.37)	2.22 (2.28)	2.09 (2.25)	2.56 (2.38)
Guilt (self-referred)	3.51 (2.47)	3.18 (2.48)	3.08 (2.45)	3.61 (2.48)
Anger (other-referred)	3.59 (2.45)	3.64 (2.50)	3.66 (2.48)	3.57 (2.47)
Disgust (other-referred)	2.41 (2.41)	2.55 (2.50)	2.44 (2.45)	2.52 (2.46)

Table 8: Mean and Standard Deviation of the dependent variables considered, divided by driving style (human, AV) and sacrifice framing (NS: Nonutilitarian Sacrifice Framing, US: Utilitarian Sacrifice Framing).

	Human-US	**Human-NS**	**AV-US**	**AV-NS**
Moral decision: utilitarian	72.34 %	85.11 %	64.95 %	84.15 %
Moral evaluation: utilitarian	2.85 (2.05)	1.98 (1.88)	2.63 (2.02)	1.94 (1.92)
Moral evaluation: Nonutilitarian	1.33 (1.54)	1.32 (1.68)	1.52 (1.92)	1.38 (1.64)
Shame (self-referred)	2.16 (2.30)	2.71 (2.28)	2.02 (2.20)	2.42 (2.34)
Guilt (self-referred)	3.20 (2.44)	3.82 (2.46)	2.95 (2.46)	3.40 (2.48)
Anger (other-referred)	3.64 (2.51)	3.55 (2.46)	3.68 (2.51)	3.59 (2.49)
Disgust (other-referred)	2.34 (2.39)	2.48 (2.43)	2.55 (2.51)	2.56 (2.50)

4.5 Discussion

In the conduction of Study 2, we pursued multiple objectives. First of all, we aimed to detect potential differences in terms of moral judgment between human- and autonomous-driving sacrificial dilemmas. Assuming that this is the first experimental attempt in comparing driving moral dilemmas with different levels of automation, 12 incidental dilemmas were appositely structured, depicting the moral agent as a potential victim of the two proposed outcomes. The dilemma set was then tested through an online survey, and a number of useful information were collected coherently with our hypothesis. As expected, a strong endorsement of the utilitarian resolution was observed both in human-driving and AV dilemmas, and the dilemma categories did not differ neither in terms of decisional times, nor in the evaluation of the proposed moral options. These results are in line with Study 1 and with the structure-based interpretation of moral dilemmas (Schein, 2020): since both the human-driving and the AV dilemmas were structured as incidental, the moral judgment process appear to be independent from the moral contextualization. Overall, we may assume that the investigation of moral judgment towards the driving activity is not sensitive to the level of automation, and morality seems shaped more consistently by the structure of the problem than its specific context. Nonetheless, we had the chance to observe a significant advantage of human-driving dilemmas in promoting the utilitarian resolution. This is also coherent with Study 1, confirming the advantage of plausible storylines in boosting utilitarian resolutions, and assuming the detrimental effect of unlikely situations on the endorsement of this moral code (Körner et al., 2019). In fact, it is plausible that the issues concerning the development of the autonomous transportation technology are still not considered as a daily-matter problem from the Italian population, reducing the perceived plausibility of AV storylines (Guo et al., 2021) in comparison with traditional *hands-on-the-wheel* vehicles. This advantage is also observable at the decision time level, confirming the (unpredicted) trend observed in Study 1. The utilitarian resolution was selected faster than the nonutilitarian one, despite the indications from the DPT (Greene et al., 2001) and the randomization of the two alternatives performed in the present study. We believe that the online experimental modality may have the potential to reduce the activation of the primary emotional response, limiting the emotional conflict between the options. Further investigation may be helpful for disclose this potential limitation of the online survey tool.

The investigation of social perception and moral judgment of autonomous transportation has strongly benefited from the sacrificial dilemma as a flexible experimental tool. Nonetheless, in the rationale of the present study (section 4.1) we described a fundamental difference between the typical sacrificial dilemma structure and its autonomous driving counterpart, standing in the moral framing of the self-sacrificial behavior. In fact, imagining as the passenger of an AV, the moral agent needs to choose between protect the higher number of pedestrians or protect themself, since the self-sacrificial behavior is framed in the nonutilitarian option (i.e., "*I die, but many survive"*). Oppositely, when a traditional nondriving dilemma is presented, the moral agent has the possibility to endorse both the utilitarian action and the self-protective one (i.e., "*I live, and many survive"*). This seems a much easier opportunity to grasp for the moral agent since they can pursue at the same time both the collective and the personal goal. Considering so, we decided to investigate the potential effect of this structural difference on moral judgment. As predicted, a clear decrease in the endorsement of the utilitarian outcome was detected when it included the need for the moral agent to sacrifice themself. This result seems intuitive, assuming the natural tendency of human-beings to self-protection (Petrinovich et al., 1999), but discloses the importance to take in consideration the self-sacrificial framing in the investigation of morality (Huebner and Hauser, 2011; Thomson, 2008). It is commonly believed that moral agent places greater value on their own life then in third-party's life (Huebner and Hauser, 2011; Lotto et al., 2014; Moore et al., 2008), but some studies claim for the suppression of selfishness in moral reasoning (Haidt, 2007), also placing greater value - and honor - to self-sacrificial acts when compared to self-protective behaviors (Sachdeva et al., 2015). This seems consistent with our results: people prefer to endorse the utilitarian resolution, but they evaluated this moral behavior as less morally acceptable (Figure 14). An interesting descriptive trend is observed by the reduced endorsement of the utilitarian option in AV dilemmas when it resulted in a self-sacrificial act (Table 8). It means that, when on board of an AV, people would be less willing to accept their own sacrifice for the greater goal, when compared to the *hands-on-the-wheel* driving situation. In our view, this tendency deserves further investigation, focusing on the perception of agency and responsibility in AV users (Awad et al., 2018b; Gill, 2021; Haboucha et al., 2017).

Considering the emerged difficulty in evaluating the primary emotional activation at the time of moral judgment, we decided to follow the investigation of moral reactions in response to a specific moral decision process, focusing on four different moral emotions (anger, disgust, shame, and guilt). Additionally, we assumed the possibility to observe differences in moral reactions between our two experimental manipulations, namely the level of automation in the dilemmas and the self-sacrifice framing. Looking at the dilemma category, our hypothesis weas partially supported by data. In fact, a higher activation of shame and guilt - as self-conscious moral emotions (i.e., self-referred) - was reported in response to human-driving dilemmas, in which the moral agent was in charge of the driving operations. Particularly for guilt, this result was predictable since this moral emotion seems elicited by contingent personal transgression in the moral realm (Sabini and Silver, 1997; Smith et al., 2002), as for example a moral transgressive behavior while driving. Results on other-condemning moral emotions (i.e., other-referred, anger and disgust) partially confirmed our hypothesis, since only disgust scores were reported higher in the AV case then in its human-driving counterpart. Nonetheless, the inconsistency between anger and disgust was expectable, as they respond to different moral cues (Gutierrez and Giner-Sorolla, 2007; Russell and Giner-Sorolla, 2011) and anger is highly related to intentional harm, which is something that seems still complex to attribute to AI and infer from their behaviors (Fritz et al., 2020; Liu, 2021). As predicted, also the sacrifice framing had a role in the perceived intensity of the moral emotions, with peculiar differences between self-protective and self-sacrificial choices. In fact, a higher intensity of self-referred moral emotions (shame and guilt) was reported in self-protective decisions (i.e., *"I live, and many survive"*), while only one other-referred emotion (anger) was significantly higher in the endorsement of self-sacrificial decisions. This result seems reasonable since our focus was on moral emotions with a negative characterization, which were activated separately and differently in consideration of the agent performing the driving maneuver (the self in the human-driving case, the AI in the AV case). Furthermore, results are consistent with the previous literature, since immoral events or injustices toward the self can elicit anger, (Haidt, 2003b; Hutcherson and Gross, 2011). Overall, we can deduce that people are most likely to pursue the utilitarian moral code, but this behavior is perceived as more shameful and blameworthy than self-sacrifice for the grater goal (Sachdeva et al., 2015).

Finally, is important to point out that the evidence in Study 2 has been collected through the administration of moral dilemmas involving an additional character, which shared the moral agent's faith in both human and autonomous driving dilemmas. Traditional sacrificial dilemmas were structured with a scarce attention towards the numerical ratio between survived and sacrificed characters (e.g., Greene et al., 2001; Moore et al., 2008), but few studies highlighted the dismissible role of numerosity in shaping moral judgment, especially in the AV case (Awad et al., 2020; Bonnefon et al., 2016, Faulhaber et al., 2019). In the present application, we had to add this additional information in the scenarios in order to maintain constant the numerical ratio between survived and sacrificed characters throughout the levels of sacrifice framing. Future applications may play an important role in disclosing new information on the role of numerosity in sacrificial dilemmas.

In conclusion, Study 2 allows us to perform a first comparison between human-driving and autonomous driving sacrificial dilemmas, in terms of decision times, moral decisions and the reported activation of four moral emotions. This comparison allowed us to bring new evidence in favor of the structure-based interpretation of moral dilemmas, since no difference were observed between levels of automation. Nevertheless, the advantage of lifelike scenarios was confirmed in boosting the likelihood of the utilitarian resolution. Additionally, important evidence weas collected on the moral framing of the self-sacrificial option, which was an overlooked factor in the development of sacrificial dilemmas. A series of useful information were then also collected in terms of self- and other referred moral emotions, which expanded in the emotional framework the evidence highlighted by Bonnefon et al., (2016) on the incongruency between moral decision and moral evaluation.

4.6 Halfway point: a quick overview of the collected evidence

Evidences collected from Study 1 and Study 2 allow us to have a wider picture on the application of sacrificial dilemmas in the context of autonomous transportation. Results from these experimental investigations highlight fundamental similarities between sacrificial self-involvement dilemmas regardless of their specific contextualization, supporting the structure-based hypothesis of moral judgment in the evaluation of critical moral scenarios (Schein, 2020). This confirmation is consistent with the DPT (Greene et al., 2001; 2004; 2008), as the main theory that combines the cognitive and the emotional perspectives for the characterization of moral decision-making: despite the dilemma storylines, the competition between automatic and more conscious reaction to these moral problems are mainly shaped by internal and structural features of the scenarios. The present interpretation is supported by the striking endorsement of the utilitarian resolution throughout the investigated dilemma categories (non-driving, human-driving, autonomous-driving), by non-different decision times between levels off driving automation, and by a scarce distinction between dilemma-contexts in terms of emotional activation. Nonetheless, a series of interesting hints agreed on the possibility that driving dilemmas may be assumed as a specific case of sacrificial moral scenarios, also leveraging on their level of realism (Bauman et al., 2014; Gold et al., 2014; Watkins, 2020). In fact, the utilitarian support emerged clearly in human-driving scenarios, more markedly than in AV dilemmas and also quicker than in non-driving ones. The application of moral problems to daily and imaginable challenges appears to enhance a more utilitarian and practical moral reasoning. This evidence needs to be taken under serious consideration during the employment of traditional but somewhat unlikely moral storylines (Moore et al., 2008). Additionally, a series of features appeared to play a sizeable role in shaping moral judgment towards

AVs, such as the framing of the utilitarian sacrifice between moral options and the emotional reaction to opposite human-machine interactions in traditional nonautonomous (human as the driver) and in futuristic autonomous vehicles (human as a passenger). We considered these insights as useful information in the hands of moral psychologists and AI ethicists for the development of controlled experimental material in the investigation of AI morality. Considering that, in Study 3 and Study 4 we decided to focus on a number of individual and structural features widely discussed in the development of AV dilemmas, which may have an impact on moral judgment towards AVs' behaviors. In this sense, the following studies will specifically focus on AV dilemmas.

Chapter 5

Study 3: The role of time constraints and prosocial orientation in the AV dilemma

5.1 The rationale of the study

In the last decades, the relationship between time and decision-making has been the subject of numerous studies (Edland and Svenson, 1993; Prelec and Loewenstein, 1991), crossing various experimental applications and focusing on several topics in the broad context of decisional sciences (e.g., Huber and Kunz, 2007; Zavala et al., 2017). In everyday life, people need to make a large number of decisions – in both relatively easy and ambiguous situations - that has to fit with the time they have available. Time can affect decision-making in different ways, and research on this topic focuses mainly on the available time while making a decision, and on how the decision rules change with the passage of time (Ariely and Zakay, 2001). Time can serve as a medium during the decision process (e.g., Pleskac and Townsend, 2010; Ratcliff and McKoon, 2008), and past decisions may have a role on later decisions, since the attributes of a certain option gain weight and became more stable (Brandts and Charness, 2000; Hoeffler and Ariely, 1999), assuming the form of a strategy method. Contextually, time can be perceived as a contextual factor when the amount at the disposal of the individual is forcedly manipulated. Time constraint is defined as an externally imposed deadline for the assumption of a particular decision (Ordóñez et al., 2015). This definition is typically used interchangeably with time pressure, which is however the subjective feeling of having not enough time to finalize a certain task (Chu and Spires, 2001). The impact of time constraints on decision processes is widely discussed (e.g., Ahituv et al., 1998; Kocher and Sutter, 2006), and when the superimposed time is perceived as costly (i.e., highly limiting), decision-makers will switch to their individual-simpler and most cognitively-effective strategy, considering only the key variables of the issue (Rieskamp and Hoffrage, 2008). This interpretation is in line with the heuristic systematic model (Chaiken, 1989) and Kahneman's two-way modes of thinking

(2011): reduced availability of time enhances the likelihood of judgments based solely on the automatic System 1 process, since the imposed deadline reduces the possibility to engage the more deliberative System 2 reasoning. In the context of ethical decisions, Moberg (2000) suggested how time constraint in the form of time delay can increase the occurrence of unethical behaviors, while a sufficient amount of time enhances the probability of ethical behaviors (Shalvi et al., 2012). Interestingly, time constraint has been found to increase the use of moral decision heuristics, but in the direction of duty-oriented moral codes (e.g., deontologism; Björklund, 2003). This result was also supported by Suter and Hertwig (2011), which found that faster responses under time constraint led to more deontological/nonutilitarian responses. This is in line with the advantage of the deontological moral code depicted in the DPT, in the context of fast and uncontrolled moral decisions (Greene et al., 2001; 2004), and with evidence collected by Frank et al., (2019) on the more frequent employment of the utilitarian moral doctrine in a deliberate decision-making process. Nonetheless, results on the role of time constraints on moral judgment are quite controversial. Tinghög et al. (2016) applied time pressure and cognitive load to investigate the role of intuition on moral judgements, but no evidence in support of a particular role of time pressure in shaping moral judgment towards a more deontological or utilitarian support. This is consistent with Haidt's Social Intuitionist Model (2011), but in contrast with Greene's theory. Additionally, Rosas and Aguilar-Pardo (2020) described a completely opposite trend, suggesting how time pressure is able to relatively increase the endorsement of utilitarian resolutions. This contrasting evidence were collected with the use of both incidental and instrumental non-driving sacrificial moral dilemmas from validated sets (Greene et al., 2001; Hauser et al., 2009; Moore et al., 2008).

Contextually, intuitive and deliberative reasoning were also deepened in the context of cooperative behaviors. An important theory was developed by Rand and defined as the Social Heuristic Hypothesis (SHH). Preliminarily, David Rand et al. (2012) claimed that intuitive behaviors are developed in everyday life, mostly (i) when cooperation results in a personal advantage during repeated interactions,

(ii) when reputation is at risk, and (iii) when bad or good behavior can result in a reward or a sanction, respectively. Subsequently, the SHH theory was experimentally tested (Rand et al., 2014), revealing a causal relationship between intuitive reasoning and cooperative behaviors in Public Goods Games, as time constraint was capable to foster action in favor of the community rather than compliance to specific norms (Cone and Rand, 2014). This effect resulted clear in one-shot anonymous interaction, where the selfish resolution was the optimal possible outcome. Nevertheless, an intense discussion arose since the definition of the SHH. Tinghög et al. (2013) reconsidered Rand's conclusions by detecting a series of methodological problem (e.g., the exclusion of individuals who fail to respond on time, the implicit assignment to experimental conditions, the presence of an illustrative example) and retesting the paradigm on a new public goods game study. No significant effect of time pressure was found, casting doubt on Rand et al. interpretation. Subsequently, Goeschl and Lohse (2018) fanned the discussion, claiming how a reduction of available time selectively reduces the contribution of cooperative agents, in line with the idea that deliberation – rather than intuition - drives cooperation.

In Study 3, we aimed to combine the evidence collected on the role of time on moral judgment and cooperative behaviors in the framework of moral AV behaviors, assuming the cooperative behavior as a prosocial action directed towards the collectivity for the achievement of a greater utilitarian goal. In this sense, we manipulated the time at the disposal of the moral agent in response to three different AV dilemmas, framed as sacrificial and incidental. Importantly, we decided to control moral decisions on the basis of the self-reported prosocial orientation. To this aim we tested the effect of time superimposition consistently with the previous literature (Goeschl and Lohse, 2018; Rand et al., 2012, 2014; Tinghög et al., 2013), administering the slider version of the Social Value Orientation scale (Murphy et al., 2011). Additional information were also collected on the five moral domains (Graham et al., 2009; 2011) and on the participants' social and professional lives. Finally, we decided to investigate the potential role of moral consistency on the moral evaluations of the proposed AV behaviors and the willingness to buy the correspondent AV. Sequential behavior paradigms, such as reiterated moral dilemma studies, aims at collecting individual responses through the presentation of a defined number of experimental stimuli of the same category (Mullen and Bonin, 2016). Performing this task, people can show more or less consistency with the first decision (e.g., Conway and Peetz, 2012). Traditionally, moral consistency is investigated performing abstract thinking and transcending actual events (e.g., Trope and Lieberman, 2010), which seems a similar form of reasoning to the one activated during the resolution of moral dilemmas.

5.2 Hypothesis

A series of hypotheses were advanced in the present study, considering the thoughts described in section 1.2 and in the rationale of the present chapter:

- Coherently with Suter and Hertwig (2011), Frank et al., (2019), Goeschl and Lohse (2018), and the Dual Process Theory (Greene et al., 2001; 2004), a lower endorsement of the utilitarian (and cooperative) behaviors was expected under time constraint, so when moral agents are forced to decide in a limited period of time, when compared to decisions with enforced delay (i.e., when moral agents are forced to wait).
- In accordance with the Social Heuristic Hypothesis (Rand et al., 2012; 2014), the endorsement of the utilitarian resolution was expected to grow coherently with individual prosocial orientation, especially under time constraint.
- Additionally, coherent results were expected between moral consistency profiles (fully utilitarian, fully nonutilitarian, non-consistent or switcher), decision times and evaluations of the proposed moral option. For example, fully utilitarian moral agents would show lower decision times (coherently to Study 1 and Study 2) and moral evaluation in favor of the AV utilitarian resolution. Willingness to buy was expected to be consistent with

this trend, with more intention to buy AVs behaving coherently with the moral consistency profile.

5.3 Method

5.3.1 Participants

An a-priori power analysis has been computed on G-power statistical software (Faul and Erdfelder, 1992) before starting the data collection, assuming a medium effect size (Cohen's d = 0.20) and a correlation of 0.50 among repeated measures, with an alpha error probability of 0.05 and 0.90 power. The system suggested a total of 204 participants, and 207 participants were recruited for the experiment. The final sample counted a total number of 206 participants: a subject was excluded as he failed to correctly answer to a check question during the completion of the experiment (see Procedure). Females accounted for 50% of the final sample (103 females). Overall, the mean age was 27.26 (SD = 8.75, range = 18–64), the mean schooling age was 16.6 years (SD = 2.7), and 42.23% of the sample were enrolled in university courses (n = 187), with 35.43% working as employees (n = 73). Most participants (90.15%) had held driver licenses (n = 183), and 77.83% of them already heard about autonomous transportation (n = 158). 62.07% were involved in social activities (e.g., volunteers, representatives, members of associations; n = 126). The study was administered through Qualtrics software (Qualtrics, Provo, UT) and approved by the local ethics committee (ID No.: 3480). Before their participation, each participant gave formal written consent, which was voluntary and unremunerated.

5.3.2 Materials

AV dilemmas

Three self-involvement, sacrificial and incidental moral dilemmas were employed for this study. The stimuli presented three critical on-road situations, in which the moral agent was depicted as the only passenger of a completely autonomous vehicle (SAE's level 5). Facing the moral dilemma, the individual had to indicate which of two proposed behaviors they considered to be the more morally acceptable for the AV to apply. In one case, the AV would have performed a nonutilitarian and passenger-protective maneuver, proceeding straight and running over five characters. In the opposite case, the AV would have opted for a utilitarian but passenger-sacrificial behavior, steering to the side of the road. This maneuver would have allowed for the protection of the higher number of characters but causing the sacrifice of the moral agent. The dilemmas employed in this study were slightly reworked from a selection of dilemmas from Study 2, removing the second passenger from the storytelling and modifying the numerical ratio between sacrificial and survived characters to 1:5. The dilemma set is reported in Table 9 and in the Appendix of the present thesis.

Table 9: The autonomous driving dilemma set administered in Study 3 (text translated from Italian).

AV Dilemma	**Scenario**	**Resolution**
Truck *AV dilemma*	You are the passenger of a fully autonomous vehicle, driving on a tree-lined avenue. A truck is proceeding in front of you, which is now slowing down for no apparent reason. A dotted line separates the road lanes, so you decide to overtake it. During the overtaking, five cyclists suddenly cross the road, appearing from behind the truck. The autonomous vehicle did not perceive them in time, and it has no time to brake.	The autonomous vehicle has two alternatives: • Proceed straight, running over the five cyclists, who will die. • Suddenly steer to the left. The five runners will not be hurt, but the autonomous vehicle will crash against a big tree, where you will die.
Fog *AV dilemma*	You are the passenger in a fully autonomous taxi vehicle. As in the last few nights, a thick fog has descended on your city, and the visibility is strongly compromised. You notice some vehicles parked on the right side of the road. Suddenly, five pedestrians appear from the right, now standing in the middle of the road. The autonomous vehicle did not perceive them in time, and it has no time to brake.	The autonomous vehicle has two alternatives: • Proceed straight, running over the five pedestrians, who will die. • Suddenly steer to the left. The five pedestrians will be unhurt, but the autonomous taxi will crash against a building, where you will die.

Workers AV dilemma	You are the passenger in a fully autonomous vehicle, driving on a two-lane highway. It is early morning, and yours is the sole vehicle on the road. Suddenly, the autonomous vehicle notices a road sign and five workers a few meters ahead in the middle of the road dealing with road maintenance work. The vehicle begins to slow down, and you realize that the brakes are not working.	The autonomous vehicle has two alternatives: • Proceed straight, running over the five workers, who will die. • Suddenly steer to the left. The five workers will not be hurt, but the autonomous vehicle will crash against a streetlamp, where you will die.

*Notes: * The two options were randomized across dilemmas*

Moral Foundation Questionnaire

In the present study, the Italian validated 30-item version of the Moral Foundation Questionnaire (MFQ; Bobbio et al., 2011) was administered, containing 15 items on the perceived relevance of a series of moral considerations (1 = Not at all relevant to 6 = extremely relevant), and further 15 items on the agreement with a series of moral judgments (1 = strongly disagree to 6 = strongly agree). The Moral Foundation Questionnaire (MFQ, Graham et al., 2011; Haidt and Joseph, 2007) was developed to assess the extent to which individuals' moral judgment is shaped by the activation of five moral domains (i.e., foundations): Harm/Care (HC), Fairness/Justice (FJ), Ingroup/Loyalty (IL), Authority/Respect (AR), and Purity/Sanctity (PS). Each of these foundations characterizes particular virtues and behavior, and is the result of human evolution (Graham et al., 2011; 2013). Among these five categories, harm/care appears to be directly involved in the context of sacrificial dilemmas, since it accounts for the ability to be empathetic and compassionate. Djeriouat and Trémolière (2014) detected a negative correlation between harm/care and utilitarianism.

Social Value Orientation (SVO)

The Slider Measure (SLM, Murphy et al., 2011) of the Social Value Orientation (SVO) was one of the measures deployed for the investigation of prosocial orientation as individual trait (Murphy and Ackermann, 2014; Murphy et al., 2011; Thielmann et al, 2020). SVO is defined by the amount of resource an individual assign to themself and to another person in a situation of interdependence (Murphy and Ackermann, 2014). The motivation that led individuals to a different allocation of resources defines different social profiles: individualistic (maximize their own gains), competitive (maximize the differences with the other person), inequality adverse (minimize the difference with the other person), or prosocial (maximize the joint payoff). SVO seems related to different aspects of interpersonal decision-making, such as resource allocation in social dilemmas (e.g., Roch et al., 2000) and propensity to cooperation (Zeelenberg et al., 2008). Several measures have been tested in the literature, such as the Triple Dominance Measure (van Lange et al., 1997), the Ring Measure (Liebrand and McClintock, 1988), but the SLM appeared to be easier to administer, more reliable and, most of all, manageable as a continuous measure. Here, respondents were requested to choose how to allocate an available resource between the self and another unspecified person, over a continuum of joint payoffs. In the present application, we administered only the six main decision items of the SLM (Figure 17), derived from goniometric representation on joint payoff allocation, described in greater details by Murphy et al., (2011) and in the analysis section (5.3.4).

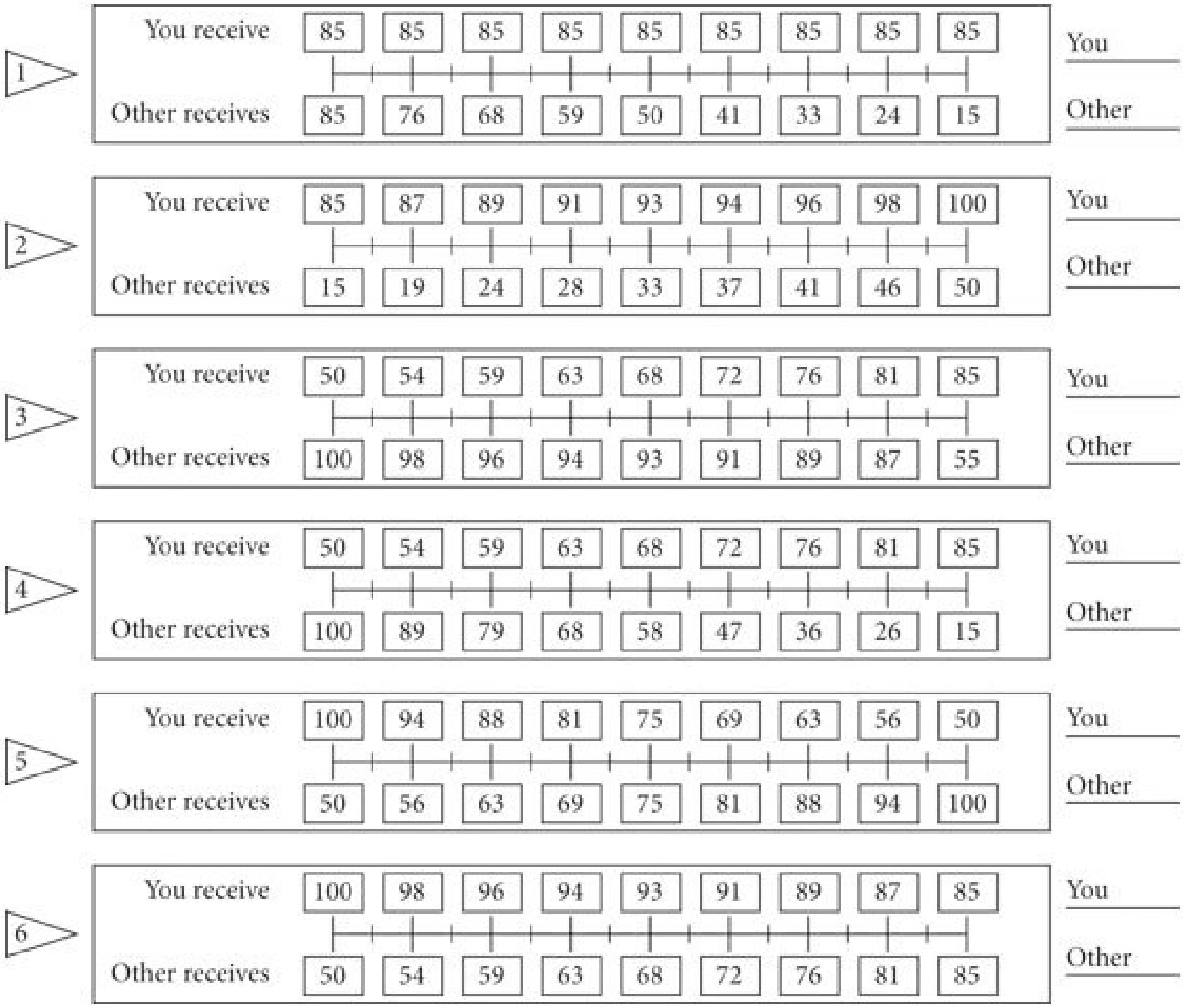

1
You receive 85 85 85 85 85 85 85 85 85
Other receives 85 76 68 59 50 41 33 24 15
You
Other
2
You receive 85 87 89 91 93 94 96 98 100
Other receives 15 19 24 28 33 37 41 46 50
You
Other
3
You receive 50 54 59 63 68 72 76 81 85
Other receives 100 98 96 94 93 91 89 87 55
You
Other
4
You receive 50 54 59 63 68 72 76 81 85
Other receives 100 89 79 68 58 47 36 26 15
You
Other
5
You receive 100 94 88 81 75 69 63 56 50
Other receives 50 56 63 69 75 81 88 94 100
You
Other
6
You receive 100 98 96 94 93 91 89 87 85
Other receives 50 54 59 63 68 72 76 81 85
You
Other

Figure 17: The six items from the SVO' slider measure (SLM).

Following the scoring procedure (see analysis section 5.3.4), individuals are assigned to one of four social categories: altruistic, cooperative, individualistic, competitive. Nonetheless, Bakker and Dijkstra (2021) underlined a strong unbalance in the SLM clustering, in favor of particular categories (cooperative and individualistic), and so recommending using the SLM as the most suitable SVO measure, although avoiding its original classification in favor of the continuous data (see analysis section 5.3.4).

Social interactions in everyday life

Previously, we defined how the SHH posit that cooperation is enhanced by time pressure as a heuristic mechanism. Describing the results obtained by their experimental investigations, Rand et al. (2012; 2014) claimed that people tend to develop intuitions in everyday contexts, shaping cooperation mainly when it is advantageous during reiterated interactions that may result in sanctions or rewards. To control for these factors, a series of information were collected on participants' social and work life, such as (i) perform social activities, (ii) have iterated social interactions in the workplace, (iii) incur in competition with peers of colleagues, (iv) risk sanctions when poor results/behaviors occurred, and (v) rely on rewards when good results/behaviors occurred.

5.3.3 Experimental procedure

Study 3 was programmed and distributed via Qualtrics software. The anonymous link provided by the program was distributed via social network and institutional communication channels on the basis of a snowball nonprobability sampling technique (Goodman, 1961; Parker et al., 2019). The data collection was performed from October 27th, 2021, and January 14th, 2022. All the participants were required not to perform the survey through smartphones or tablets, but only using laptops, in order to avoid problems of data comparability between multiple devices (Krebs and Höhne, 2021). The experimental procedure is graphically described in Figure 18.

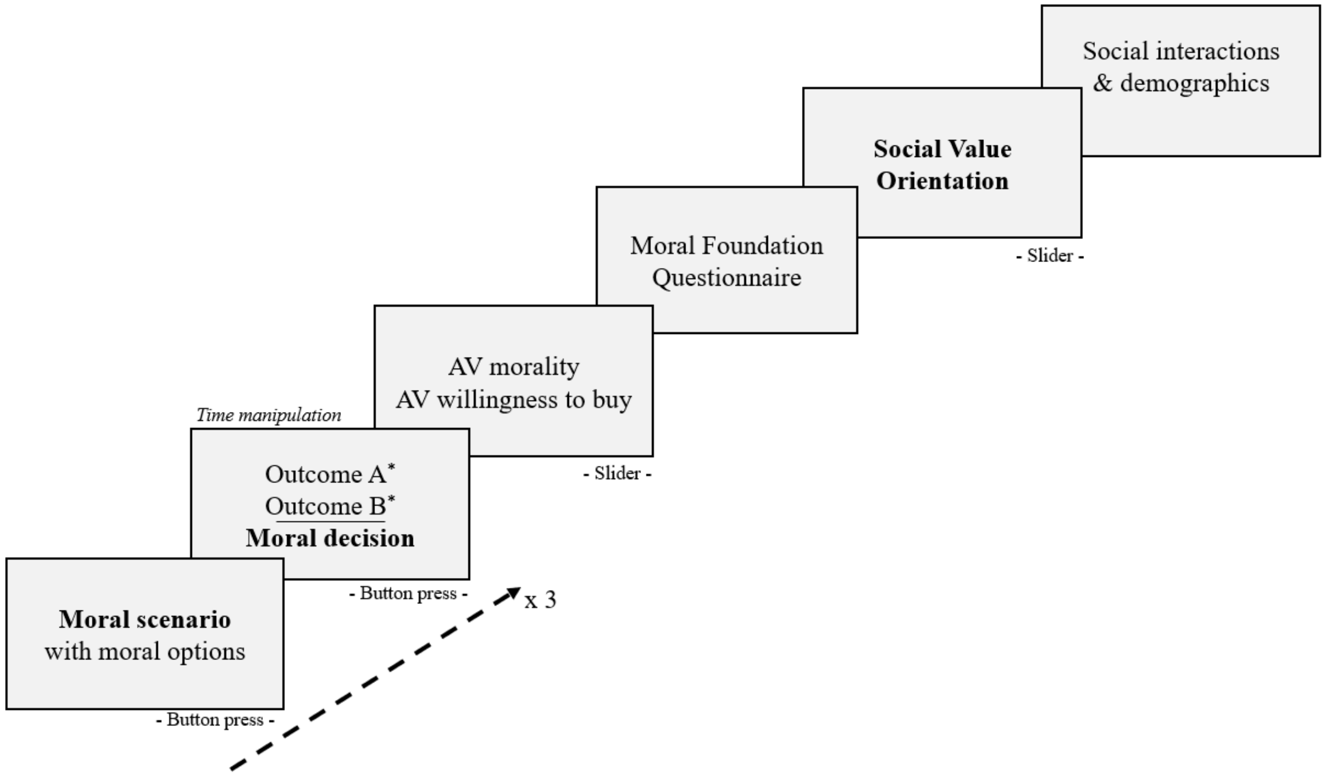


Figure 18: The experimental procedure of Study 3.

The mean completion time of the experimental procedure was 18.19 mins (SD = 14.19 mins). Before the beginning of any experimental activity, the participants were requested to read and fill out an informed consent about their participation and data protection regulation. First of all, each participant was randomly assigned to one of the three experimental conditions, controlling for gender balance. A detailed explanation of the dilemma section was then presented, coherently with the assigned condition (Table 10). All the participants had a maximum of 75 seconds to read each moral dilemma (scenario and moral outcomes), which was presented at the center of the screen, in black type (font Arial, size 10) against white background. At that point, the two moral outcomes were left on the screen with the correspondent selection keys.

Table 10: The distribution of the sample between experimental conditions, controlled by gender.

	Female	Male	*N*
Time constraint (max 8 secs)	35	34	*69*
Time delay (min 60 secs)	34	33	*67*
No time limitation (control)	34	36	*70*
N	*103*	*103*	**206**

Participants in the time constraint condition were asked to respond as quickly as possible having a maximum of 8 seconds at their disposal. This amount of time was defined as an intermediate value between research endorsing the SHH (10 seconds, Rand et al., 2012; 2014) and research against this theory (7 seconds; Tinghög et al., 2013; 2016). Subjects in the delay condition were asked to think carefully about their decision, being forced to wait at least 60 seconds before answering. Previous research opted for a 10 seconds threshold in the delay condition, but we raise this limit to stress the experimental manipulation. In both these latter conditions, a timer counted down/up the time passed from the presentation of the decision page, in order to boost the external superimposition. Nonetheless, participants in the constraint condition were allowed to express their decision also after the conclusion of the countdown, as suggested by Tinghög et al., (2013). Finally, participants in the control condition were requested to perform the same activity of the previous two groups, but without any kind of time limitation. After the presentation of each dilemma, a comprehension check question was presented, to assess the understanding of basic features regarding the presented storytellings. One subject failed in responding correctly to one of these check questions, causing their removal from the final sample. After the completion of the dilemma section, further information was collected towards autonomous transportation, such as: moral acceptability of the two driving styles (bipolar slider, 0 = self-protective

AVs are the most morally acceptable solution, and 100 = utilitarian AVs are the most morally acceptable solution); opinion on which of the two driving styles will be effectively implemented in the future, and willingness to buy self-protective and utilitarian AVs. Subsequently, the Italian version of the 30-item Moral Foundation Questionnaire was administered (Bobbio et al., 2011), followed by the Slider measure of SVO (SLM, Murphy et al., 2011), and concluding with a series of questions on daily social interactions and general information (socio-demographic).

5.3.4 Analysis

The statistical analysis was conducted in the R environment (version 4.1.1; R Core Team, 2021). As preliminary steps, first a single index for SVO was computed for each participant following the procedure suggested by SLM's authors (Murphy et al., 2011), through the formula:

$$SVO° = arctan\left(\frac{(\bar{A}_{other} - 50)}{(\bar{A}_{self} - 50)}\right)$$

Notice that $\bar{A}$ is assumed to be the mean allocation for the other and for the self, respectively. Minimum competitive scores is set at -12.04 points, while maximum altruist score is set at 57.15 points. Considered the unbalance proportion of the derived SVO profiles and the methodological indication from Bakker and Dijkstra (2021), the SVO score was employed as a continuous variable, and subject were grouped in 'prosocial' (combining the *altruistic* and the *cooperative* profiles) and 'proself' (combining the *individualistic* and the *competitive* profiles). Contextually, the moral consistency profiles were defined on the basis of subjects' moral decisions. Individuals who endorsed always the same moral code throughout the three dilemmas were labelled as 'fully utilitarian' and 'fully nonutilitarian', while individuals who changed their moral decision at least once were defined as 'switchers'. A number of descriptive information on several variables are reported in the results section (Table 12) and in the

Appendix (tables A3.2 and A3.3), divided by SVO and moral consistency profiles. Moving to the statistical analysis, a correlation analysis was computed considering several factors (schooling, moral evaluation, willingness to buy AVs, social interactions and MFQ's subscales total scores) and reported in a graphical form (Figure 18). Consequently, a series of statistical models were implemented for testing the experimental hypothesis, setting as dependent variables: (M_1) moral decision, (M_2) decision times, (M_3) moral evaluation and (M_4) willingness to buy an AV programmed to follow one of the two proposed moral codes. Given the nature of the data, different kinds of statistical models were fitted to the data through the R package lme4 (Baters et al., 2015), such as both generalized and mixed effects linear models, as well as simple linear model in the case of moral evaluation. The models presented in the main analysis ($M_1 – M_4$) are the result of four corresponding forward stepwise model comparisons, which considered models with a number of different predictors. The chosen one was selected on the basis of the Akaike Weights comparison procedure (Wagenmakers and Farrell, 2004). Post hoc pairwise comparisons were considered when requested, using the R package *emmeans* (Lenth, 2020). Bonferroni's correction was set as an adjustment method. A 98% acceptance interval was considered in terms of decision times, which were transformed in their logarithmic form consistently with Lotto et al., (2014). The final dataset and further supplemental information are retrievable in the OSF project folder: https://bit.ly/3cksq6Q.

5.4 Results

Figure 19 chromatically describes the correlation between a number of individual attitudes and moral judgments. Expectably, higher the willingness to buy an AV programmed to follow a particular behavior (self-protection or utilitarian), higher was the moral evaluation of the correspondent behavior (self-protection: $r = .20$; utilitarianism, $r = .38$). Interestingly, the higher the competition in everyday life, the lower the moral evaluation of utilitarian AV ($r = -.18$). Oppositely, the availability to purchasing utilitarian AVs was positively correlated with the number of daily interactions $r = .18$). Prosocial orientations were inversely correlated with willingness to buy self-protective AVs (*r* = -.18) and positively with harm/care moral foundation, which underlies kindness and gentleness (*r* = .23). Finally, religiosity was observed to be positively correlated with three moral foundations: ingroup/loyalty (*r* = .38), authority/respect (*r* = .37), and purity/sanctity (*r* = .44).

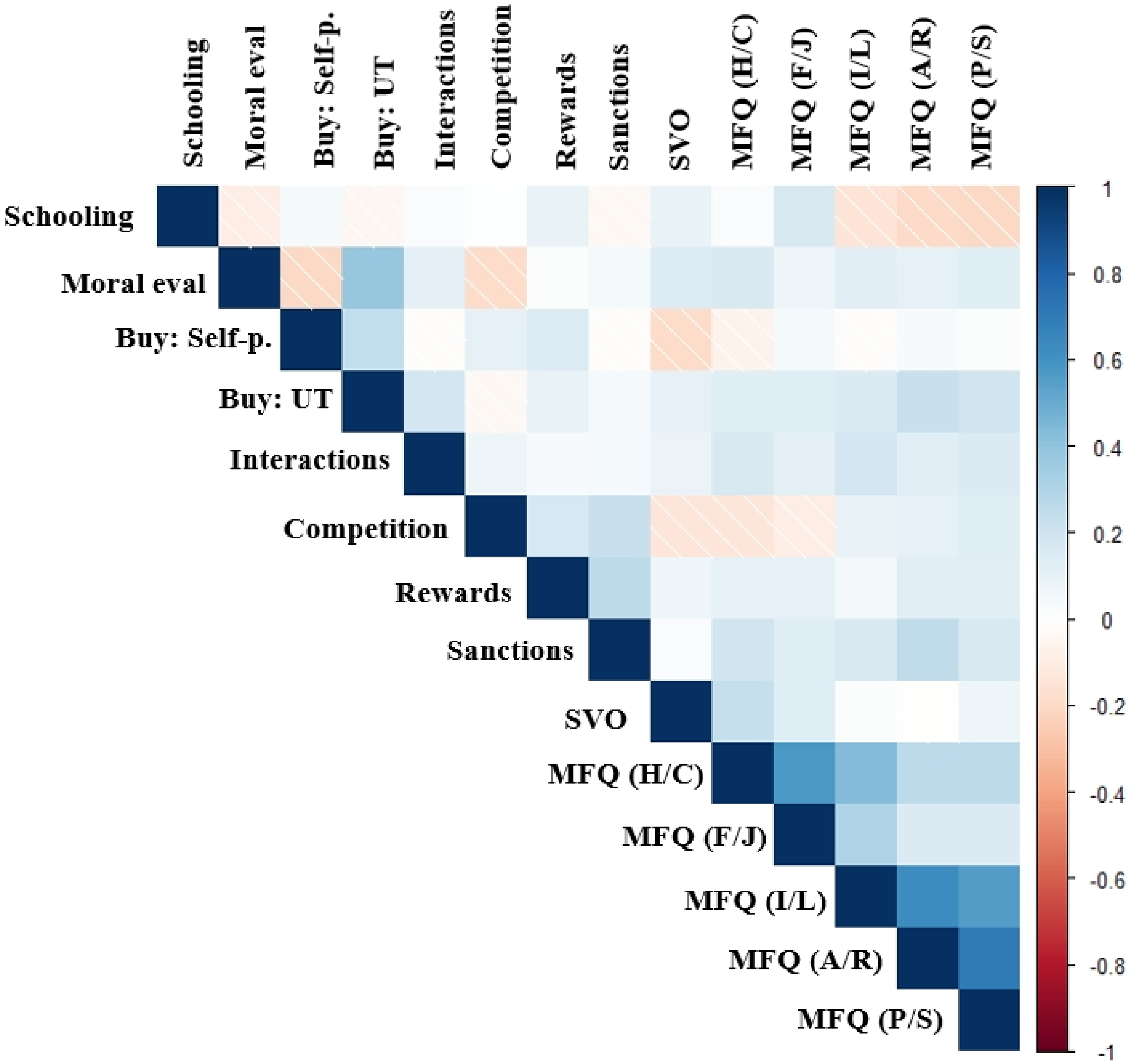


Figure 19: Heat map plot describing the correlations between variables on a chromatic form. for the 7-meters throw. Each cell represents the correlation between the row variable and the column variable. The level of agreement is chromatically represented from dark red (Pearson's r = -1, negative correlation), to white (r = 0, no correlation) and dark blue (r = 1, positive correlation). Significant correlations are reported as diagonally slashed cells.

Notes: the full names of the variables are schooling, moral evaluation (0 = self-protective AV, 100 = utilitarian AV), willingness to buy self-protective AV, willingness to buy utilitarian AV, iterated interactions, competition with peers/colleagues, rely on rewards after good behaviors, risk of sanctions after poor behaviors, Social Value Orientation, MFQ (Harm/Care, Fairness/Justice, Ingroup/Loyalty, Authority/Respect, Purity/Sanctity).

Table A3.1 (Appendix) summarizes the predictors included in the selected models (M_{1}- M_{4}), coherently with the obtained estimates and the corresponding effect sizes. The binomial family distribution was set as a reference point for implementing a generalized linear model (M_{1}), with the moral decision as the dependent variable. Following the evidence obtained from the computed onward stepwise regression, the selected model considered experimental conditions (time constraint, time delay, control) dilemma order, SVO profiles (prosocial, proself), and gender as fixed effects, as well as experimental conditions in interaction with (i) dilemma order and (ii) SVO profiles. Results showed no significant effect of time manipulation on the moral decision ($\chi^2{}_2 = 2.28, p = .320$). Nonetheless, some worth noting descriptive trends arose, which have been deepened in the discussion of the present study. The endorsement of the utilitarian behavior was significantly higher responding to the last dilemma when compared to the first one ($\chi^2{}_2 = 7.50, p = .023$; D1 – D3: z = -3.01, $p = .008$). This trend was especially visible in the control condition, where no time limitation was superimposed ($\chi^2{}_4 = 14.35, p = .006$; Figure 20).

Table 11: Descriptive information on the presentation of the three AV dilemmas, in terms of decision time (mean and standard deviation, in brackets) and percentage of endorsement of the utilitarian AV behavior, both overall and specifically for each dilemma position in the randomized experimental order (first, second, third). The information are reported by experimental condition (time constraint, time delay, control condition) and overall.

	Decision time	**Moral decision (overall, %)**	**Moral decision (Dilemma 1, %)**	**Moral decision (dilemma 2, %)**	**Moral decision (dilemma 3, %)**
Time conditions					
Time constraint (N = 69)	5.42 (3.48)	42.51	42.02	49.27	56.52
Time delay (N = 70)	5.41 (5.15)	59.22	61.43	55.71	57.14
Control (N = 67)	7.52 (9.18)	48.06	35.82	56.71	55.22
Overall (N = 206)	6.10 (6.44)	52.26	46.60	53.88	56.31

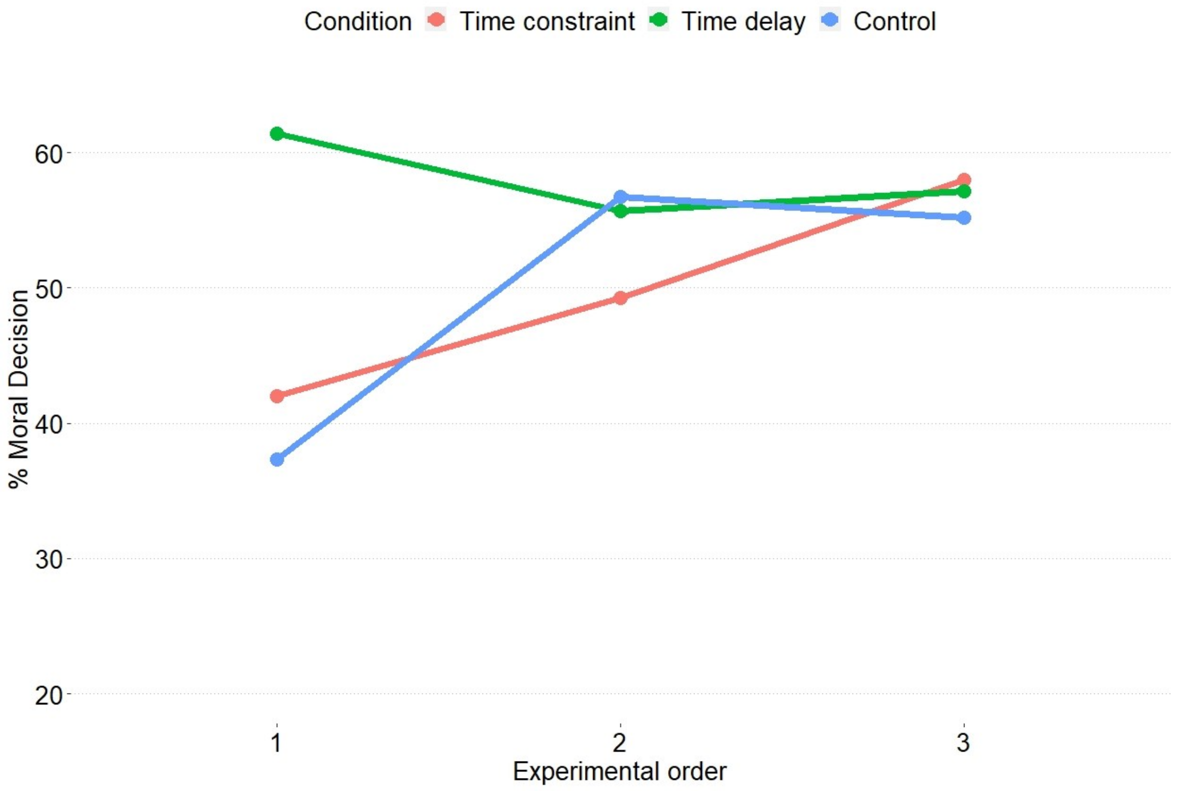


Figure 20: Percentage of endorsement of the utilitarian AV behavior per experimental condition and dilemma order.

In terms of prosocial orientation, a main effect of SVO profile was detected ($\chi^2{}_1 = 7.40$, $p = .006$), in the form of a higher utilitarian endorsement for prosocial individuals (54.89% Vs. 33.33%). The original hypothesis on the interaction between prosocial orientation and time superimposition was not sustained by data ($\chi^2{}_2 = 0.87$, $p = .645$). In this direction Figure 21 describes the likelihood of utilitarian endorsement assuming the continuous SVO scores, as suggested by Bakker and Dijkstra (2021). A descriptive discussion of the observed trend has been developed in the discussion of this study.

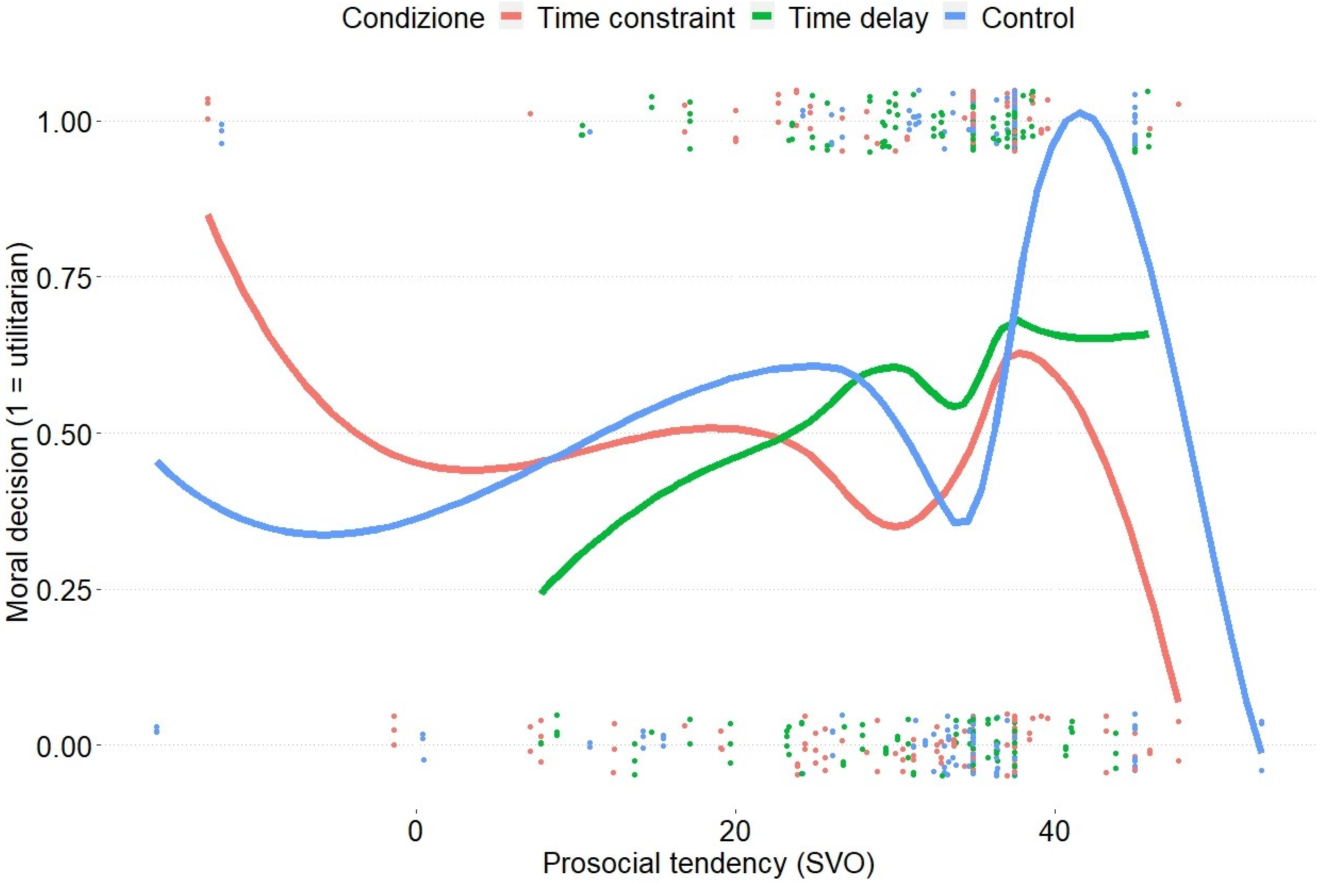


Figure 21: Smoothed curves on the likelihood of endorsement of the self-protective (= 0) and utilitarian AV behavior (= 1) by SVO score, divided by experimental condition

Subsequently, a series of inferential analyses were performed to profile the derived moral consistency in terms of decision times, moral evaluation, and willingness to buy AVs programmed to follow one of the two proposed behaviors. As clarified in the analysis section, all the following results have been obtained by fitting mixed (M_2, M_4) and simple linear (M_3) models selected from three specific Akaike model comparison analyses. The decision time mixed linear model (M_2) assumed the moral profile (fully utilitarian, fully nonutilitarian, switcher), the experimental condition, and the dilemma order as fixed effects, as well as the interaction between experimental condition and dilemma order. Fully utilitarian individuals were faster than switchers in taking their decisions ($\chi^2{}_2 = 6.14$, $p = .042$; UT – switchers: t = -2.43, $p = .042$). Consistently with Study 1 and Study 2, decision time speeded up throughout the experimental session ($\chi^2{}_2 = 100.33$, $p < .001$). Predictably with the experimental design, this trend was not observed in subjects under the Time Delay condition, which had to wait 60 seconds before the selection of the moral option ($\chi^2{}_2 = 17.45$, $p < .001$; Figure 22).

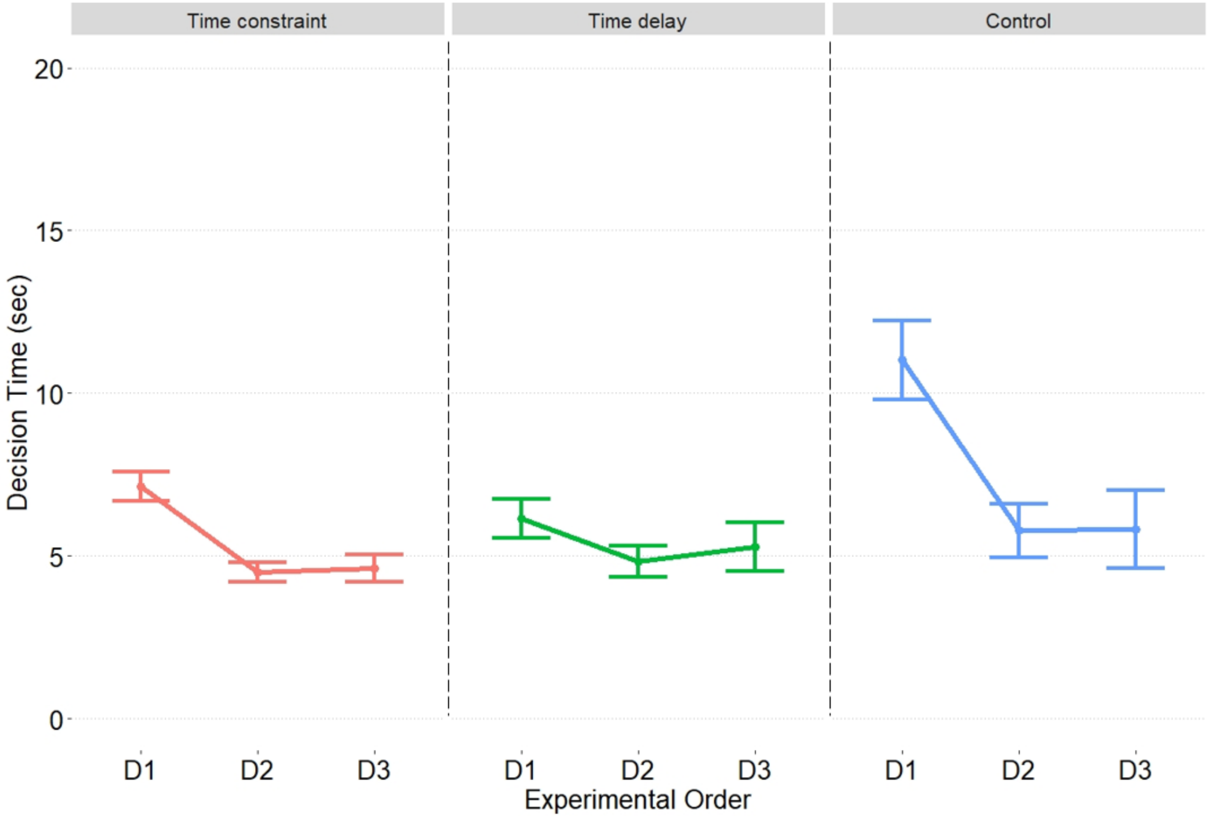


Figure 22: Mean and standard errors of decision times, divided by experimental order (D1, D2, D3) and experimental conditions (time constraint in red, time delay in green, control in light blue)

Finally, moral evaluation was set as dependent variable of a simple linear model, assuming moral consistency profile and gender as predictors (M_3). Contextually, 'willingness to buy' was set as the dependent variable of a mixed effects linear model (M_4), considering the moral consistency profile and the AV driving type (self-protective, utilitarian) as fixed effects, as well as their interaction. Fully utilitarian and switchers agents were more prone to evaluate the utilitarian AV as the most morally acceptable option when compared to fully nonutilitarian individuals ($F_2 = 25.73$, $p < .001$; Table 12). Finally, fully utilitarian and nonutilitarian agents were more willing to purchase AVs when the vehicle was programmed coherently with their preferred moral code ($\chi^2{}_2 = 25.45$, $p < .001$).

Table 12: Mean and Standard Deviation (in brackets) OF total SVO scores (from -12.04 to 57.15 points) moral acceptability of the two proposed AV behaviors (0 = the self-protective option is the most acceptable behavior, 100 = the utilitarian option is the most acceptable behavior) and willingness to buy AV programmed as self-protective ('self-p') and utilitarian ('ut'). The information are reported by SVO profiles (proself: prosocial: individualistic and competitive: altruistic and cooperative), moral consistency profiles (fully utilitarian, fully nonutilitarian, switcher), and overall.

	N	**SVO scores** **(-12.04 to 57.15)**	**Moral acceptability** **(0 = NUT, 100 = UT)**	**Willingness to buy** **(SELF-P)**	**Willingness to buy** **(UT)**
SVO profiles					
Proself	22	8.9 (10.8)	58.1 (32.2)	64.4 (32.2)	39 (29.2)
Prosocial	184	35.2 (5.2)	69.5 (29.3)	50.2 (33.1)	48.8 (32.6)
Moral profiles					
Utilitarian	79	33.4 (9.6)	83.0 (22.4)	43.7 (33.8)	56.2 (32.7)
Nonutilitarian	71	30.5 (12.0)	50.5 (31.0)	55.7 (32.7)	37.6 (30)
Switcher	56	33.3 (7.9)	70.1 (25.3)	57.9 (31.4)	48.8 (31.5)
Overall	206	32.3 (10.2)	68.2 (29.7)	51.7 (33.2)	47.8 (32.3)

5.5 Discussion

Following the evidence collected throughout the two previous studies on the moral perception of autonomous transportation, in Study 3 we focused on two individual features that may play an important role in the endorsement of a particular AV behavior, which are the time availability and the prosocial orientation.

Focusing on the role of time superimposition on moral judgment, results did not show any statistical difference between the three experimental conditions, in which participants were asked to express their moral decision within 8 seconds (i.e., time constraint), after 60 seconds (i.e., time delay), or with no time limitations (i.e., control condition). This result seems consistent with Tinghög et al., (2016), which claimed that moral judgment is immune from intuitive processing and time pressure, contrasting previous research on the positive effect of time constraint on the rate of deontological/nonutilitarian responses (Suter and Hertwig, 2011; Greene et al., 2001; 2004). Consistently with Tinghög et al. (2016), a possible interpretation could refer to the moment in which moral rules are formed. Mallon and Nichols (2011) stated how Dual Process Theory (DPT) has the limitation to neglect the possibility that there can be moral rules formed and reinforced in time, that can be effortlessly recalled and applied during intuitive moral reasoning. This interpretation is convergent to the Social Heuristic Hypothesis (Rand et al., 2012), which claims how moral reasoning under time pressure is solved through the employment of the easiest possible moral approach, which can also be different from the traditional intuitive deontological reasoning. Consequently, the biggest part of our daily moral reasoning has already been shaped in the past, as part of our moral development process. In these terms, moral reasoning cannot be easily triggered or conditioned by time constraints in response to a specific event, but moral decisions may be considered assuming the involvement of more stable individual characteristics. Conceivably, another potential explanation can be related to the operation of a weak experimental manipulation, which failed to reach the hypothesized effects. The adopted time manipulation was consistent with previous investigations (Rand et al., 2012; Tinghög et al., 2013, Trémolière et al., 2017), and adapted to the textual nature of the experimental material (Rosas and Aguilar-Pardo, 2020). Possibly, a laboratory setting would have helped to have a more careful experimental control on the presentation of the dilemmas and on potential distractors. Nonetheless, interesting insights for future investigations on the relationship between time and moral decision-making arose from descriptive results. From this perspective, the proportion of utilitarian decisions among the three experimental condition endorses the interpretation of deliberation as a motivation for utilitarian and cooperative behaviors, which would be consistent with a considerable line of research (Frank et al., 2019; Goeschl and Lohse, 2018; Suter and Hertwig, 2011) and with the DPT (Greene et al., 2001; 2004). Furthermore, clearer evidence in this sense was collected observing the first showed AV dilemma. A stronger effect of the experimental manipulation at this level would be supported by the number of studies that operated with one-shot interactions or a single stimulus (Frank et al., 2019; Rand et al. 2014; Goeschl and Lohse, 2018). In the same direction, a further interesting information came by the overall improvement of the utilitarian endorsement under time constraint throughout the experimental session, coherently with a progressive reduction of decision times. Considering all the described evidence, our idea is that is still possible that deliberative moral reasoning facilitates utilitarianism, albeit this descriptive trend needs to be further investigated. Nonetheless, when moral agents already had the possibility to focus on the moral problem (also under time constraint), subsequent dilemmas would be solved faster and with less "moral conflict". The reduction of decision time during the experimental procedure is evident since Study 1, which confirms – when investigating moral judgment - the necessity to focus on the definition of a right number of stimuli while conducting an iterated-moral dilemmas experiment (see section 3.5).

The individual prosocial orientation was the second individual factor investigated in the present study, hypothesizing a higher favor towards the utilitarian/cooperative resolution for higher scores of Social Value Orientation (SVO, Murphy et al., 2011), and especially under time pressure (Rand et al., 2012, 2014). Overall, results showed that individuals categorized as ‘prosocial’ (SVO’s altruistic and cooperative profiles) were more likely to endorse the utilitarian AV behavior when compared to the ‘proself’ group (SVO’ individualistic and competitive profiles). This evidence was quite predictable, since utilitarian AVs – which promote the greatest overall safety - are widely perceived as the most prosocial version of the autonomous transportation technology, and consequently the most appropriate vehicle for public and private use (Gogoll and Müller, 2017; Shariff et al., 2017). Additionally, in the last years prosocial behavior has been described on the basis of several definitions and applications, and some of them - emphasizing consequences (Schroeder and Graziano, 2015) and intentions toward the collectivity (Eisenberg and Miller, 1987; Pfattheicher et al., 2022) - seem in line with the utilitarian approach. Nonetheless, our hypothesis was specifically referred to the effect of time superimposition on the enhancement of utilitarian/cooperative AV behaviors, coherently with the SHH which described the positive relationship between time constraint and cooperative actions (Rand et al., 2012; 2014). This theory was not confirmed by data, which showed no differences between prosocial and proself individuals in the endorsement of the greater goal-solution (i.e., utilitarian). Albeit in line with the hypothesis of no-relation between time limitations and cooperativeness (Tinghög et al., 2013), these results have to be taken with a grain of salt. Bakker and Dijkstra (2021) have been critical about restricting SLM profiling to a categorical factor, since it appears to be inconsistent with the SVO definition itself, which defines the measure as the assigned weight of benefit in situations of interdependence (Messick and McClintock, 1968; Murphy and Ackerman, 2014). Additionally, they claim that despite the SLM seems to be the most reliable measure of SVO, it has the strongest orientation to favor particular prosocial categories, namely the ‘central’ cooperative and the individualistic profiles, potentially leading to underestimate the relation between the prosocial measure and the outcome. For this reason, we opted to combine the two couples of prosocial and proself profiles in two macro-categories, but still, the sample was unbalanced in favor of prosocial profiles (89%). Future studies may try to control this factor a-priori, stratifying the sample between the four SVO profiles aiming for more balanced groups. Despite the few evidence collected on the correlation between SVO scores and social interactions/work habits, this would be possible focusing on specific professional environments in which competition (or cooperation) challenge individual orientation towards prosocial and altruistic behaviors (Grant and Shandell, 2022; Kilduff et al., 2016; Vianello et al., 2010). Nevertheless, also here descriptive information leave some space for considering the possibility of retesting this hypothesis in future investigations. In fact, looking to how the likelihood of the utilitarian endorsement changes relatively to time availability and SVO scores (as suggested by Bakker and Dijkstra, 2021), an interesting trend appears looking at the time delay condition (Figure 20). Here, we observed how the likelihood of endorsement of the utilitarian AV behavior appeared to be not completely independent from SVO scores, describing a potential positive correlation. If demonstrated across future investigation, this would be coherent with the hypothesis of deliberate reasoning stimulating cooperative behaviors (Goeschl, and Lohse, 2018; Suter and Hertwig, 2011; Greene et al., 2001; 2004).

Finally, we assumed the possibility to observe distinctive behaviors (i.e., decision times) and evaluations towards AVs (e.g., moral acceptability and willingness to buy) coherent with the described moral consistency profile. Three moral profiles were retrieved by the given responses to the three AV dilemmas: fully utilitarian, fully nonutilitarian, and switchers (for the individuals who changed their moral decision at least one time). The consistently utilitarian profile appeared to be coherent with expectations, showing lower decision times, moral acceptability mainly directed towards the utilitarian AV and higher interest to buy utilitarian AV when compared to a self-protective model. This result demonstrates how moral consistency can be considered as a relative proxy between individual moral judgment and attitudes towards AV technology, especially in the description of ‘utilitarian’ moral agents. In fact, no clear pattern was observed for consistently nonutilitarian and switchers individuals, mainly in the expression of interest in purchasing an AV in the future. This result is consistent with the ‘social dilemma of self-driving cars' described by Bonnefon et al., (2016), which described the contrast between moral evaluation and availability to share an AV programmed to follow the utilitarian moral code. Here, results show a recognizable pattern especially for utilitarian agents, but the willingness to buy an AV was not enough 'polarized' throughout the three moral profiles (descriptively, around the center of the evaluation continuous scale). Interestingly, the utilitarian approach was evaluated as more acceptable by fully utilitarian and switchers moral agents, while nonutilitarian individuals did not take a specific stance between self-protective and utilitarian models, slightly preferring the self-protective AV for themselves. These results help to describe attitudes towards AVs assuming the stability of moral judgment throughout the experimental session, a characteristic that has also been investigated in Study 4.

In conclusion, evidence collected from Study 3 converge in describing a negligible role of time superimposition and prosocial orientation on influencing the endorsement of utilitarian and cooperative AV behaviors. Albeit utilitarianism and prosocial orientation both relate to ensuring overall welfare, this result may be explained by a number of previous research focused on the differences between utilitarian judgment and the interest of providing the greatest collective good (Bartels and Pizarro, 2011; Côté et al., 2012; Dovidio et al., 2006; Wilson et al., 1996). Nevertheless, the observed descriptive trends do not allow us to satisfactory address our research questions, justifying the need for further exploration concerning the role of decision time availability and prosocial orientation in shaping moral judgment and attitudes towards autonomous transportation.

Chapter 6

Study 4: The role of information availability and perspective-taking: moral judgment behind the Veil of Ignorance

6.1 The rationale of the study

Personal perspective has been recognized as an important bias in individuals' moral decisions, also playing a major role in the definition of social perception and expectancies towards autonomous transportation (Bonnefon et al., 2016; Shariff et al., 2017). An emblematic example in this sense is the 'social dilemma of self-driving cars' (Bonnefon et al., 2016), assuming how people consider the most moral AV behavior the one who reduce the number of endangered lives even if it costs the life of the passenger but prefer a self-protective vehicle when imaging as the passengers of the vehicle. A possible solution to this discrepancy is suggested by Martin et al. (201b): moral dilemmas involving AVs should be resolved considering both the passenger and the pedestrian perspectives. Nonetheless, this rarely happens in the literature, and the majority of the experimental applications focus on AV occupants/passengers' safety, underestimating the perception of other road actors (e.g., pedestrians; Borenstein et al., 2019). This may be mainly related to the stronger mediatic emphasis given to road accidents involving AVs, describing critical events mainly from the perspective of the AV passenger in relation to their monitoring activity during the autonomous driving (e.g., Petrović et al, 2020; Randsazzo, 2019). Despite this lack of attention, a number of studies highlighted the importance of considering the role of perspective-taking (PT) while evaluating the morality of AV behaviors. Kallioinen et al., (2019) investigated moral beliefs towards autonomous and nonautonomous vehicles in a virtual environment, assuming the perspective of car occupants (drivers or AV passengers), pedestrians and uninvolved observers. No stronger differences were observed in terms of level of automation (consistently with Study 2), but PT seemed to influence moral judgment, boosting the endorsement of self-protective resolutions.

Later, Mayer et al., (2021) confirmed this effect in a vignette-based study, claiming an overall advantage of the utilitarian resolution but a tendency to self-protection when the number of lives at stake was within the 1:5 ratio.

An interesting strategy for the investigation of the role of PT in shaping moral judgment has been retrieved from the theory of the political philosopher John Rawls. In 1971' *A Theory of Justice*, Rawls describes his idea of 'social contract' as the agreement between a group of individuals for the sake of collective fairness (Rawls, 1971/2009). For the definition of the leading principles of this agreement, Rawls suggested that individuals' reasoning has to take place in a hypothetical setting, the *Original Position,* which is located behind the so-called *Veil Of Ignorance* (VOI). Behind the VOI, individuals (i.e., rational agents) have the possibility to take impartial and disinterested decisions for the sake of fairness, since they are deprived of contextual and personal information about the self and about the other people affected by these decisions. In other words, only when individuals are unaware of their own and others' individual characteristics, social positions and relationships, they are capable to reach an agreement aiming at the fairest and egalitarian version of society (Maxcy, 2002; Moehler, 2018). This is a typical 'decision under ignorance' situation, which is typically investigated in behavioral economics to deepen situations in which the rational agent knows the full set of alternatives, but has no information about their effects (e.g., Arrow and Hurwicz, 1982; Krug et al., 2020). Rawls claimed that, behind the VOI, the most appropriate decisional process to follow should be the *maximin* strategy ("maximize the minimum"). Following this decisional rule means ensuring the greatest possible benefit to the least-advantaged member of the group when the disposable outcomes are uncertain and have even odds of happening (Rawls, 1971/2009). Rawls assumed that the nature of this decision was anyway selfish, since the agent is aware that valuing the condition of the less-advantaged person may be far-sighting for the self in particular and unlucky situations (Ashford and Mulgan, 2013). The Rawslian decisional process has been disputed by John Harsanyi, which enforced the idea of impersonality over impartiality

(Harsanyi, 1975; 1978). According to the philosopher, individuals are *Bayesian* agents behaving for maximizing their expected individual utility (on the basis of the Rational Choice Theory by Coleman, 1994). Nevertheless, when key contextual and personal information are concealed, their behavior would change, likely aiming to follow the average utility principle, as an equal partition of the questioned resource between the parts. In other words, when individuals cannot clearly favor themselves, they will opt to prioritize collectivity in a more utilitarian -albeit still primarily selfish - sense (Moehler, 2018). Rawls and Harsanyi describe two different (although theoretically convergent) decisional processes under a clearly defined state of ignorance, hypothesizing two VOIs possessing slightly different features. Rawlsian VOI can be defined as a *Thick Veil,* where the agent is completely deprived of contextual and individual information on the self and the others (i.e., the 'No Knowledge Formula', Parfit, 2011). Divergently, behind the Harsanyian VOI rational agents are aware of their existing role in society, being able to dispose of this additional information behind a *Thin Veil* (i.e., the 'Equal Chance Formula', Parfit, 2011). Recently, the hypothetical VOI environment has been interpreted and then applied as an experimental setting in the investigation of moral reasoning, in which the individual had little knowledge about the context and attributes. As a first attempt, Huang et al., (2019) demonstrated that the VOI's impartial thinking has the potential of affecting moral reasoning, boosting the endorsement of the utilitarian moral code as an attempt to maximize self-beneficial outcomes. Importantly, Martin et al. (2021a; 2021b) followed this experimental approach but criticized previous investigation on morality of AVs. First of all, the authors marked a flaw in the methodology adopted by Bonnefon et al. (2016), since participants were subjected only to a partial PT (the AV passenger perspective), reducing the generalizability of their conclusions, and underestimating the role of PT on moral reasoning (Kallionen et al., 2019; Mayer et al., 2021). Secondly, they also highlighted how Huang et al. approach (2019) was characterized by uneven odds of being each of the characters involved in the scenario, injecting selfishness in the moral reasoning. Indeed, Huang's study presented a scenario with 10 characters, in which the moral agent had a 1-to-10 chance of being the AV passenger and a 9-to-10 chance of being one of the pedestrians. To fill these gaps, Martin and colleagues developed a between-subject study comparing partial and full PT accessibility, assuming even odds of being each of the characters involved. Results confirmed the increased likelihood of utilitarian reasoning when contingent information are blurred *behind the VOI.*

The present research aims to deepen the effect of VOI reasoning (i.e., full PT) on moral judgments towards AV, taking few steps forward in the optimization of the technique when operationalized in applied experimental settings. In this sense, we investigated the endorsement of three different AV behaviors through a *funnel* within-subject approach to perspective-taking and accessible information. Firstly, impartiality was stimulated adopting the Rawslian version of the VOI (no contingent information, *Thick Veil*). Then, impersonality was triggered adopting the Harsanyian version of the VOI (no personal information, *Thin Veil*). Finally, all the contextual and individual information were disclosed in the *No Veil* version of the AV dilemma, where moral agents moved from a full to a partial PT about the moral scene. The proposed AV behaviors where in line with the decision strategies suggested by Rawls and Harsanyi, namely the maximin and the utilitarian codes, as well as with the nonutilitarian resolution, typically adopted in the AV dilemmas. These strategies were assumed as equivalent to three potential and adoptable AV rules of choice when facing critical road events: i) prioritize the safety of the passenger (nonutilitarian and self-protective rule), ii) minimize the number of casualties (utilitarian rule) or iii) optimize the worst possible outcome between all the alternatives (maximin rule). To allow for the presentation of a three-options moral dilemma, the traditional AV dilemma was adapted to the Thomson's '*Bystander three options'* version of the *Switch* dilemma, also defined as a *moral trilemma* (Thomson, 2008). This moral stimulus has been recently discussed (Di Nucci, 2013), with the aim of disentangle the permissibility of self-sacrifice when pursuing the utilitarian resolution. In fact, this dilemma give the moral agent the possibility to divert the course of the trolley not only against a third character, but also in the opposite direction, admitting the self-sacrifice in the name of the greater goal. Huebner and Hauser (2011) tested this tool in an experimental setting, confirming the dominance of the utilitarian resolution independently from the presence of an altruistic self-sacrificial option. In the present study, the three alternatives were aptly manipulated in each VOI condition in order to be coherent with the amount of detectable information, in the form of numerical individual chances of survival. Further specification on the adopted stimuli are presented in the 'stimuli and experimental material' section (6.3.2).

To the best of our knowledge, Study 4 is the first experimental application which distinguishes between Rawslian and Harsanyian VOI reasoning in the investigation of moral judgment towards autonomous transportation, aiming to differentiate their proposed decisional strategy (maximin Vs. utilitarian) when moral reasoning is stimulated under full and partial PTs and with a different amount of contextual and individual information. Full PT has been enhanced in the two VOI scenarios (*Thick* and *Thin* veils), while partial PT has been disclosed in the last *No veil* scenario, in which moral agents have been divided in two partial-perspective conditions (the AV passenger and one of the pedestrians). Further clarification of the experimental procedure is presented in the correspondent section (6.3.3). Additionally, considering the sequential paradigm nature of Study 4 and results collected from Study 3, moral consistency profiles were again deepened, so to describe how moral consistency throughout a VOI experimental application has the ability to shapes moral acceptability and willingness to share potential AV moral behaviors. Finally, the Interpersonal Reactivity Index (IRI) was employed to describe potential relations between individual empathy, socio-demographic characteristics of the sample, and evidences collected in terms of moral judgment. Utilitarian reasoning has been observed to be enhanced by reduced empathetic concern (Gleichgerrcht and Young, 2013; Patil and Silani, 2014) and reduced personal distress (Aridağ and Yüksel, 2010), demonstrating a relationship between interpersonal characteristics and moral judgment.

6.2 Hypothesis

A series of hypotheses were advanced in the present study, considering the evidence collected in Study 3, as well as the thoughts described in the rationale of the present chapter and in section 1.2:

- Coherently with the reference theories (Di Nucci, 2013; Huebner and Hauser, 2011; Rawls 1971/2009; Harsanyi 1975; 1978), different decision strategies should be adopted throughout the different VOI conditions. Specifically, the maximin rule should be the electing choice behind the Rawslian Thick Veil, while moral agents should more frequently follow the utilitarian approach behind the Harsanyian Thin Veil. Oppositely, when all the information are available to the participant in the No veil condition, a higher endorsement of the self-protective/nonutilitarian resolution is expected when compared to the full PT conditions.
- Given the evidence collected from Study 3, the derived moral consistency profiles were expected to be consistent with the moral acceptability of the proposed AV behaviors and willingness to buy AVs programmed in the described ways (i.e., aim to protect the AV passenger, to minimize the total number of casualties, or to maximize the protection of the last-advantage character). For example, fully utilitarian moral agents would show higher moral evaluation of utilitarian AVs, as well as higher willingness to purchase this kind of technology for themselves.

6.3 Method

6.3.1 Participants

An a-priori power analysis has been computed on G-power statistical software (Faul and Erdfelder, 1992) before computing the data collection, assuming a medium effect size (Cohen's d = 0.20) and a correlation of 0.50 among repeated measures, with an alpha error probability of 0.05 and 0.95 of power. The system suggested a total number of 220 participants, and 251 participants were recruited for the experiment. The final sample counted a total number of 239 participants: 12 subjects were excluded as them failed to correctly answer to a check question during the completion of the experiment (see Procedure). The final sample was composed of 50.21% females (*n* = 120). Overall, the mean age was 28.28 years (SD = 8.26, range = 18-63), the mean schooling age was 16.94 years (SD = 2.78), and 51.46% of the participants were enrolled in university courses (n = 123), with 24.68% working as employees (n = 59). Most participants (88.28%) had held diver licenses (n = 211), and the majority (92.89%) already heard about autonomous transportation technology (n = 122). 53.55% of the sample have been involved in a road accident at least once in a lifetime (*n* = 128), and only 5.85% (*n* = 14) was involved in at least one road accident in the last 12 months. When participants were asked to imagine themselves in a hypothetical on-road scenario, half of the sample described themselves in the role of the driver (50.6%, n = 121), the 28.87 as passengers (*n* = 69), and 20.50% as pedestrians (*n* = 49). The study was approved by the local ethics committee (ID No.: 4420), developed through Qualtrics software (Qualtrics, Provo, UT), and distributed via Prolific platform (Palan and Schitter, 2018; Peer et al., 2017). Before their participation, each participant gave formal written consent, which was voluntary, and participant were remunerated for their time. An hourly rate contribution of 12.70€ was ensured, and on average, each individual was rewarded with 3.20€ for their participation.

6.3.2 Materials

Three self-involvement, sacrificial and incidental moral trilemmas were developed for this study. All the three scenarios, in a textual and vignette-based form, depicted the same traffic event involving an AV driving on an urban road with a single passenger on board. Here, the vehicle is approaching a road intersection, where three pedestrians are crossing the road right in front of the AV. Due to a non-human-related malfunction to the traffic lights coordination, the pedestrian are approaching the road crossing in the exact moment in which the AV is crossing the intersection. The dynamic of the event does not allow the vehicle to brake safely, leading to an unavoidable crash. The storyline and the vignette were slightly readapted from Martin et al. (2021a, see below). In the Thick and the Thin Veil scenario, participants were informed that they could be the AV passenger or one of the pedestrians crossing the road. Oppositely, in the No Veil scenario, the conditional tense was replaced by the present indicative tense, with participant assigned to one of the two partial PT conditions (AV passenger or pedestrian). The textual scenario was always paired with a 2D vignette representing the situation, readapting the labels to the level of information disclosed in the text (Figure 23).

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

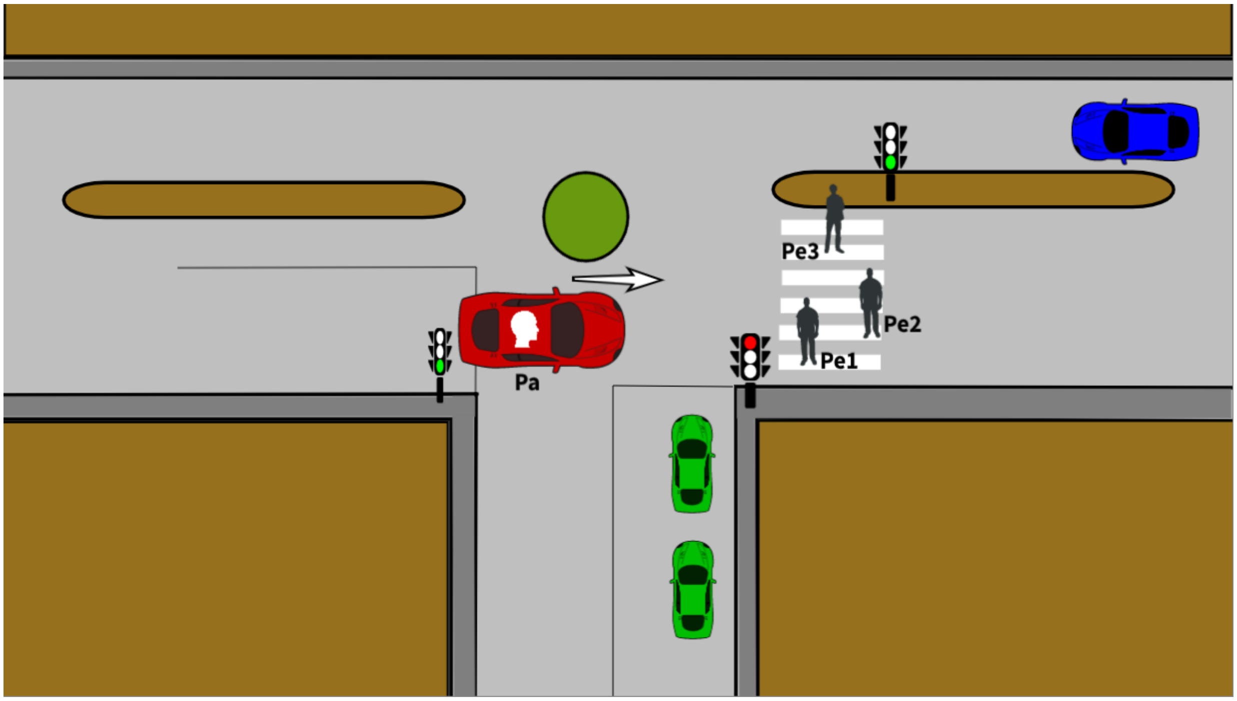


Figure 23: The AV trilemma vignette, deployed in the Thick and Thin veil conditions. In the context of the No Veil scenario, and balanced on the experimental condition, one of the labels (Pa or Pe1) was replaced by 'You' (Pa = Passenger; Pe1/Pe2/Pe3 = Pedestrians).

Each AV trilemma was associated with three alternative outcomes with even odds of realization, which corresponded with three potential AV behaviors in reaction to the critical event. In each alternative, the four characters (one AV passenger and three pedestrians) were put in danger with different individual probability risk, which were aptly manipulated to be consistent with the investigated rules of choice (maximin, utilitarian, nonutilitarian). In order to clearly distinguish the decisional rules and to mask personal information, the alternatives were presented in the form of numerical individual chances of survival. The technique of presenting risk though probabilities has been widely thorough, especially in the field of risk communication (Bonner et al., 2021; Gigerenzer and Galesic, 2012; Waters et al., 2007), and a series of best practices have been suggested in the literature and assumed in the present study. Probabilistic risk communication appears to be more effective when the information is presented as numbers instead of words (Trevena et al., 2006). Low numeracy has to be taken under consideration, since it may have a detrimental role on individuals, potentially leading to an overestimation of risk (Hill and Brase, 2012; Weinstein et al., 2004). Nonetheless, it seems to have a larger impact on risk described as frequencies than in the equivalent probability format (Peters et al., 2006; 2011; Schapira et al., 2004). Coherently with the described evidence, the trilemma's resolutions were described as the chances-of survival for each involved characters (i.e., per each character, higher the percentage, higher the chance of survival), and each alternative led to an expected outcome, which was consistent with one specific AV behavior. The nonutilitarian behavior favored the lowest number of characters (1 out of 4) and the lowest expected utility (Morgenstern and Von Neumann, 1944). In the No Veil scenario, it may could turn out to be also the self-protective option depending on the experimental condition. Oppositely, the utilitarian behavior resulted in the protection of the highest number of characters (3 out of 4) and the highest expected utility. Differently, the maximin behavior resulted to be the option which raised the chance of survival of the character with the highest risk, but distributing it among the other characters, resulting in an expected utility value just below the utilitarian option (Table 13). The three dilemmas were presented in a fixed order, from the full PT trilemma with less situational awareness (Thick Veil scenario) to the partial PT scenario (No Veil) with full disclosure of contextual and personal information. In the first scenario (Thick veil), the chances-of-survival were disjointed from characters' roles (i.e., no label were provided). Roles (i.e., labels) were then disclosed in the Thin Veil trilemma, but the moral agent was not associated to any of the four characters. Finally, the moral agent was assigned to a specific perspective in the No veil scenario, in which contextual and personal information were no longer blurred (Table 13). The chances-of-survival were kept constant throughout the three dilemmas to reduce the risk of confusion.

Table 13: The three outcomes depicting the three potential AV behaviors in the three scenarios (Thick, Thin, No veil). The percentages in each cell indicate the chance-of-survival for each character (columns) in each AV behavior (rows).

Thick Veil Scenario

AV behavior	*Unknown character*	*Unknown character*	*Unknown character*	*Unknown character*
Nonutilitarian	99 %	1 %	1 %	99 %
Utilitarian	1 %	99 %	99 %	99 %
Maximin	42 %	38 %	38 %	90 %

Thin Veil Scenario

AV behavior	*Passenger*	*Pedestrian 1*	*Pedestrian 2*	*Pedestrian 3*
Nonutilitarian	99 %	1 %	1 %	99 %
Utilitarian	1 %	99 %	99 %	99 %
Maximin	42 %	38 %	38 %	90 %

No veil Scenario

AV behavior	*Passenger OR You*	*Pedestrian 1 OR You*	*Pedestrian 2*	*Pedestrian 3*
Nonutilitarian	99 %	1 %	1 %	99 %
Utilitarian	1 %	99 %	99 %	99 %
Maximin	42 %	38 %	38 %	90 %

Notes: The top table was recalled in the Thick veil scenario (in which all the roles/labels in the scene were hidden, both for the self and other characters involved); the central table was presented in the Thin veil scenario (in which the roles/labels were disclosed, but not for the participant); while the bottom table was associated with the No veil scenario (in which all the roles/labels were disclosed, both for the self and other characters involved). The sample was divided in two parts in the last scenario: a group was asked to assume the passenger's perspective (Passenger), and the other the pedestrian's one (Pedestrian 1). The name of the behaviors are here presented for demonstration purposes only, but they were hidden to the participants during the experimental procedure. In each scenario, Behavior 1 corresponded with the nonutilitarian outcome, Behavior 2 with the utilitarian outcome, and Behavior 3 with the maximin outcome.

Additionally, the Interpersonal Reactivity Index (IRI) was administered to assess empathy and sympathetic feelings through a multidimensional approach (Albiero et al., 2006; Davis, 1980; 1983). The Italian validation of the measure disposes of 25 items on a 5-point Likert scale (from 1 = Does not describe me well, to 5 = Describes me very well), grouped into four subscales: Perspective Taking (7 items), Fantasy (5 items), Empathetic Concern (6 items), and Personal Distress (7 items). Higher is the sum score, higher the individual emphatic inclination. Given the importance of sympathetic feelings in understanding morality and harm norms (e.g., Carlo et al., 2010; Irwin, 2013; Maibom, 2009; Pérez-Manrique-Gomila, 2018), Eisenberg et al. (2001) suggested to derive a self-reported sympathetic concern index leveraging on IRI 'subscales of Empathic Concern and Perspective Taking.

6.3.3 Experimental procedure

The experiment was programmed via Qualtrics software and distributed via Prolific platform. The anonymous link provided by Qualtrics was embedded in the platform for reaching the requested sample size, assuming the stratification of the age factor and a balanced number between gender and experimental conditions in the No Veil scenario's partial perspective (AV passenger and pedestrian). The data collection was performed in 2022, from January 16th to January 25th. All the participants were required not to perform the survey through smartphones or tablets, but only using laptops, in order to avoid problems of data comparability among multiple devices (Krebs and Höhne, 2021). To test the comprehensibility of the experimental task, a pilot study was conducted on 10 participants, that confirmed the feasibility of the experimental design and the intelligibility of the experimental material. The experimental procedure is graphically described in Figure 24. The mean completion time of the experimental procedure was 13.36 mins (SD = 6.37). Before the beginning of any experimental activity, the participants were requested to read and fill out an informed consent about their participation and data protection regulation. In order to receive the hourly rate contribution at the end of the experiment, a personal Unicode had to be entered at the beginning of the experimental session, which was provided by Prolific itself. The presentation of the AV trilemmas was anticipated by the collection of socio-demographics, driving habits information and previous knowledge about the AV technology. Additionally, two numerical literacy questions were administered to the sample, controlling for basic knowledge of proportions and percentages. Twelve participants were excluded from the sample on the basis of this control check, since they failed at least one of the two questions.

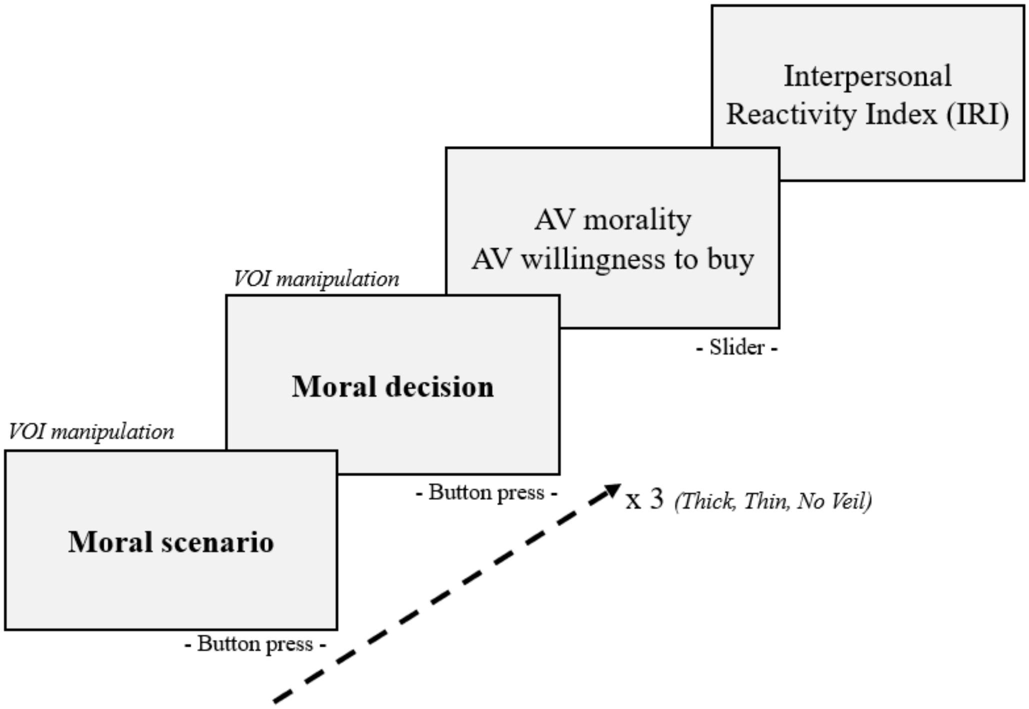


Figure 24: The experimental procedure of Study 4. The three moral scenarios (i.e., trilemmas) were administered in fixed order (Thick Veil, Thin Veil, and No Veil scenario).

At the beginning of the trilemma section, a detailed explanation of the task was provided. Participants were informed about the characteristics of the upcoming hypothetical traffic situation, presented paired with a vignette. In the depicted event, the participants knew that they could have been one of the characters involved in the scene. The moral trilemmas were presented at the center of the screen, in black type (font Arial, size 10) against white background, and participants had unlimited time to read the storylines and watch the vignettes. At this point, the Thick Veil and the Thin Veil scenarios were presented in this order, as the two full PT trilemmas with progressive amount of disclosed information (see the stimuli section for a detailed explanation, and the Appendix to retrieve the stimuli).

After reading each storyline, the three potential AV behaviors in the form of chances-of-survival were presented, asking participants to select the most morally rightful outcome in their opinion. At this point, the sample was randomly divided into two partial PT groups (AV passenger and pedestrian), controlling for gender balance. This decision was taken to investigate potential differences in moral reasoning while assuming two different perspectives. Consequently, the No Veil scenario was presented, with different perspectives considering the experimental condition (see the Appendix). Following the trilemma section, participants were requested to (i) indicate which perspectives they assumed while answering to the full PT scenarios (AV passenger, pedestrian, or eagle-eye), (ii) grade the moral acceptance and (iii) the willingness to buy AVs programmed to follow three different behaviors when reacting to trilemma-like situations (0 = completely unacceptable/ unwilling to buy, 100 = completely acceptable/ willing to buy). Participants evaluated AVs programmed: to prioritize the AV passenger before anyone else (passenger-protective AV, coherent with the nonutilitarian outcome), to prioritize the protection of the highest number of characters (utilitarian AV), and to distribute the risk among the characters involved, with the aim to improve the protection for the most endangered character (maximin AV). Lastly, the Interpersonal Reactivity Index (IRI) was administered to assess empathy.

Table 14: The distribution of the sample between experimental conditions, controlled by gender.

	Female	Male	*N*
Thick Veil scenario	120	119	*239*
Thin Veil scenario	120	119	*239*
No Veil scenario	120	119	*239*
AV Passenger perspective	60	60	*120*
Pedestrian perspective	60	59	*119*

N	*120*	*119*	**239**

6.3.4 Analysis

The statistical analysis was conducted in the R environment (version 4.1.1; R Core Team, 2021). Given the experimental questions, a series of statistical models were implemented. For descriptive purposes, a correlation analysis was computed considering moral acceptability, willingness to buy the automated technology, IRI total sub scores, and the sum of Emphatic Concern and Perspective Taking subscales total scores i.e., 'Sympathetic feelings'). Descriptive information on moral evaluation, willingness to buy are reported in Table 12, and IRI total sub scores are reported in the Appendix (table A4.1), divided by moral consistency profiles. Then, moral decisions to the three AV trilemmas (Nonutilitarian, Utilitarian, Maximin) were assumed as three separate binomial dependent variables in three separate generalized mixed-effects linear models ($M_1 – M_3$), setting the participants as random intercept. Then, two further mixed-effects linear models (M_4, M_5), were implemented for the investigation of potential differences in terms of moral acceptability and willingness to buy AVs programmed to follow the three proposed behaviors (now made explicit). In these latter two models, moral consistency profiles (fully utilitarian, fully maximin, switchers) were considered as potential fixed effect. The fully nonutilitarian profile was excluded from the analysis, considering its scarce numerosity ($n = 3$). The models presented in the main analysis ($M_1 – M_5$) are the result of four corresponding forward stepwise model comparisons, which considered models with a number of different predictors. The chosen one was selected on the basis of the Akaike Weights comparison procedure (Wagenmakers and Farrell, 2004). Post hoc pairwise comparisons were considered when requested, using the R package *emmeans* (Lenth, 2020). Bonferroni correction was set as an adjustment method. The final dataset and further supplemental information are retrievable in the OSF project folder: https://bit.ly/3cksq6Q.

6.4 Results

Figure 25 chromatically describes the correlation between a moral evaluation, willingness to buy and IRI sub scores. Unsurprisingly, positive moderate-to-high correlations were observed between moral evaluation and willingness to buy per each king of AV programming (passenger-protective: *r* = .40; utilitarian: *r* =.39; maximin: *r* = .68). In line with the literature, a negative but low correlation was observed between acceptability of the utilitarian moral code and personal distress (*r* = -.11). Interestingly, lower interest in buying passenger-protective AVs were correlated with higher fantasy (*r* = -.13), emphatic concern (*r* = -.25) and sympathetic feelings (*r* = -.18).

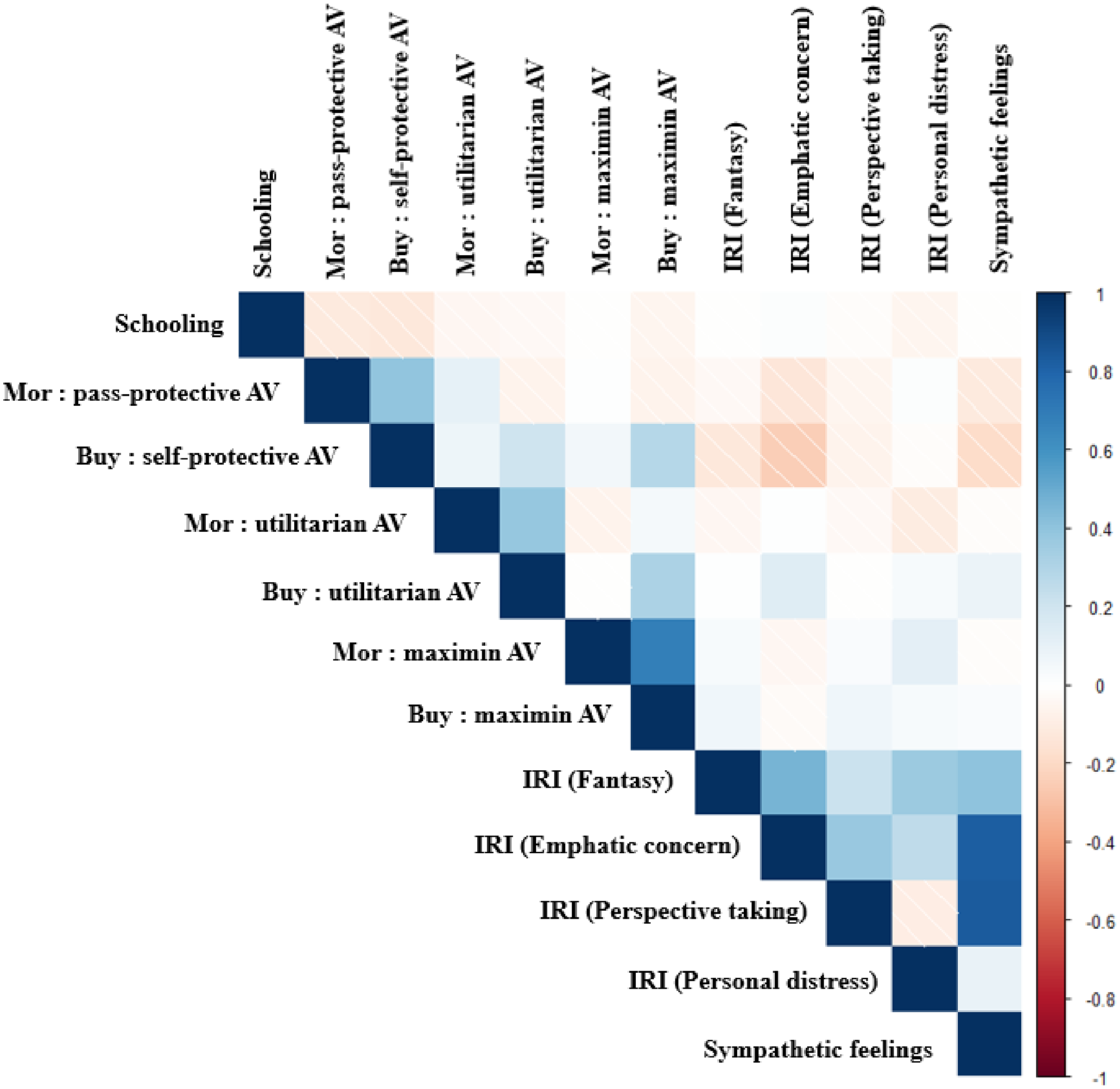


Figure 25: Heat map plot describing the correlations between variables on a chromatic form. for the 7-meters throw. Each cell represents the correlation between the row variable and the column variable. The level of agreement is chromatically represented from dark red (Pearson's r = -1, negative correlation), to white (r = 0, no correlation) and dark blue (r = 1, positive correlation). Significant correlations are reported as diagonally slashed cells.

Table A4.1 (Appendix) summarizes the predictors included in the selected models (M_1- M_5), coherently with the obtained estimates and the corresponding effect sizes. The binomial family distribution was set as a reference point for implementing three generalized linear models (M_1 – M_3), considering the three moral decision (M_1: nonutilitarian, M_2: utilitarian, M_3: maximin) as dependent variables and the participants as random intercepts. Following the evidence obtained from the computed onward stepwise regressions, all the selected models considered the VOI type (Thick veil, Thin Veil, No Veil) as fixed effects, as well as the interaction with the partial PT in the No Veil scenario (AV passenger, pedestrian). Additionally, the models assuming the utilitarian (M_2) and the maximin (M_3) decisions as dependent variables also admitted Gender (female, male) as fixed effects. The percentages of endorsement of the three AV behaviors in each of the VOI scenarios is summarized in Table 15 and described in Figure 26.

Table 15: Percentages of endorsement of the three proposed AV behaviors, divided by VOI type (Thick veil, Thin veil, No veil) and specified – between dotted lines - for the experimental condition in the No veil scenario (AV's passenger, Pedestrian). The overall percentages are presented in the last column.

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

A significant role of VOI type was detected in the endorsement of all the three AV behaviors. Oppositely to our prediction, the utilitarian moral code was mainly preferred in the Thick veil trilemma ($\chi^2{}_2 = 9.75$, $p = .007$) when compared to the Thin veil ($z = 3.45$, $p = .002$) and the No veil scenario ($z = 3.75$, $p < .001$). The inversion of the predicted trend was also observed in the endorsement of the maximin decisional strategy, which was the more frequently selected outcome responding to the Thin veil scenario ($\chi^2{}_2 = 8.31$, $p < .015$) when compared to Thick veil ($z = 2.99$, $p = .008$) and the No veil trilemma ($z = 2.90$, $p = .011$). As expected, a higher preference for the nonutilitarian behavior was observed in the No Veil scenario ($\chi^2{}_2 = 23.59$, $p < .001$), where the moral agents' perspectives were disclosed

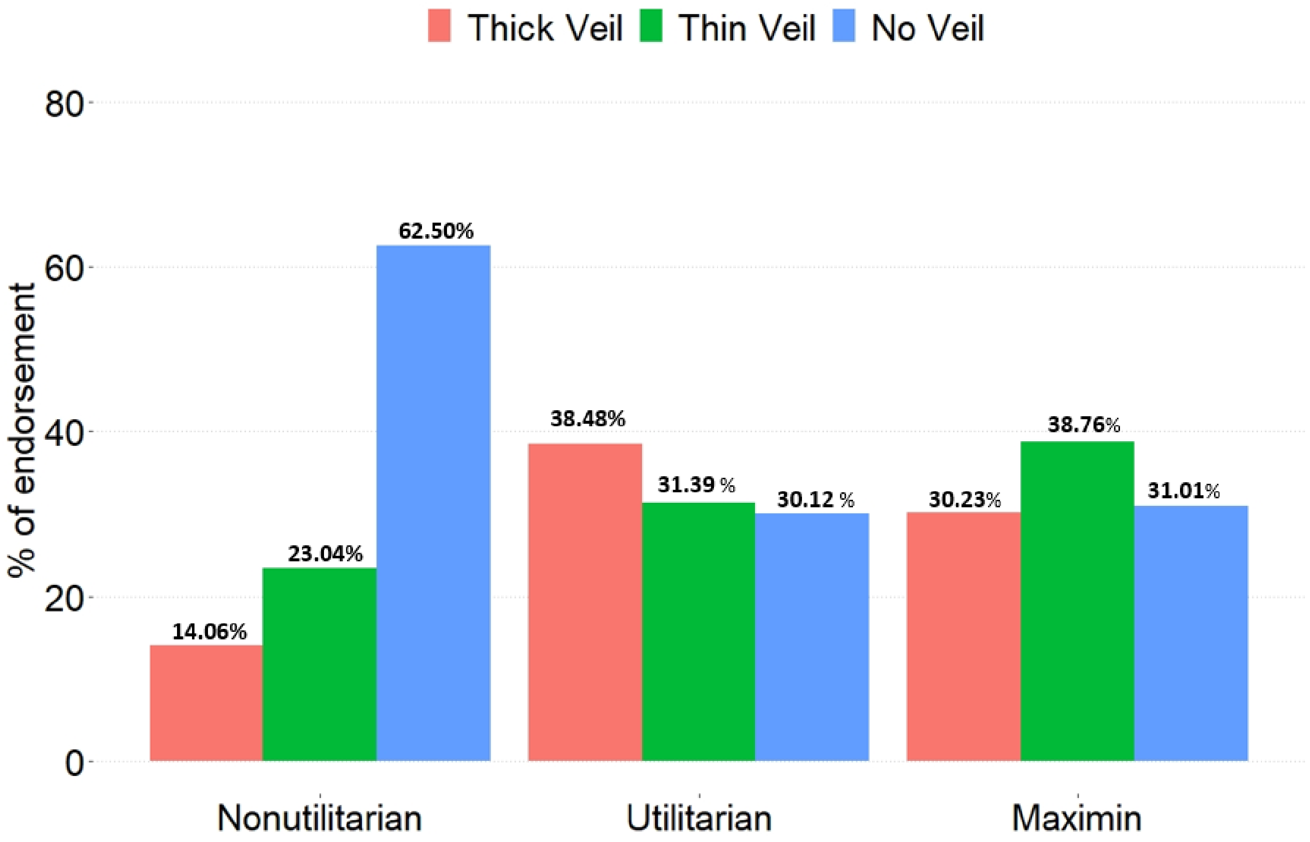


Figure 26: Bar chart on the relative percentage of endorsement of the three proposed AV behaviors (Self-protective, Utilitarian, Maximin), throughout the three types of veils (Thick veil, Thin veil, No veil scenario).

In all the three binomial models, the significant interaction between VOI type and partial PT helped in disclosing the effect of perspective taking on the endorsement of each AV behavior (Figure 27). In fact, the preference for the nonutilitarian and utilitarian behaviors in the No Veil scenario was well explained by the disclosure of the self-protective feature in the two options. The tendency for favoring the nonutilitarian AV behavior in the No Veil scenario was a specifical characteristic of the passenger perspective condition ($\chi^2{}_3 = 13.20$, $p = .004$), in which the endorsement of the nonutilitarian outcome led to the protection of the moral agent more than in the Thick (z = 5.11, $p < .001$) and Thin scenario (z = 4.78, $p < .001$). Despite this effect, the nonutilitarian option was assumingly the least preferred behavior for an AV, confirming the overall favor for the utilitarian resolution in moral judgment (Table 15). This was also the case of the No Veil scenario, mainly thanks to the contribution of the pedestrian perspective condition ($\chi^2{}_3 = 45.44$, $p < .001$; z = 6.35, $p < .001$), in which the utilitarian option also turned out to be the self-protective one. Interestingly, when all the information were disclosed, the maximin decisional strategy appeared to be highly preferred by moral agents' in the perspective of the pedestrian than of the AV passenger ($\chi^2{}_3 = 17.08$, $p < .001$; z = 3.64, $p = .004$). As expected, no differences were detected between partial PTs in the Thick and Thin scenarios, since the experimental condition was revealed only in the last No Veil trilemma. Additionally, a significant overall effect of gender was detected in the endorsement of the utilitarian outcome, with a favor towards men (45.00% Vs. 26.89%; $\chi^2{}_1 = 9.25$, $p = .002$), as well as in the endorsement of the maximin strategy, with a favor towards women (62.18%% Vs. 48.05%; $\chi^2{}_1 = 13.78$, $p < .001$).

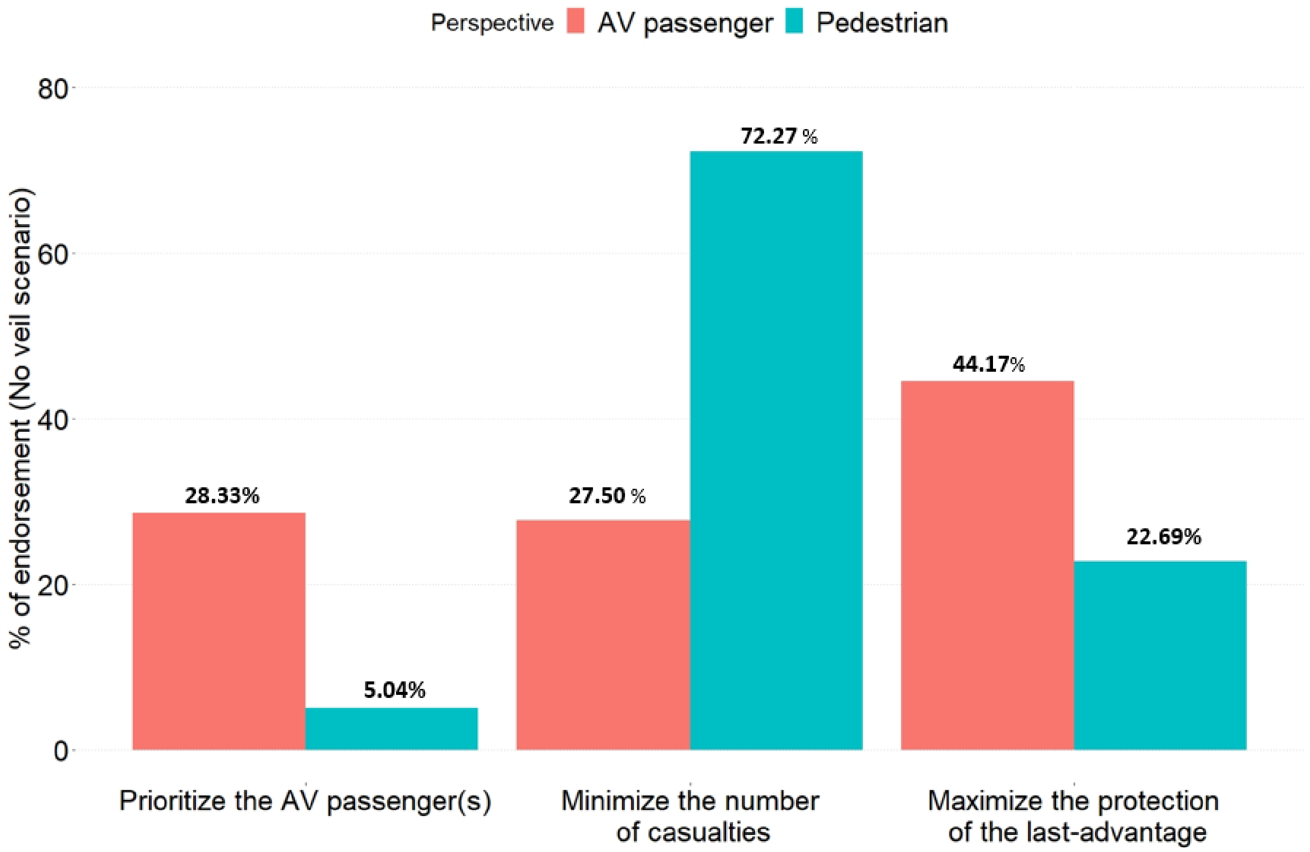


Figure 27: Bar chart on the total percentage of endorsement of the three proposed AV behaviors when the self-protective feature was revealed in the No veil scenario (Prioritize the AV passenger/Nonutilitarian, Minimize the number of casualties/Utilitarian, Maximin), divided by experimental condition (AV's passenger perspective, pedestrian perspective).

Finally, two mixed effects linear models were implemented for the investigation of moral acceptability (M_4; 0-100) and willingness to buy AVs programmed to follow the three described behaviors (M_5; 0-100). On the basis of the evidence collected from the stepwise regressions, the final models considered the AV Behavior (Prioritize the AV passenger, Minimize the number of casualties, Maximize the protection of the last-advantaged) and the Moral consistency profile (Fully utilitarian, Fully maximin, Switchers) as fixed effects, as well as their interaction. The utilitarian AV behavior was evaluated as the greatest moral resolution (M_4; $\chi^2_2 = 342.63$, $p < .001$; Figure 28) both when compared to the passenger-protective ($z = 4.75$, $p < .001$), and when compared to the minimax approach ($z = 6.54$,

$p < .001$). Overall, Switchers evaluated the three AV behaviors more morally than fully utilitarian (M_4; $\chi^2_2 = 17.65$, $p < .001$; z = 3.94, $p < .001$) and fully maximin individuals (z = 2.75, $p = .019$). In terms of moral acceptability, a significant interaction effect was revealed (M_4; $\chi^2_4 = 67.32$, $p < .001$). Mean scores are retrievable in Table 16. As expected, fully maximin individuals had a greater moral evaluation of the maximin AV behavior when compared to the fully utilitarian agents (z = 4.10, $p = .001$). Inversely, individuals with an utilitarian profile had a greater endorsement for the utilitarian AV behavior when compared to maximin individuals (z = 6.25, $p < .001$). Throughout the moral profiles, the passenger-protective behavior was consistently evaluated as less moral than the utilitarian and the maximin approaches. Seemingly, the utilitarian AV behavior was perceived as more moral than the maximin one, except from fully maximin agents (z = 1.21, $p = 1.00$).

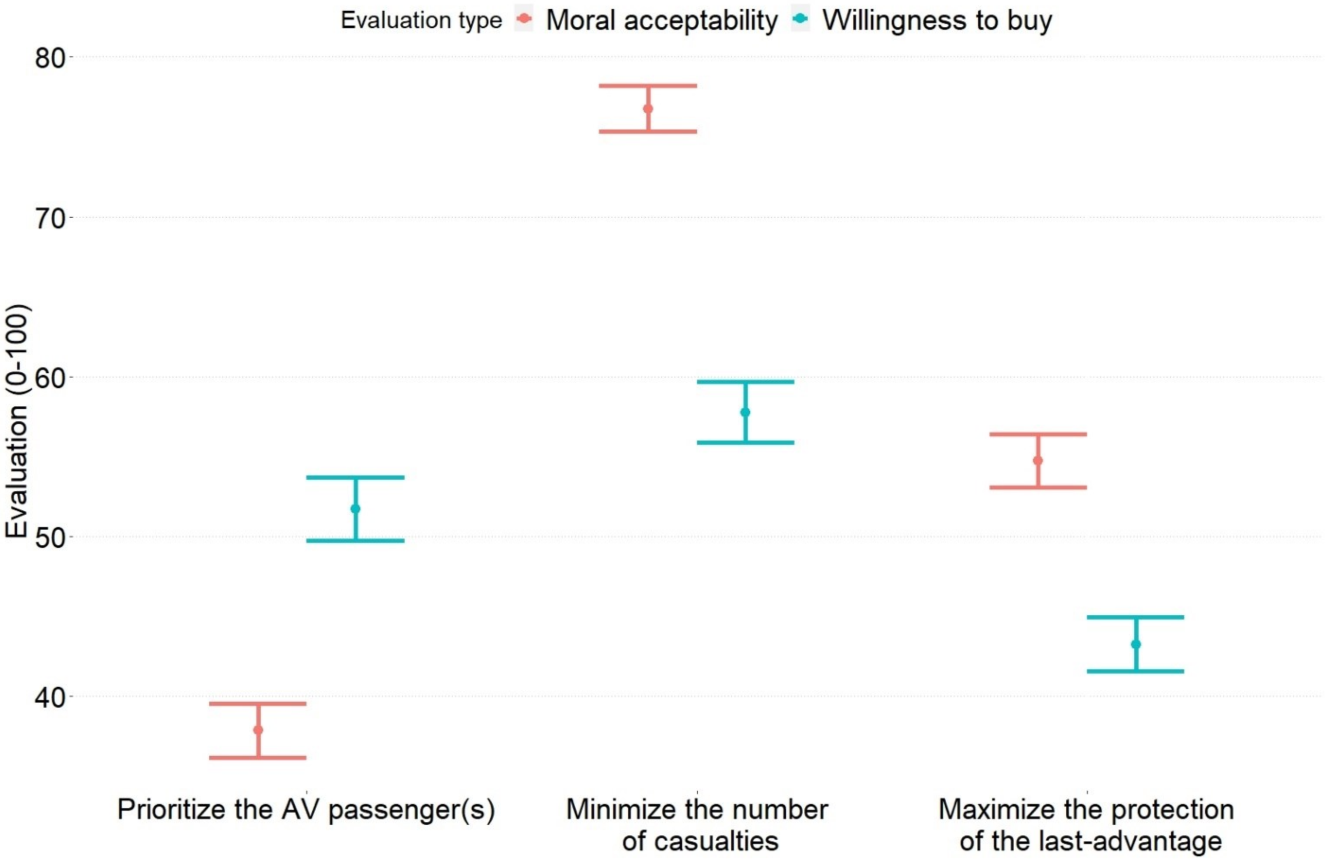


Figure 28: Error bars plot representing means and standard errors of the evaluations of moral acceptability and willingness to buy for the three proposed AV behaviors (Prioritize the AV passenger/Nonutilitarian, Minimize the number of casualties/Utilitarian, Maximin).

The advantage of the utilitarian AV behavior was also confirmed in terms of willingness to buy (M_5; $\chi^2{}_2 = 43.19$, $p < .001$; Figure 28), but only in the comparison between the utilitarian and the maximin strategies ($z = 2.70$, $p = .021$), while no differences were observed on the availability to purchase a utilitarian or a passenger-protective vehicle. Switchers were also more willing to purchase AVs than fully maximin individuals (M_5; $\chi^2{}_2 = 7.37$, $p = .025$; $z = 2.68$, $p = .024$), but not more than fully utilitarian moral agents ($z = 0.53$, $p = 1.00$). The interaction effect between AV behavior and Moral consistency profile was also confirmed in M_5 ($\chi^2{}_4 = 35.00$, $p < .001$). Fully maximin individuals had no preference between AVs algorithms in terms of purchasing availability, while – as expected - the utilitarian AV behavior was preferred from fully utilitarian to the nonutilitarian ($z = 7.87$, $p < .001$) and to the maximin algorithm ($z = 4.44$, $p < .001$). Switchers had no specific preference for utilitarian AVs but preferred the passenger-protective vehicle to the maximin one ($z = 3.37$, $p = .029$, see Table 16).

Table 16: Mean and Standard Deviation of evaluations of moral acceptability (0 = completely unacceptable) and willingness to buy (0 = unwilling to buy), divided by the three main consistency profiles (Consistent Utilitarian, Consistent Maximin, Inconsistent). The 'Consistent nonutilitarian profile was not considered because of its scarce numerosity (n = 3). The overall information is presented in the last row.

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

6.5 Discussion

Study 4 aimed to investigate the role of perspective-taking on moral reasoning in the evaluation of AV moral behaviors. To fulfill this goal, we decided to operationalize Rawlsian and Harsanyian theoretical Veil of Ignorance (VOI) settings, shaping it in the form of three moral trilemmas. Our goal was to observe if the availability of a crescent amount of contextual and personal information may significantly affect the endorsement of a particular AV moral code. On the basis of the reference theories (Harsanyi, 1975; 1978; Rawls, 1971/2009), the Rawslian ‘No knowledge’ VOI formula should have led to the application of a different decision strategy than the Harsanyian ‘Equal Chance’ VOI formula (Parfit, 2011). In fact, a stronger endorsement of the maximin strategy was expected behind the so-called Thick Veil, where moral agents had no information about themselves and the other characters involved in the scenario. Inversely, a stronger endorsement of the utilitarian moral code was expected behind the so-called Thin Veil, where moral agents were yet not aware of their own role in the critical situation, but each role was now associated – through apposite labels - with the respective chance-of-survival. Responding to these two full perspective-taking (PT) VOI trilemmas, moral agents indeed recalled the opposite decisional strategy of what was expected, contrarily to our research hypothesis. In fact, a stronger endorsement of the utilitarian moral code was detected behind the Rawslian Thick Veil, while the maximin criterion was highly selected when responding to the Thin Veil form of the AV dilemma. This solution appears to aim to a more ‘democratic’ distribution of risk among all the characters involved, when compared to the more ‘economical’ utilitarian resolution for the minimization of the total number of casualties. Consistently with the present results, we can deduce that the addition of the individual labels (i.e., roles) as the simplest contextual information in the VOI trilemma has the role to improve the likelihood of the maximin moral reasoning, towards a more distributive approach to risk management and – specifically - towards morality of AV behaviors when facing critical road events. Instead, the lower likelihood of the maximin decisional rule in the Thick veil scenario (32.63% Vs. 63.61%) may be explained by Moehler (2018), who suggests that - behind the Rawlsian VOI - rational agents may act to maximize the expected individual utility (Briggs, 2014), independently of their social positions. Overall, this trend is anyway consistent with the observed preference for the utilitarian AV behavior throughout the three trilemmas, confirming the hypothesis of a general advantage of the utilitarian moral code when (i) a limited number of personal features are available in a full PT moral dilemma (Huang et al, 2019;

Martin et al., 2021a; 2021b), and when (ii) the comprehensibility of the utilitarian moral behavior is highly accessible (Kusev et al., 2016). Nonetheless, the utilitarian endorsement seems to be negatively affected by the amount of contextual and personal information, and in this context by the specification of each character's traffic role. This result deserves further investigation, by focusing – for example – on other features that may be progressively disclosed throughout the experimental approach (e.g., gender, age, social position, negligence; Awad et al., 2018a), and also stressing moral agent's driving habits.

As expected, a growing interest in the nonutilitarian AV behavior was observed when all the contextual and personal information were disclosed in the No Veil trilemma. In this specific situation, the nonutilitarian resolution allowed the protection of the AV passenger, while with the utilitarian option the moral agent had the opportunity to protect the three pedestrians. The incremented likelihood of the nonutilitarian behavior confirmed the effectiveness of the implemented experimental manipulation, bringing new evidence on the significant role of partial PT when reasoning about morality of AVs. In fact, considering the moral decisions taken behind the Thick and Thin Veils, we can assume that the growing likelihood of the nonutilitarian resolution was not justified by a proper favour towards this moral code, but rather by the disclosure of moral agents' perspectives in the traffic scenario. Also 'outside' the VOI, minimizing the total number of casualties was still the preferred outcome to pursue, but here participants were awarded with a specific perspective (AV passenger or pedestrian), being finally able to discern their individual advantage from the general collective goal. Consistently, the allocation of a specific (i.e., partial) perspective revealed the tendency towards self-protective behaviors, both in the pedestrian perspective and mainly in the AV passenger perspective, and despite the confirmation of the overall low attractiveness of nonutilitarian option. The weight of PT on moral reasoning is consistent with Martin et al. (2021a): if full PT accessibility (in which the moral agent could play the role of each of the characters involved with equal probability) has the ability to boost the likelihood of the utilitarian behavior, assuming a partial PT clearly reduce the interest towards this moral code. Martin et al. (2021a) opted to investigate the sole AV passenger partial PT, highlighting a relevant detrimental effect of induced personal risk on the endorsement of the utilitarian resolution. Study 4 underlined that this effect is leaded by the search for self-protection, revealing how the proportion of preference is affected by the assumed perspective. Interestingly, when individuals from the AV passenger perspective were not willing to endorse the 'selfish' nonutilitarian option for pursuing their own self-protection, they opted for the maximin AV behavior more than the utilitarian one, still improving their chance to survive. Further studies may stress the role of perspective-taking in the resolution of moral trilemmas, especially in the direction of self-protecting behaviors. In this sense, the moral trilemma appears to be a useful experimental tool. Di Nucci (2013) empirically tested Thomson's interpretation (2008) on how presenting a three-options trilemma may affect the proportion of utilitarian endorsement in the traditional binary Switch problem. Despite some potential criticisms about the employed methodology, Di Nucci detected a reduced favor toward the utilitarian moral code when the self-sacrificial trilemma was presented before the traditional non-sacrificial binary dilemma. Possibly, the trilemma may have shifted the respondents' moral reasoning towards a less individualistic approach, easing their identification with the sacrificed character on the secondary railway, and so encouraging the impermissibility of the utilitarian option (see Di Nucci, 2013). In this study, this can be a potential explanation for the detected preference towards a more distributive approach to risk (i.e., the maximin strategy) over the utilitarian moral code, specifically when embracing the AV passenger partial perspective. Importantly, the personal involvement effect has been widely described in decision-making processes (Greene et al, 2001; Kusev et al., 2016; Lotto et al., 2014). When moral agents are directly involved in the dilemma (e.g., the *Push dilemma*, Thomson, 1985), the moral problem is perceived as more emotionally salient and cognitively demanding, negatively affecting the endorsement of the utilitarian resolution (Moore, 2008). Overall, this study confirmed that partial PT and self-involvement in sacrificial scenarios affect moral reasoning, enhancing self-protective behaviors when compared to full PT 'behind' the VOI. We believe the evidence collected in Study 4 will be helpful for obtaining a more precise description of morality towards autonomous transportation, considering that the perspective assumed may considerably affect the way through which people perceive the implementation of this new technology.

Finally, and consistently with evidence collected from Study 3, we expected (i) evaluation of moral acceptability and (ii) willingness to buy AVs programmed to follow the three described behaviors to be coherent with the derived moral consistency profiles. Three moral profiles were obtained from the expressed moral decisions, describing moral agents: who consistently followed the utilitarian moral code (i.e., fully utilitarian, 34.72%), who consistently followed the maximin strategy (i.e., fully maximin, 18.83%), and who changed their decisional rule at least one time throughout the experiment (i.e., switchers, n = 45.19%). As expected, fully utilitarian and fully maximin individuals evaluated the correspondent AV behaviors as more acceptable, respectively to minimize the number of total casualties and to maximize the protection of the last-advantage person. Consistently with Study 3 and with previous literature (Awad et al., 2018a; Bonnefon et al., 2016; Martin et al., 2021a), the utilitarian approach was coherently evaluated as the more morally acceptable decisional rule from fully utilitarian moral agents, as well as from switchers individuals. Coherently, the fully maximin group rated the maximin strategy as more moral than the fully utilitarian group, despite fully maximin individuals evaluated the maximin and the utilitarian resolution as equally moral. This trend appears to be in accordance with the willingness to buy AVs, with fully utilitarian agents who expressed a higher interest in utilitarian AVs when compared to passenger-protective and maximin vehicles but preferring a passenger-protective vehicle rather than a maximin one. This trend was confirmed in the switcher moral group, but overall, the willingness to share the autonomous technology was lower than the correspondent moral evaluation. Considering the gap between moral evaluation and availability to purchase AVs, this result confirms the 'dilemma of self-driving cars' discovered by Bonnefon et al. (2016). Nonetheless, an AV programmed to minimize the number of casualties was not only the moral landmark for fully utilitarian moral agents, but also their preferred moral code in case of future purchasing (Table 16). In this sense, Martin et al. (2021a; 2021b) suggested that assuming a full perspective on the moral problem leads to reduce the observed distance between moral evaluation and willingness to share a utilitarian AV, easing the dilemma observed in previous studies (Bonnefon et al., 2016). Future investigations on the topic may focus on the advantage of controlling moral reasoning with moral profiles, to better describe individuals' attitudes towards features of autonomous transportation and of AV behaviors (e.g., through the administration of the Oxford Utilitarianism Scale; Kahane et al., 2018). The present study showed the importance of take in consideration individuals' decisional profile and moral consistency throughout the experimental session, as the interpretation of moral acceptability and the shareability of the utilitarian AV seem to be conditioned by these profiles.

In the present study, the operationalization of Rawslian and Harsanyian VOIs may seem rather simplistic, as it condense two complex and tangled social and political theories in a controlled experimental setting. Furthermore, the Rawls Vs. Harsanyi dispute towards social contractualism and the fairest decisional strategy still have a relevant impact on political philosophy (e.g., Frohlich et al., 1987; Gaus and Thrasher, 2015; Moehler, 2018). In this sense, is it worth to point out that this experimental application does not aim at solving the remaining discrepancies between the interpretations of the two theories, but the VOI construct has served as an inspiration to deepen the impact of perspective-taking and information accessibility in the evaluation of morality of AV behaviors. Additionally, in order to properly distinguish the employment of different decisional strategies (representing three different way of AV behaviors), the trilemma version of the traditional the Switch moral dilemma was selected as an experimental stimulus (Di Nucci, 2013; Thomson, 2008). Up to date, the majority of the evidence collected on the moral judgment and social perception of autonomous transportation has been collected through traditional binary moral scenarios (Bonnefon et al., 2016; Gill, 2020; Huang et al., 2019). We believe that future studies may continue to investigate the usefulness of this experimental stimulus in the field of AI ethics, which could allow to overcome a series of limitations regarding the stark distinction between the utilitarian and nonutilitarian decisional criteria (e.g., Evans et al., 2020; Rhim et al., 2021). Finally, opting for the operationalization of the VOI theoretical environment, we decided to refer to a numerical representation of chances-of-survival to distinguish the three decisional strategies. The use of percentages in risk communication has been widely discussed (e.g., Gigerenzer et al., 2007; Peters et al., 2011), we carefully controlled the experimental material following the suggestions described in the literature, but potential limitations have to be taken into account. Through the development of a series of partial PT studies, De Melo et al. (2021) revealed how the perceived risk of the situation can moderate moral decisions. Further studies may opt for manipulate the numerical risk, revealing for example the presence of potential numerical thresholds between decisional strategies that may switch the moral decision, or even propose a different but still effective method to describe the VOI's features without the numerical representation.

In conclusion, evidence collected from Study 4 shed new light on the significant role of perspective-taking on moral reasoning, in the evaluation of AV moral behaviors. With this study, we presented for the first-time a within-subjects approach to VOI moral reasoning, distinguishing the features of the VOI setting following the theories of John Rawls and John Harsanyi (Harsanyi, 1975; 1978; Rawls, 1971/2009). Results revealed the use of different – albeit somewhat unpredicted – decisional rules on the basis of a different number of contextual and personal information. The definition of three moral consistency profiles also allowed us to continue the investigation already began in Study 3 on moral evaluation and willingness to share the autonomous technology, highlighting the usefulness of moral profiling in better understand the individual interest towards AVs, especially when investigated through moral dilemmas-experimental approaches.

Chapter 7

7.1 General discussion and future perspectives

Technological innovation can have profound effects on moral systems. Describing the impact of technical revolution during the medieval age, White (1962) underlined the role of basic innovations (e.g., the stirrup) as technological facilitators of the development of new social systems and moral norms, leading to produce new effective responses to societal challenges. The history of morality in human values systems has been shaped by economic and technological progress as well as by collective beliefs (Harari, 2016), best representing the link between the individual and society itself (Rokeach, 1973). In this framework, the current introduction of new AI technologies surely has a transformative effect on reshaping social values and moral norms, requiring for a redefinition of social perspectives and moral attitudes towards these novelties (Danaher and Skaug Sætra, 2022; Klenk, 2022; van de Poel and Kudina, 2022). In the field of mobility, the optimization of the revolutionary autonomous transportation systems can be eligibly acknowledged as one of the main technological challenges that upcoming societies will face in the next decades, reconsidering driving operations as an exclusive responsibility of intelligent systems (Fagnant and Kockelman, 2015; Gilbert, 2012). Nevertheless, positive or negative beliefs towards this new technology will have the potential to boost or decelerate the implementation of autonomous vehicles (AVs), being fundamentally related with the ways through which future users (i.e., stakeholders) will evaluate AVs' behaviors when addressing driving situations with a given risk entity. In the last decade, literature has focused on exploring and assessing preference towards autonomous transportation, focusing on the social implications of this technology (e.g., Gruel and Stanford, 2016; Othman, 2022; Moody et al., 2020), as well as on ethical aspects, individual perception and moral evaluation of AVs behaviors, mainly when facing unpredicted and critical situations (e.g., Bonnefon et al., 2016; Gill, 2021; Li et al., 2016; Robinson et al., 2021; Shariff et al., 2017). Considering the future impact of the autonomous driving revolution on social systems and moral values, empirical investigations in this direction has a great potential, since they can offer a comprehensive overview on general attitudes towards AVs, also highlighting specific features that play a significant role in shaping moral perception of intelligent transportation systems. In this framework, the present thesis aimed at shading new light on moral judgment towards AVs from a cognitive, emotional, and social perspectives, especially focusing on one of the most employed experimental tools in this field, namely the sacrificial moral dilemma (Awad et al., 2018a; Bonnefon et al., 2016; Frank et al., 2019).

The reliability of the tool in the context of AVs

As a first main objective, we aimed at deepening the reliability of the sacrificial dilemma as the elective experimental tool in the investigation of AV morality. In our vision, this topic has been widely underestimated in previous comprehensive discussions (Bonnefon et al., 2016), despite this step seems fundamental, before interpreting moral attitudes derived from AV dilemmas on the basis of the traditional moral framework. In this sense, Study 1 took a step backwards in terms of technological innovation, aiming at describing potential effect of the traditional human-driving (i.e., nonautonomous) context in sacrificial and incidental dilemmas. Subsequently, Study 2 focused on the comparison between nonautonomous and autonomous driving vehicles. In both the studies, our attention was focused on traditionally used indices to evaluate emotional reasoning (Bonnefon et al., 2006¸ Greene et al., 2001; 2004; Lotto et al., 2014; Palmiotti et al., 2020; Pletti et al., 2016), such as decisional times, moral judgment, moral acceptability of the AV behaviors, as well as self-referred emotional evaluations (Study 1: valence and arousal; Study 2: moral emotions). Overall, moral reasoning did not seem to be heavily affected by the driving context, corroborating the hypothesis of the ‘structure-based’ interpretation of moral dilemmas (Schein, 2020). In this sense, when moral storylines are controlled for a number of substantial features (e.g., number of casualties, affective relationships), their structure affects moral reasoning more than its specific context. In our view, this result is a further confirmation of the reliability of the experimental conclusions collected until today in the investigation of AV dilemmas (Awad et al., 2018a; Bonnefon et al., 2016), therefore endorsing the sacrificial moral dilemma as a reliable and generalizable tool in the investigation of AV morality. Nonetheless, results allows us to bring additional and new evidence on this topic, in general on the sacrificial dilemma tool, and specifically on the role played by automation on moral reasoning.

Addressing the considered variables, responses were faster when selecting the utilitarian option, and a higher proportion in the endorsement of the utilitarian resolution was observed when answering to human-driving scenarios. From our perspective, picturing human-driving situation is an easier task for moral agents, when compared with traditional nondriving (e.g., the "Crying baby" dilemma, Greene et al., 2001) or futuristic autonomous driving scenarios (Bonnefon et al., 2016). When moral judgments are applied to conceivable events, it is possible that cognitive demands decrease (Conway and Gawronski, 2013; Schein, 2020; Sütfeld et al., 2019), somewhat simplifying the decision process and allowing moral decision to emerge faster and sharper, in favor of the utilitarian resolution for the attainment of the greater collective goal. In the assumed reference population, the plausibility of AV storylines may still appear as scarce, suggesting the possibility to add the consideration of AI literacy in the evaluation of AV morality (Long and Magerko, 2020; Luccioni and Bengio, 2020). Future studies could take this factor into account, considering the possible detrimental effect of implausible storylines on the endorsement of the utilitarian moral code (Körner et al., 2019).

A steady reduction of decision times were also observed throughout the studies and regardless the investigated dilemma categories (nondriving, manual driving, autonomous driving). In our view, this evidence may encourage a new methodological discussion on iterated moral dilemma investigations, especially on the proper number of stimuli that has to be administered in order to describe the individual's moral inclination in the most precise way. This appears as a problem of statistical power (e.g., Baker et al., 2020; Lerche et al., 2017), that can take interesting cues from human behaviors in repeated economic games (e.g., one-shot Vs. iterated prisoner dilemma, Colman et al., 2018; Darwen and Yao, 1993). In fact, investigating cooperative behaviors, the use of single trial does not give the possibility to fairly generalize results or define a behavioral strategy (Callaway et al., 2022; Raihani and Bshary, 2011). Therefore, iterated methodologies are typically used in moral psychology, but resulting in a wide variability in the number of dilemmas between applications. Theoretically, a high number of stimuli would not be a problem, since it allows us to get closer to the true investigated parameter. Nonetheless, the specific nature of textual moral problems (e.g., time-consuming, cognitively demanding) may have a detrimental effect, distorting the parameter estimation in time and consequently worsening the determination of individual's moral inclination. This distortion could be related to several influential factors (e.g., cognitive effort, boredom, etc.), that are typically underestimated in the composition of the dilemma set. Future investigation need to stress this methodological question, to detect the best possible trade-off between statistical requests and applicative limitations to investigate moral reasoning in a more reliable way.

Other important insights have been collected from Study 2, Study 3 and Study 4 on the role of specific moral dilemma's features in shaping moral judgment towards autonomous transportation. In Study 2 we focused on the role of self-sacrifice framing on the proportion of endorsement of the utilitarian moral code in human-driving and AV dilemmas. In self-involvement dilemmas, traditional nondriving storylines typically frame the self-protective option within the utilitarian behavior, enhancing the likelihood of the utilitarian endorsement since it allows to protect both the self and the higher number of characters (Greene et al., 2001; 2004; Lotto et al., 2014; Moore et al., 2008). This is not the case in the AV dilemma, where pursuing the greater collective goal is conditioned to the acceptance of your own self-sacrifice (Bonnefon et al., 2016). Expectably, results showed that the endorsement of the utilitarian behavior was negatively affected by self-sacrifice, highlighting the impact of this factor on the moral scenario. Despite this preference and consistently with the literature (Sachdeva et al., 2015), individual placed greater moral value in the self-sacrificial behavior. This evidence reveal the need to carefully consider this feature when developing sacrificial AV dilemmas, since it can significantly affect moral reasoning acting on the proportion of agreement with the utilitarian resolution.

In Study 4 two important characteristics of AV dilemma were stressed, namely, the amount of disposable contextual and personal information and the perspective assumed by the moral agent when judging the AV behavior. The role of these features was deepened operationalizing the so-called Veil Of Ignorance (VOI), as a hypothetical setting differently described by Rawls (*Think Veil,* 1971/2009) and Harsanyi (*Thin Veil*, 1975; 1978). The authors suggested that, when moral agent assume a full Perspective-Taking (PT) and when a number of contextual and personal information are concealed in the decision process, the selected decisional strategy would pursue the fairest behavior for the greatest possible collective benefit. In Rawls' view, the fairest and proper behavior would follow the maximin strategy, while Harsanyian interpretation endorsed the utilitarian code. Oppositely, when moral agents have the chance to assume a specific perspective in the scene (partial PT), the self-protective behavior will improve its likelihood of being selected, despite its moral framing (within the utilitarian or nonutilitarian behavior). The research hypothesis were answered developing a novel methodological approach to VOI reasoning investigation (Huang et al., 2019; Martin et al., 2021a), as a funnel within-subject design to PT and information availability, readapting the moral trilemma structure in the context of autonomous transportation (Thomson, 2008). Results showed an improved endorsement of the utilitarian AV behavior when moral agents possessed only basic information about the event, while the disposal of contextual (Thin Veil) and personal (No Veil) information readjust the proportion of endorsement towards different decision strategies (maximin and self-protective, respectively). Most of the previous investigations depicted the individual as the AV passenger (e.g., Bonnefon et al., 2016), but studies on VOI and PT demonstrated the significant role of partial PT on moral judgment (Huang et al., 2019; Kallioinen et al., 2019; Martin et al., 2021a). Despite the overall preference for the utilitarian moral code, Study 4 highlighted the importance to be reckoned of the perspective assumed in the development of moral reasoning, since it significantly change the perception of AV behavior and of its impact on the collectivity. Further investigations on the topic may also take a step forward, deepening into the complexity of moral motivations underlying moral decisions (e.g., short open-ended questions), which may potentially reveal interesting new aspects in the investigation of moral reasoning and moral behavior.

Additionally, Study 3 investigated how the available decisional time shapes moral judgement, individually and in interaction with individual prosocial orientation. The inspiration came from research on the effect of time superimposition on moral judgment (Rosas and Aguilar-Pardo, 2020; Suter and Hertwig, 2011; Tinghög, 2016), and on the potential contribution of cooperativeness endorsing the utilitarian moral behavior (Goeschl and Lohse, 2018; Rand et al., 2012; 2014; Tinghög, 2014). A wide variety of interpretation on the role of time and social behaviors were suggested by the authors: some authors believes that time pressure (Rosas and Aguilar-Pardo, 2020) and prosocial orientation (Rand et al., 2012; 2014) should stimulate utilitarianism and cooperativeness, whereas other authors argued for a positive effect of time availability on utilitarian/prosocial behaviors (Frank et al., 2019; Goeschl and Lohse, 2018; Greene et al., 2001; 2004; Suter and Hertwig, 2011). Results suggest that moral judgment towards AV behaviors seems immune both from intuitive and deliberative processing (Tinghög et al., 2016), and no differences between 'prosocial' and 'proself' individuals (Murphy et al., 2011) were detected in the endorsement of the greater goal-solution (i.e., utilitarian; Tinghög et al., 2013). Nonetheless, the role of time constraint is widely discussed and undoubtedly deserves further deepening. In this sense, descriptive information collected from Study 4 allows us to leave room for future investigation on the topic. In fact, future studies may focus on the observed tendency towards utilitarianism in the time delay condition (Frank et al., 2019; Goeschl and Lohse, 2018; Suter and Hertwig, 2011), and especially on the hypothesized reduction of moral conflict throughout the experimental session, consistently with the discussed need for a proper definition of the total number of stimuli in iterated moral dilemma designs. Coherently, also the role of prosocial orientation in moral reasoning deserves further exploration, considering the descriptive rising of utilitarian endorsement in the three experimental conditions (Figure 21). Future investigations may consider additional individuals' characteristics to improve social profiling, stressing for example the role of personal advantages resulting from competitive or cooperative behaviors in the workplace (Rand et al., 2012; 2014).

The emotional perspective

Within the first two studies, the emotional dimension was also thorough, combining the investigation of moral reasoning with the evaluation of self-reported emotional activation (Study 1) and self-referred moral emotion's perception (Study 2). The role of emotions has been widely deepened in the field of morality (Greene et al., 2001, 2004, 2008), recognizing its significant contribution in the development of moral judgment with several different theories (see sections 2.1.1 and 2.1.2). On a number of occasions, iterated moral dilemma studies have deepened emotional activation and emotional valence through self-assessment techniques (Lotto et al., 2014; Pletti et al., 2016; Sarlo et al., 2012), detecting higher arousal levels in the endorsement of the utilitarian (consistently with Greene's Dual Process Theory, DPT) and a significant role of self-involvement on emotional state. Inversely, to our knowledge no iterated moral dilemma studies have ever focused on the distinction between other (anger and disgust) and self-referred (shame and guilt) moral emotions (Huebner et al., 2009).

In Study 1 we aimed to confirm that evidence collected on moral reasoning through incidental nondriving dilemmas could be also generalized in the context of driving scenarios. Apparently, the structure-based interpretation of moral dilemmas was also endorsed from the emotional perspective, with no differences among moral contexts and an overall stronger emotional activation when selecting the nonutilitarian behaviors. Interestingly, results did not support the stronger emotional activation described in the literature in case of self-involvement scenarios (Moore et al., 2008; Lotto et al., 2014). We believed that the involvement factor deserved further deepening, since it appears to be even more critical behind the wheel of a traditional nonautonomous vehicle or as the passenger of an AV. With this aim, in Study 2 we investigated the potential role of sacrifice framing in shaping moral reasoning towards autonomous and nonautonomous vehicles, as well as in the related emotional experience. Results were consistent with Study 1, as nonutilitarian decisions triggered emotional activations more effectively, and especially in terms of self-conscious moral emotions (i.e., shame and guilt). Specific activation patterns for the four investigated moral emotions were observed between different levels of automation and sacrifice framing, confirming the sensitivity of this indices in describing emotional experience facing driving moral events (Greenbaum et al., 2020; Haidt, 2003b). Importantly, the two studies converged on the faster selection of the utilitarian behavior in terms of decisional times. This result was unexpected, since Greene's DPT (2009) described this behavior as controlled by a more pragmatic and deliberative process, in opposition with the more emotionally driven intuitive approach. Overall, the observed discrepancy between decisional times and self-reported emotional activation highlight a potential limitation in investigating emotional experience via online survey experiment. In fact, this technique has been equated to lab-based methodologies in a wide range of studies (Bridges et al., 2020; Buso et al., 2021; Dandurand et al., 2008). Additionally, Qualtrics has been recognized to be a valid method for collecting generalizable and reliable online data (Belliveau et al., 2022; Roulin, 2015). In our view, it is possible that in this particular case – so developing an iterated textual dilemma-based methodology - the online survey approach resulted to be less capable to stimulate the moral conflict between two opposed moral behaviors, and consequently being less sensitive in detecting fast and non-discrete emotional variations (e.g., arousal and valence). In conclusion, assuming the described value of considering self-reported emotional evaluations in the investigation of moral judgment towards AV behavior, future studies may aim to compare these findings with new evidence collected from different methodological approaches (e.g., laboratory and simulative settings, as well as with the help of neuroimaging tools; Kallionen et al., 2019), to optimize the description of emotional state during moral reasoning in the context of autonomous transportation.

Profiling AV' social perception

Moral research towards AV behaviors has recently risen to the headlines mainly due to the groundbreaking six-experiment research paper from Bonnefon, Shariff and Rahwan (2016), which highlighted a fundamental incongruency between moral evaluation and willingness to share the autonomous transportation technology. The authors evidenced that, despite considering the utilitarian AV behavior as the fairest approach to follow when facing dilemma-like situations, people would prefer to purchase a self-protective vehicle for themselves, regardless of the potential negative effect on the collectivity. This evidence led to a series of moral and social discussions (e.g., Haboucha et al., 2017; Kaur and Rampersad, 2018; Maurer et al., 2016; Moody et al., 2020), since it disclosed a significant barrier to the introduction of AVs into the mass market (Shariff et al., 2017). It also seemed to be an obstacle to the massive implementation of autonomous driving on international roadways, which we have seen to be the more economic, safer and practical solution to properly boost the advantages of full autonomous driving systems (see section 1.1). The observed phenomenon has been described as 'the social dilemma of self-driving cars', describing the detachment between moral action and moral evaluation in the social perception of autonomous transportation. We decided to deepen the problem throughout the four studies, aiming to achieve a wider description of the phenomenon assuming different perspectives.

Study 1 confirmed the distance between decision and evaluation outside also the AV context, reporting a scarce moral acceptability of the outcome selected in the moral dilemma, regardless the derived moral code (utilitarian or nonutilitarian) and the level of personal involvement. Despite the overall low evaluation of all the proposed moral options, a small moral preference for the selected behavior was still observed. The negative trend towards the evaluation of the expressed moral decision was confirmed when AV dilemmas were introduced in Study 2, which seems consistent among the investigated levels of automation (human-driving, autonomous driving). Interestingly, the manipulation of the self-sacrifice framing has revealed new aspects of this phenomenon. In fact, when the sacrifice of the moral agent aimed to the protection of the higher number of individuals, the moral evaluation of this utilitarian behavior was described as less immoral than its nonutilitarian counterpart. This result was coherent with the evidence collected from moral emotions in the same study: people seemed more inclined to pursue the utilitarian behavior, but this action was perceived as more immoral, shameful and blameworthy if it assume the protection of the moral agent at the expense of more lives (Sabini and Silver, 1997; Sachdeva et al., 2015; Smith et al., 2002).

In the subsequent studies, we deepened moral judgment towards AVs behavior considering additional features, such as time availability and prosocial orientation (Study 3), and information availability and perspective-taking (Study 4). Consistently with the methodology employed in the two previous studies, these two experiments were designed as two iterated dilemma paradigms, with three moral stimuli each. This technique has been the subject of theorical and methodological discussion, also analyzing, across dilemmas, the switch in the preferred moral judgment (Bostyn and Roets, 2022). In this context, in the investigation of the distance between action and evaluation, we decided to deepen the potential role of moral consistency when evaluating AV behaviors and willingness to share the autonomous technology. To this aim, we derived the correspondent 'moral profiles' from moral agents' decisions, investigating if (and how) decisional consistency shapes evaluation of moral acceptability and willingness to share of different AVs algorithms. Overall, and consistently with previous literature (Awad et al., 2018a; Bonnefon et al., 2016; Martin et al., 2021a), the utilitarian behavior was steadily evaluated as the most acceptable decisional rule, both from fully utilitarian and from morally inconsistent agents (i.e., switchers). Importantly, moral decision and moral evaluation were observed to be 'closer' to each other in the case of utilitarian moral consistency, detecting a potential reduction of the 'social dilemma' in this moral profile. In fact, fully utilitarian agents evaluated the correspondent moral behavior as the most acceptable (Study 3), and also reported a greater interest in sharing the autonomous technology when programmed to favor the highest number of people (Study 4). In this latter case, we believe that the assumption of different perspectives when reasoning about the moral problem may have ‘soften’ the contrast between moral beliefs and moral action, drawing closer the two values. Oppositely, this tendency was not observed in non-utilitarian profiles (fully nonutilitarian in Study 3, fully maximin in Study 4), which appeared to be less willing to purchase self-protective or maximin AVs, as also less prone to positively evaluate their moral decision, compared to their fully utilitarian peers. Interesting insights came from the opinions expressed by the inconsistent group (i.e., switchers), which possessed the more hybrid approach in the perception of AV behaviors. In fact, both in Study 3 and in Study 4 they reported grater moral value for the utilitarian dilemma resolution, despite of a non-clarified preference for a specific autonomous driving algorithm. In our view, moral inconsistent individuals were those who perceived the AV dilemma as more challenging and effective, since they did not indistinctively follow a certain moral code throughout the experimental sessions but evaluated each dilemma on the basis of different features and following non-equal values (Campbell, 2017). Overall, we believe that iterated dilemma-based research towards perception of AV (and AI in general) may benefit from the profiling of moral agents, since it allows to disclose the role of moral reasoning in the interpretation of individual perception of AVs evolution. The definition of individual attitudes can be further integrated, leveraging on additional instruments that may disclose – for example - the altruistic core of utilitarianism (Kahane et al., 2018), the individual level of AI literacy (Wang et al., 2022), the role of trust (Jessup et al., 2019; Jian et al., 2000) and acceptance in autonomous systems (e.g., Davis et al., 1989; Zhang et al., 2019).

Concluding remark

The present dissertation aimed to deepen the sacrificial moral dilemma as the elective experimental tool in the investigation of morality of autonomous transportation. The evidence collected from four studies highlight the important contribution of dilemma’s structure conditioning moral reasoning towards opposite levels of driving automation, as well as disclosing the impact of specific methodological and contextual features on shaping the likelihood towards an autonomous vehicle (AV) following a utilitarian moral code and decisional strategy. Useful insights were also collected in terms of the corresponding emotional activation and the social evaluation of AV' algorithms. Further studies will be needed to continue the optimization of the sacrificial moral dilemma tool in the investigation of morality and autonomous driving technology, and in general for stressing the pros and cons of this methodology for the retrieving social requirements, expectations, and attitudes towards this new revolutionary mode of transportation. In fact, any technological revolution is capable to improve general welfare without the compliance of its stakeholders. The moral alignment between humans and machines seems a fundamental step in the direction of trust, intelligibility and acceptance of automated driving systems.

Data Availability Statement: The data that support the findings of this thesis are available in Open Science Framework at: https://bit.ly/3cksq6Q

APPENDIX

Experimental Material: Study 1

Traditional non-driving sacrificial and incidental dilemma / other-involvement (TO, Lotto et al., 2014)

**Anchor Dilemma*

1) **Cargo**

You are in charge of the Maritime Police. You have received a message that the captain and two seamen of a cargo ship are healthy carriers of a lethal and contagious virus. They do not know that they are infected and are traveling towards a small gulf where nine fishermen live. The communications system of the cargo ship is not working and, therefore, they cannot be contacted.

a) You let the captain of the cargo and the two seamen moor in the small gulf, infecting and killing the nine fishermen.
b) You launch a missile and sink the cargo ship to avoid it mooring and thus spreading the virus. You know that the captain and the two seamen will die, but the nine fishermen will be saved.

2) **Crane**

You are a building worker who is maneuvering a crane on a building site. You have just started your day on the site when you realize that the cable of the crane is about to break. Attached to the cable is an enormous steel beam that is directly above a crew of six who are working on the outside of a building in construction.

a) As there is no way to warn them, you let the cable break and the steel beam kills the six workers.

b) You slightly move the crane arm to another area of the construction site. You know that two workers work there, who will be crushed by the steel beam and will die, but the six workers will be saved.

3) **Fire**

You are a fireman who is trying to save some people trapped in a fire which has broken out on the 20th floor of a skyscraper, which is luckily deserted. Six people are trapped in a room on the 21st floor and are already beginning to suffocate from the smoke, while due persons are coming down alone from the 19th floor on foot.

a) You let the fire reach the six people on the 21st floor of the building, who will die.
b) Activate the automatic closure of the fire doors. You know that this will mean that the persons coming down will be unable to escape from the building. He will die, but it will give you time to save the other six.

4) **Nurse***

You are a nurse who is in charge of a machine that controls drug dosage levels in patients' blood. Because of a technical failure, the machine is supplying a lethal dose of the drug to three patients. Another patient, in a single room, is hooked up to the same machine and has not undergone any variation in dosage.

a) As there is no way to fix the machine, you let the lethal dose be delivered to the three patients, who will die.

b) Press the button to block the drug supply to the four patients. You know that the overdose of drugs will be redirected to the patient in the single room, who will die, but the other three will be saved.

5) **Soldier**

You are a soldier in the Gulf War. An armed group has taken three civilian hostages and threatens to kill them. You have been able to discover where the hostages are being held and you must act quickly before they are killed. You have discovered that a tanker transporting oil is about to pass in front of where the hostages are being held.

a) As there is no time to break in the hideout, you let the three civilian hostages be killed by the kidnappers.
b) You shoot at the tanker so that it explodes, causing the kidnappers to leave their hideout. You know that the explosion will kill the driver, but it will make it possible to enter into action and save the three civilians.

6) **Motorboat**

You are driving your motorboat in a small bay when your attention is drawn to cries of help from six people who are drowning at the end of a very narrow channel that is right in front of you. Between you and the people who are drowning, to one side of the channel, is another two persons who are calmly swimming.

a) You do not enter the narrow channel where the six people are drowning, leaving them to die.

b) You steer towards the end of the channel at high speed. You know that the two persons who are swimming will be hit by the motorboat, but the other six people will be saved.

7) **Hospital**

You work as the night caretaker in a small provincial hospital. During one of your rounds, you realize that, because of a laboratory accident, some highly toxic fumes are spreading through the ventilation system towards a room in which there are six patients. In another room in the same ward, there are two patients.

a) You let the toxic fumes spread through the ventilation system towards the room with six patients, who will die.

b) You activate a switch that allows the toxic fumes to be diverted away from the room. You know that the fumes will be directed to the single room where the patient will die, but the other six will be saved.

8) **Quarantine**

The healthy carrier of a contagious and lethal disease is being held in quarantine in the hospital. Suddenly the ventilation system breaks down and there is no longer a change of air in the room. The emergency system will shortly be activated, and an internal window will be opened. This window opens into a ward in which three patients are being treated for various illnesses.

a) You let the emergency system be activated. The disease will spread through the three patients, who will die.

b) You block the emergency system by pressing a button that will keep the window closed. You know that the healthy carrier will suffocate, but the three patients will be saved from mortal contagion.

9) **Ferris wheel**

You are the safety officer in charge of a fun park. One of the metal arms of the ferris wheel suddenly breaks because of a structural defect. Three people are stranded in a cabin 80 meters up in the air. Another person is in a cabin just a few meters from ground level and is able to get off alone. The whole structure is falling down.

a) You let the ferris wheel collapse and fall, killing the three persons in the cabin.

b) You put the ferris wheel in motion to bring the cabin with the three people down. You know that the person who is getting off now will go up again and will die as the ferris wheel collapses, but the other three will be saved.

10) Trolley*

You are in charge of a work crew who are doing repair work for the railways. In the distance, you see a trolley and realize that the driver has lost control of it. If the trolley continues on it will end up running into three workers who are working on the tracks. On a secondary track, there is one worker.

a) You let the trolley continue its run on the main track, killing the three workers.

b) You pull a lever on the interchange which will divert the trolley onto the secondary track. You know that it will run into and kill the worker, but the other three workers will be unhurt.

Traditional non-driving sacrificial and incidental dilemma / self-involvement (TS, Lotto et al., 2014))

**Anchor Dilemma*

1) Bomb in the bank vault

You are in the head office of your bank together with five other people. Suddenly, the director calls you because he has discovered a bomb in an office on the ground floor. He knows you are a bomb disposal expert and asks you to defuse it. You realize immediately that there is not enough time to evacuate the people in the bank before the bomb explodes.

a) As there is no time to disarm the bomb, you let the bomb explode, killing you and the other five people.
b) You throw the bomb into the basement where there is the security vault. You know that the explosion will kill the security guard in the vault, but you and the other five people will be saved.

2) Underground cave

A very large man is leading you and other two explorers out of an underground cave on the west coast of Scotland. Because of his large size, the man gets stuck in the narrow opening at the mouth of the cave. In a short time, it will be high tide, and you will all drown except for the man, who has his head outside of the cave.

a) Since the passage is blocked, you let the high tide overwhelm you, and all three drowned.

b) You use the explosives you brought with you to widen the cave opening. You know that the man stuck will die, but you and the other two explorers will be able to escape.

3) Electric cable

A car accident causes a devastating explosion inside a long tunnel. You and another five survivors are finding your way through the debris to get out of the tunnel. Because of structural damage, a high-tension electric cable has snapped and is beginning to swing towards you. The asphalt is soaked with petrol. In the opposite direction, you see other two survivors coming towards you.

a) You let the high-tension electric cable touch the asphalt. The explosion will kill you and the other five survivors.

b) You use a rubber-soled shoe to divert the electric cable in the opposite direction. You know that the two persons coming towards you will be electrocuted, but you and the other five will be unhurt.

4) Atomic energy plant*

You are a worker in an atomic energy plant. After an explosion, there has been a leakage of radioactivity in your work area, where there are another two people. The doors are blocked because the security system has been activated, but the decontamination and communications systems have broken down. If you remain exposed to the radiation you will die in a short time.

a) Since is impossible to ask for help, you and the other two people get stuck in your work area. The radiation will kill all of you.

b) You activate the ventilation system to divert the radioactivity towards another work division. You know that there is a worker in the area who will die from the radiation, but you and the other two workers will be saved.

5) Dam

You are a specialist in underwater welding. You are working with your team of two other people in a small dam at a depth of 30m under water. Because of a problem with the flow regulator, about 100 cubic meters too much water has flowed into the dam. The pressure is too great, and you are unable to surface.

a) You activate the alarm, even if you know that the rescue team will not arrive in time. You and the other two people will die.

b) You activate the opening of the sluice gates. The water will flow into the valley where there is an observatory. You know that a researcher works there, and he/she will be swept away in the overflow water, but you and the other three will be saved.

6) Window

You are a fireman, and you are trying to save five people from inside a burning building. The only window from which the people can be evacuated is jammed and will not open. The fire will reach you in a short time. Outside on the window ledge of the floor below, there are two persons who are waiting to be saved.

a) Since the window is stuck, you let the flames reach you and the other five people, killing all of you.

b) With an ax, you smash the window to get out. You know that when it falls, the heavy glass will kill the two persons on the lower floor, but you and the five people in danger will be able to escape.

7) Waterfront*

You are a worker specialized in maneuvering a crane and are part of a work team that is loading containers into a ship. You have just lifted a container from the wharf when you realize that the cable of the crane is breaking, and that the container is about to crash down on you and the other five workers in the team.

a) Since there is no other way out, you let the cable of the crane break. The container will crush you and the other five workers.

b) You move the arm of the crane away from you. You know that the container will fall in an area in which there are two workers who will die crushed, but you and the other five workers will be saved.

8) Bodyguard

You are the bodyguard for an important politician. At the end of a rally, as you are getting into the car together with another bodyguard, the secret services inform you that a terrorist is heading towards you at high speed in a car filled with dynamite. With the binoculars, you see a car at a distance of several hundred meters.

a) You let the car filled with dynamite run onto you. The explosion will kill you, the politician, and the other bodyguard.

b) You shoot at the car coming towards you by aiming at the petrol tank. The explosion will hit a traffic policeman, who, unaware of the danger, is patrolling the area. You know that he will die, but you and the other two will be unhurt.

9) Roller coaster

You are at Luna Park and you have decided to take a ride on the roller coaster. You get into the carriage together with two other people. After a couple of circuits, the speed starts to increase dramatically right at the point that the carriage does a loop the loop. The technician in charge tells you over the loudspeaker that the mechanism which controls the brakes is not responding.

a) You let the carriage increase its speed until it gets off the tracks. You and the other two people will die in the crush.

b) You pull the emergency handle in the carriage that will make it divert onto another track. You can see there is a man working there. The carriage will run into him and he will die, but you and the other two will be unhurt.

10) Bull

You are at Pamplona with three colleagues and the Encierro has just finished, that is the running of the bulls through the streets of the city. While you are leaving, you realize that a bull has escaped from the corral and is heading towards you, attracted by your colleague's red bag. You have your shoulders to a wall and there is no time to escape.

a) You watch helplessly to the bull running onto the red bag, aware that you and the other two people will die in the clash.

b) You throw the red bag in the opposite direction. You know it will land near another person and that the bull will head for him/her, killing the person, but the three of you will be unhurt.

New human driving-type sacrificial and incidental dilemma / other-involvement (DO)

**Anchor Dilemma*

1) Driving Trolley*

You are driving your car, reaching a road area with a road sign indicating work in progress in proximity. In the middle of the road, there are three workers, dealing with road maintenance work. In an alley on your right, you can recognize a single worker. There are no other roads to take. You try to break when you realize that the car's brakes are not working.

a) You let the car proceed on the main road running over the three workers, who will die.

b) You suddenly steer right. You know that your car will run over the single worker in the alley, killing him, but the three workers on the main road will be unhurt.

2) Panoramic Road*

You are driving your car on a panoramic road. A cycle path is on your right, parallel to the roadway, and right now there is only a single cyclist running on it. You go straight on the road, while you suddenly see three workers in the middle of the road engaged in removing a small obstacle from the path. You are too close to them, there is no more time to brake.

a) You let the car proceed straight, running over the three workers, who will die.

b) You suddenly steer right. You know that your car will run over the single cyclist on the cycle path, who will die, but the three workers on the main road will be unhurt.

3) City hall

You receive a call from a city hall's office: some important papers need your sign. You get into your car and start driving towards the city hall. You are proceeding on a downhill road, and you can see a runner coming in the opposite direction. Suddenly three cyclists cross the road right in front of you. You are under the speed limit, but there is no more time to brake.

a) You let the car proceed straight, running over the three bikers, who will die.

b) You suddenly steer left. You know that your car will run over the single runner coming from the opposite direction, who will die, but the three cyclists will be unhurt.

4) Tourists

You are a taxi driver, driving four tourists to their destination. One of them is sitting beside you on the front seat, and the other three are sitting on the back seats. You all wear the seat belt properly. You are driving on a narrow road, and behind your car, there is a biker, when suddenly three runners cross the road right in front of you. You are too close to them, there is little time to brake.

a) You let the car proceed straight, running over the three runners, who will die.

b) You brake suddenly. You know that the biker behind you will fall off his motorbike crashing into the back of your car. He will die, but the three runners will be unhurt.

5) Collision

You are driving your car. Three cyclists are running in front of you, while a single biker is proceedings behind your car. Suddenly, one of the three cyclists slides along a curve, falling off his bike and causing the other two to fall too. They are on the asphalt, very close to your car. Even if under the speed limit, your cruising speed does not give you the time to brake safely.

a) You let the car proceed straight, running over the three cyclists, who will die.

b) You brake suddenly. You know that the biker behind you will fall off his motorbike crashing into the back of your car. He will die, but the three cyclists will be saved.

6) Biker

You are driving your car on a two-lane road, and you can see a biker on his motorbike proceeding in the opposite position. On your right, there are a series of parking spots. You reach a bulky pickup parked on a parking spot when suddenly three pedestrians appear from behind. Even if under the speed limit, they are too close and there is no more time to brake.

a) You let the car proceed straight, running over the three pedestrians, who will die.

b) You suddenly steer left. You know that your car will run over the single biker coming from the opposite direction, who will die, but the three pedestrians will be unhurt.

7) Pedestrian

It is night. You are driving your car and a thick fog has fallen on your city. You are driving on a one-way street, and two cycle paths flank the road. On the left path, you can see a single cyclist. You are approaching a crosswalk when suddenly three more cyclists cross the road. Because of the thick fog, you did not notice them, and now there is no more time to brake.

a) You let the car proceed straight, running over the three cyclists, who will die.

b) You suddenly steer left. You know that your car will run over the single cyclist on the left cycle path, who will die, but the other three cyclists will be unhurt.

8) Sun

It is a sunny morning, and you are driving your car right in the direction of the sun. On your left, you can see a person calling from a phone box. Suddenly a sunbeam breaks through the windshield of your car, preventing you from clearly seeing the road. As soon as you regain your sight, three persons start crossing the road right in front of you. There is no more time to brake.

a) You let the car proceed straight, running over the three pedestrians, who will die.

b) You suddenly steer left. You know that your car will run over the person calling from the phone box, who will die, but the other three pedestrians will be unhurt.

9) **Overtaking**

It is night. You are driving your car to a restaurant nearby for a business dinner. You are proceeding on a two-lane road, and you can see a motorbike and a red car coming in the opposite direction. Suddenly, three pedestrians cross the road right in front of you. Even if under the speed limit, they are too close and there is no more time to brake.

a) You let the car proceed straight, running over the three pedestrians, who will die.

b) You suddenly steer left. You know that your car will run over the single biker coming from the opposite direction, who will die, but the three pedestrians will be unhurt.

10) Road Tunnel

You are driving your car, and you are going through a road tunnel right below the railway station. The tunnel is almost finished, and you can notice, just outside the tunnel, a single person waiting at the bus stop few meters to the left. You come out of the gallery when suddenly three skaters cross the road from your right. Even if under the speed limit, they are too close and there is no more time to brake.

a) You let the car proceed straight, running over the three skaters, who will die.

b) You suddenly steer left. You know that your car will run over the single person waiting at the bus stop, who will die, but the three skaters will be unhurt.

New human driving-type sacrificial and incidental dilemma / self-involvement (DS)

**Anchor Dilemma*

1) Autostop

You are driving your car when you see two people trying to hitch a ride. You decide to give them a ride, so you stop and let them get in the car. After a few kilometers, suddenly you see a track coming from your right with the only driver inside. Even if under the speed limit, your driving speed does not give you the time to brake safely. On the left sidewalk, you see a person driving his bike.

a) You let the car proceed straight. The truck will crash against your car killing you and the two passengers.

b) You suddenly steer left. You know that your car will run over the single biker on the sidewalk, who will die, but you and the two passengers will be saved.

2) Ravine

It is morning, and you are driving your car on a road that runs along a ravine. Suddenly, a dog cuts you off in traffic. To avoid it you lost control of the vehicle, which is now headed off the road in the direction of two runners and the ravine behind them. There is no more time to brake. On the right, you see a single person seated at the bus stop.

a) You let the car proceed straight, running over the two runners, who will die. Your car will drive into the ravine, and you will die in the crash.

b) You suddenly steer right. You know that your car will run over the single person waiting at the bus stop, who will die, but you and the two runners will be unhurt.

3) Taxi

You are a taxi driver, driving a passenger on a tree-lined avenue. Suddenly one of your tires bursts because of a hole in the asphalt. You have lost control of the car, which is now headed off the road in the direction of a cyclist and a big tree behind him. There is no more time to brake. On the right, you see a person calling from a phone box.

a) You let the car proceed straight, running over the cyclist, who will die. Your car will crash against the big tree, and you and the passenger will die.

b) You suddenly steer right. You know that your car will run over the single person calling from the phone box, who will die, but you, the passenger, and the cyclist will be unhurt.

4) Concrete mixer

It is early in the morning, and you are driving your truck on a two-lane road. There is only a car coming in the opposite direction. Suddenly you notice two workers and a bulky concrete mixer a few meters from you in the middle of the road, dealing with road maintenance work. You begin to slow down when you realize that the truck’s brakes are not working.

a) You let the truck proceed straight, running over the two workers, who will die. The truck will crash against the concrete mixer, and you will die.

b) You suddenly steer left. You know that your truck will crash against the car coming from the opposite direction, killing its driver, but you and the two workers will be saved.

5) Streetlamp

It is night and you are driving your car. Suddenly you see a truck with the main beams coming in your direction: you cross its beams with your sight, which blinds you for few seconds. As soon as you regain your sight, you discover that your car is now headed off the road in the direction of two bikers, who are taking a break from driving under a big streetlamp. There is no more time to brake. On the right roadside, you see a hitchhiker.

a) You let the car proceed straight, running over the two bikers, who will die. The car will crash against the streetlamp, and you will die.

b) You suddenly steer right. You know that your car will run over the hitchhiker, who will die, but you and the two bikers will be unhurt.

6) Snow

You are driving your car. It is winter, and the asphalt is still partially frozen after a heavy snowfall. You are approaching a low traffic uphill on your left when suddenly you see two persons sledding down the road in your direction. You try to brake when you realize that the car’s brakes are not working. On the right sidewalk, you see a pedestrian.

a) You let the car proceed straight, hitting the two persons on the sled, who will die. Your car will not brake on the frozen asphalt and will hit a big tree, where you will die.

b) You suddenly steer right. You know that your car will run over the pedestrian, who will die, but your car will slow down in an open field and you and the two persons on the sled will be saved.

7) Rush

You are driving your car on a tree-lined avenue. On your left run a cycle path, where you can see a single cyclist. You are following a red car, which is now braking for no apparent reasons. The two road lanes are separated by a dotted line, so you decide to overtake the red car. During the overtaking, two pedestrians suddenly appear from the right. There is no more time to brake.

a) You let the car proceed straight, running over the two pedestrians, who will die. Your car will swerve crushing against a big tree, and you will die.

b) You suddenly steer left. You know that your car will run over the cyclist on the cycle path, who will die, but you and the two pedestrians will be unhurt.

8) Highway*

It is night and you are driving your truck on the highway. You are overtaking a motorbike when you suddenly notice a black car with two passengers standing in the middle of the lane. You suppose that the car had a malfunction. You try to brake when you realize that the car's brakes are not working.

a) You let the car proceed straight, hitting the black car and killing its passengers. Your truck will swerve crushing against the guard-rail, and you will die.

b) You suddenly steer right. You know that your car will hit the motorbike, killing its driver, but you and the two passengers of the black car will be unhurt.

9) Rain

It is night and you are driving your car. A violent storm has hit your city for a few hours, it is still raining, and the asphalt is slippery. You are approaching a traffic light when suddenly two cyclists cross the road right in front of you. You try to brake, but the tires slide on the asphalt. On the right sidewalk, you see a pedestrian.

a) You let the car proceed straight, running over the two cyclists, who will die. Your car will continue to slide on the asphalt, crushing against a near building, and you will die.

b) You suddenly steer right. You know that your car will run over the pedestrian on the sidewalk, who will die, but your car will slow down in an open field and you and the two cyclists will be saved.

10) Fog*

You are a taxi driving. It is night and you are driving a passenger. As in the last nights, a thick fog has descended on your city and the visibility is strongly compromised. Nonetheless, you can see a pedestrian on the right sidewalk. Suddenly you notice a cyclist crossing the road right in front of you. Because of the thick fog, you did not notice him and now there is no more time to brake.

a) You let the car proceed straight, running over the cyclist, who will die. Your car will swerve, crushing against a streetlamp, and you and your passengers will die.

b) You suddenly steer right. You know that your car will run over the pedestrian on the sidewalk, who will die, but you and your passenger will be unhurt.

Filler nonsacrificial dilemma (NS)

**Anchor Dilemma*

1) Gift*

You have been invited to the birthday party of someone you do not know well. You are not really very keen on going, or on spending a lot of money on a present, because he/she is not a close friend. You find a brand-name jumper in excellent condition in a second-hand shop and buy it. On opening the present, the receiver seems embarrassed by such a generous big present.

a) You smile satisfied and reassure the receiver, saying that an important occasion like his/her birthday deserves a special present. You know that he/she will then think of you as a generous person.

b) As the receiver opens the gift, you immediately let him/her know the truth about the jumper, explaining that today second-hand fashion is becoming more and more trendy.

2) Cake*

You have often talked with your friends about making a delicious cake which is an old family tradition. You have finally decided to give them a surprise and have invited them to come round after dinner. In the afternoon, you get down to work on the recipe and realize that you do not have any dark chocolate which is what gives this cake its very particular taste.

a) You use milk chocolate which you already have, so you do not have to go out again. You know that your friends do not know the recipe and will not realize it.

b) Even if you are not very keen on going out, you go buy dark chocolate at the supermarket. You know that the expectations are high, so the cake has to be perfect.

Experimental material: Study 2

Human-driving sacrificial (self-involvement) and incidental dilemmas

A1)

You are driving a car with a passenger on a two-lane country road, and a post office truck is proceeding in front of you. The truck begins to slow down, stopping on the right side of the lane. A dotted line separates the two road lanes, so you decide to overtake it. During the overtaking, four cyclists suddenly cross the road, appearing from behind the truck. You did not see them coming, and you have no time to brake.

You have two alternatives:

a) You let the car proceed straight, running over the four cyclists, who will die.

b) You suddenly steer left. The four cyclists will be unhurt, but your car will crash against a streetlamp, and you and your passenger will die.

A2)

You are a taxi driver. It is morning, and you are driving a passenger. A violent storm has hit your city for a few hours, it is still raining, and the visibility is strongly compromised. You can notice a bus stop few meters ahead, when suddenly four pedestrians appear from the right, now standing in the middle of the road. Because of the heavy rain, you did not notice them, and now you have no time to brake.

You have two alternatives:

- You let the car proceed straight, running over the four pedestrians, who will die.
- You suddenly steer left. The four pedestrians will not be hurt, but your taxi will crash against the bus stop, and you and your passenger will die.

A3)

You are driving a car with a passenger on a two-lane highway. It is late night, and only one car is coming in the opposite direction. Suddenly, you notice a small van on the side of the road and four workers a few meters from you in the middle of the road dealing with road maintenance work. You begin to slow down when you realize that the brakes are not working.

You have two alternatives:

- You let the car proceed straight, running over the four workers, who will die.
- You suddenly steer left. The four workers will be unhurt, but your car will crash against the guardrail, and you and your passenger will die.

<u>Manual, nonutilitarian sacrifice</u>

B1)

You are driving a car with a passenger on a two-lane city road. A cycle path is on your right, parallel to the roadway, and right now, two cyclists are riding on it. You are following a taxi, which is now pulling over to the right to load a passenger. A dotted line separates the two road lanes, so you decide to overtake it. During the overtaking, two pedestrians suddenly cross the road, appearing from behind the taxi. You did not see them coming, and you have no time to brake.

You have two alternatives:

- You let the car proceed straight, running over the two pedestrians, who will die. Your car will swerve, crashing against a building, and you and your passenger will die.
- You suddenly steer left, running over the two cyclists on the cycle path, who will die, but you, your passenger, and the two pedestrians will be saved.

B2)

You are a taxi driver. It is night, and you are driving a passenger. As in the last few nights, a thick fog has descended on your city, and the visibility is strongly compromised. You notice two pedestrians on the right sidewalk. Suddenly, you notice two cyclists crossing the road right in front of you. Because of the thick fog, you did not notice them, and you have no time to brake.

You have two alternatives:

- You let the car proceed straight, running over the two cyclists, who will die. Your taxi will swerve, crashing against a building, and you and your passenger will die.
- You suddenly steer left, running over the two pedestrians on the sidewalk, who will die, but you, your passenger, and the two cyclists will be saved.

B3)

You are driving a car with a passenger on a two-lane highway. It is early morning, and a motor cyclist with a passenger is coming in the opposite direction. Suddenly, you notice a bulky concrete mixer and two workers a few meters from you in the middle of the road, diverting the traffic flow. You begin to slow down, and you realize that the brakes are not working.

You have two alternatives:

- You let the car proceed straight, running over the two workers, who will die. Your car will swerve, crashing against the concrete mixer, and you and your passenger will die.
- You suddenly steer left, hitting the motor cyclist and its passenger, who will die, but you, your passenger, and the two workers will be saved.

Autonomous-driving, utilitarian sacrifice

C1)

You and another person are the passengers of a fully autonomous vehicle, driving on a tree-lined avenue. A truck is proceeding in front of you, which is now slowing down for no apparent reason. A dotted line separates the road lanes, so you decide to overtake it. During the overtaking, four runners suddenly cross the road, appearing from behind the truck. The autonomous vehicle did not perceive them in time, and it has no time to brake.

The autonomous vehicle has two alternatives:

- Proceed straight, running over the four runners, who will die.
- Suddenly steer to the left. The four runners will not be hurt, but the autonomous vehicle will crash against a big tree, and you and the other passenger will die.

C2)

You and another person are the passengers in a fully autonomous taxi vehicle. As in the last few nights, a thick fog has descended on your city, and the visibility is strongly compromised. You notice some vehicles parked on the right side of the road. Suddenly, four pedestrians appear from the right, now standing in the middle of the road. The autonomous vehicle did not perceive them in time, and it has no time to brake.

The autonomous vehicle has two alternatives:

- Proceed straight, running over the four pedestrians, who will die.

- Suddenly steer to the left. The four pedestrians will be unhurt, but the autonomous taxi will crash against a big tree, and you and the other passenger will die.

C3)

You and another person are the passengers in a fully autonomous vehicle, driving on a two-lane highway. It is early morning, and yours is the sole vehicle on the road. Suddenly, the autonomous vehicle notices a road sign and four workers a few meters ahead in the middle of the road dealing with road maintenance work. The vehicle begins to slow down, and you realize that the brakes are not working.

The autonomous vehicle has two alternatives:

- Proceed straight, running over the four workers, who will die.
- Suddenly steer to the left. The four workers will not be hurt, but the autonomous vehicle will crash against a big tree, and you and the other passenger will die.

Autonomous-driving, nonutilitarian sacrifice

D1)

You and another person are the passengers in a fully autonomous vehicle, driving on a two-lane city road. A cycle path is on your right, parallel to the roadway, and right now, two cyclists are riding on it. You are following a car, which is now slowing down to park on the right. A dotted line separates the two lanes, so you decide to overtake it. During the overtaking, two pedestrians suddenly cross the road, appearing from behind the car. The autonomous vehicle did not perceive them in time, and it has no time to brake.

The autonomous vehicle has two alternatives:

- Proceed straight, running over the two pedestrians, who will die. The autonomous vehicle will swerve, crashing against a big tree, and you and the other passenger will die.
- Suddenly steer left, running over the two cyclists on the cycle path, who will die, but you, the other passenger, and the two pedestrians will be saved.

D2)

You and another person are the passengers in a full autonomous taxi vehicle. A violent storm has hit your city for a few hours, it is still raining, and visibility is strongly compromised. You notice two pedestrians on the right sidewalk. Suddenly, two cyclists appear from the right, now in the middle of the road. The autonomous vehicle did not perceive them in time, and it has no time to brake.

The autonomous vehicle has two alternatives:

- Proceed straight, running over the two cyclists, who will die. The autonomous vehicle will swerve, crashing against a streetlamp, and you and the other passenger will die.
- Suddenly steer left, running over the two pedestrians on the sidewalk, who will die, but you, the other passenger, and the two cyclists will be saved.

D3)

You and another person are the passengers in a fully autonomous vehicle, driving on a two-lane country road. It is night, and two cyclists are coming in the opposite direction. Suddenly, the autonomous vehicle notices a big truck and two workers a few meters ahead in the middle of the road, diverting the traffic flow. The vehicle begins to slow down, and you realize that the brakes are not working.

The autonomous vehicle has two alternatives:

- Proceed straight, running over the two workers, who will die. The autonomous vehicle will crash against the truck, and you and the other passenger will die.
- Suddenly steer left, running over the two cyclists from the opposite line, who will die, but you, the other passenger, and the two workers will be saved.

Experimental material: Study 3.

<u>Autonomous-driving sacrificial and incidental dilemma / self-involvement (reworked by Study 2)</u>

1) Truck

You are the passenger of a fully autonomous vehicle, driving on a tree-lined avenue. A truck is proceeding in front of you, which is now slowing down for no apparent reason. A dotted line separates the road lanes, so you decide to overtake it. During the overtaking, five cyclists suddenly cross the road, appearing from behind the truck. The autonomous vehicle did not perceive them in time, and it has no time to brake.

The autonomous vehicle has two alternatives:

- Proceed straight, running over the five cyclists, who will die.
- Suddenly steer to the left. The five runners will not be hurt, but the autonomous vehicle will crash against a big tree, where you will die.

2) Fog

You are the passenger in a fully autonomous taxi vehicle. As in the last few nights, a thick fog has descended on your city, and the visibility is strongly compromised. You notice some vehicles parked on the right side of the road. Suddenly, five pedestrians appear from the right, now standing in the middle of the road. The autonomous vehicle did not perceive them in time, and it has no time to brake.

The autonomous vehicle has two alternatives:

- Proceed straight, running over the five pedestrians, who will die.
- Suddenly steer to the left. The five pedestrians will be unhurt, but the autonomous taxi will crash against a building, where you will die.

3) Workers

You are the passenger in a fully autonomous vehicle, driving on a two-lane highway. It is early morning, and yours is the sole vehicle on the road. Suddenly, the autonomous vehicle notices a road sign and five workers a few meters ahead in the middle of the road dealing with road maintenance work. The vehicle begins to slow down, and you realize that the brakes are not working.

The autonomous vehicle has two alternatives:

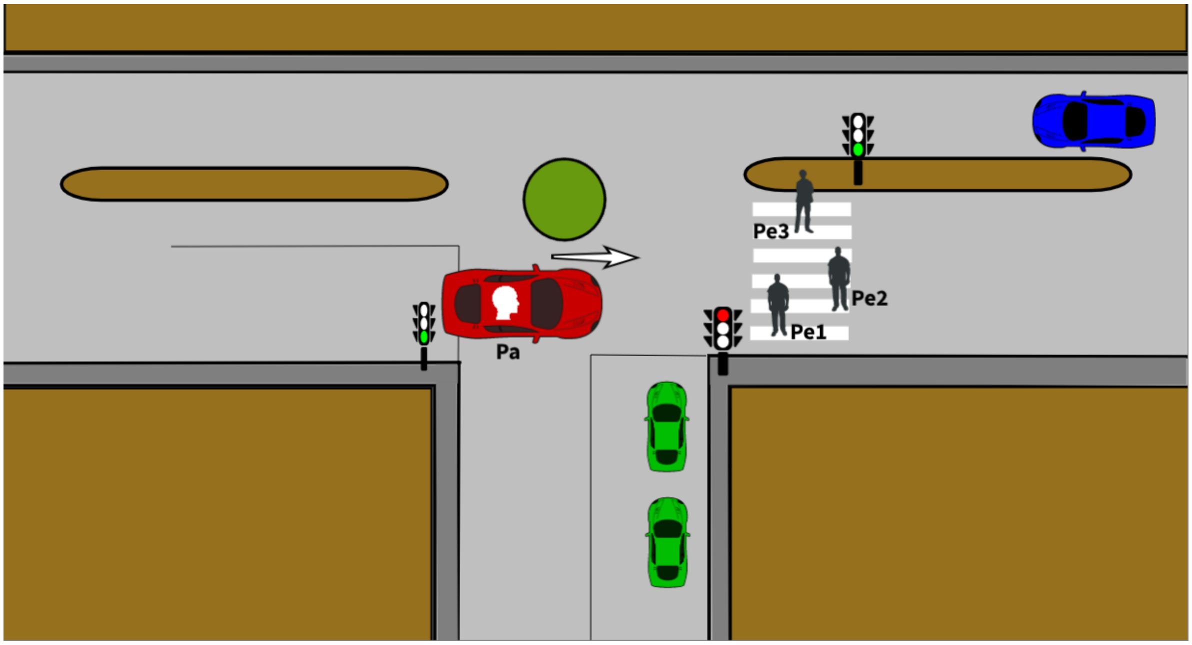
- Proceed straight, running over the five workers, who will die.
- Suddenly steer to the left. The five workers will not be hurt, but the autonomous vehicle will crash against a streetlamp, where you will die.

Experimental material: Study 4.

Autonomous-driving sacrificial and incidental trilemmas / self-involvement (reworked from Martin et al., 2021; Thomson, 2008)

Thick and Thin Veil Scenario:

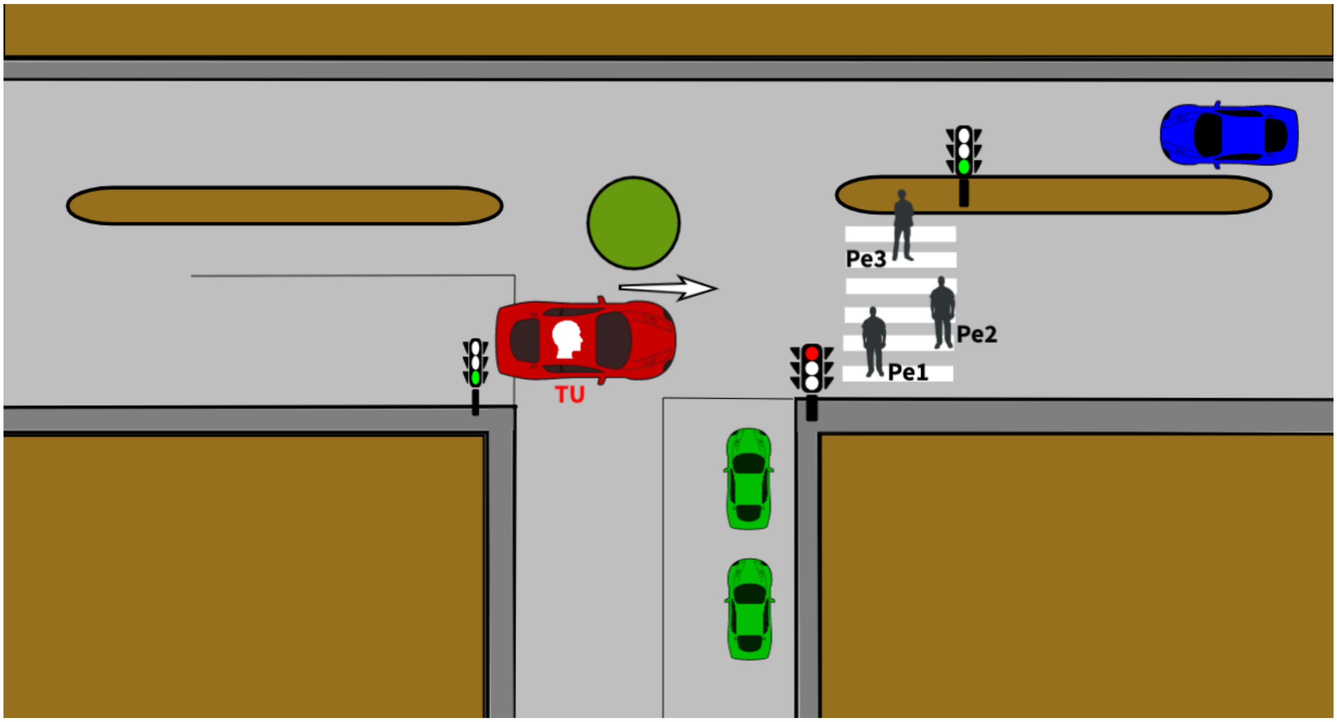
YOU could be the sole passenger (Pa) in an autonomous self-driving vehicle traveling at the speed limit down an urban road. OR you could be one of the three pedestrians now crossing the road. Pe1 and Pe2 are in the middle of the road, whereas Pe3 is just behind them. Because of a traffic lights malfunction,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



The image of the AV trilemma in the Thick and Thin veil scenarios (Pa = Passenger; Pe1/Pe2/Pe3 = Pedestrians).

No Veil Scenario, AV passenger perspective:

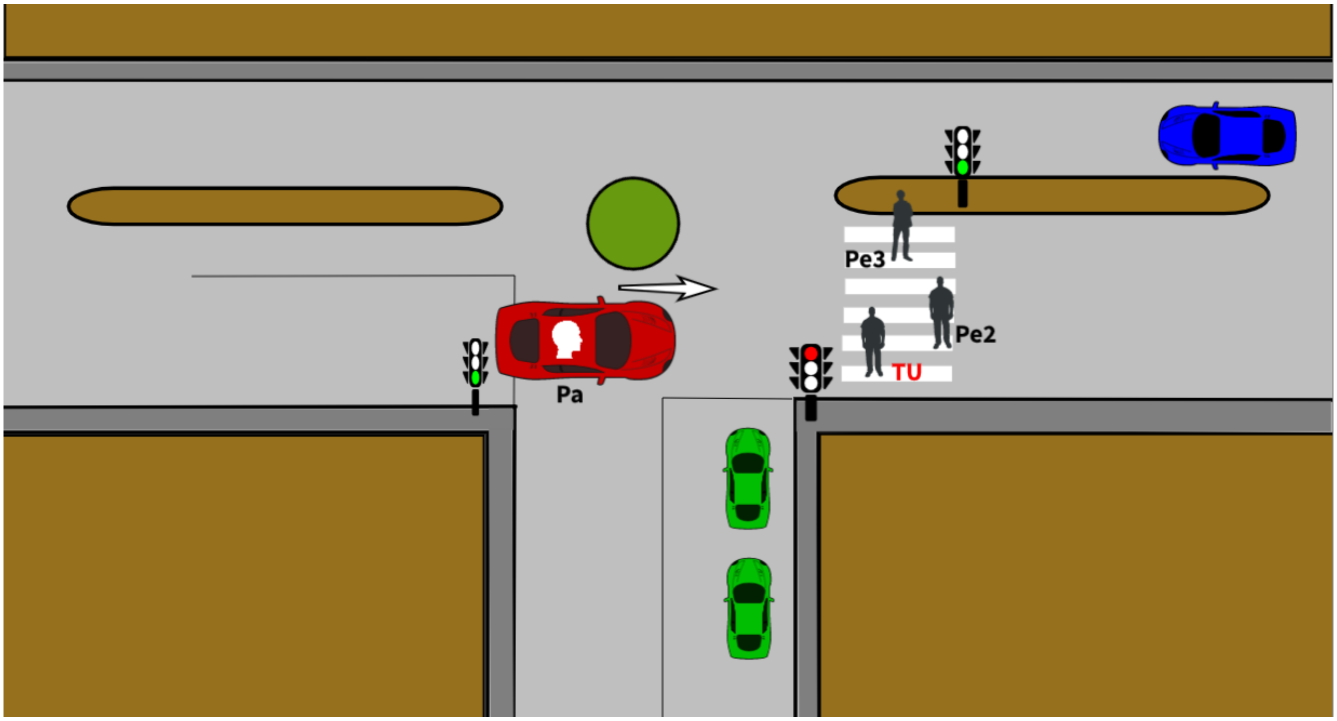
YOU are the sole passenger (Pa) in an autonomous self-driving vehicle traveling at the speed limit down an urban road. Three pedestrians are now crossing the road. Pe1 and Pe2 are in the middle of the road, whereas Pe3 is just behind them. Because of a traffic lights malfunction, the pedestrians are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



The image of the AV trilemma in the passenger perspective's version of the No Veil Scenario (TU = Passenger and moral agent; Pe1/Pe2/Pe3 = Pedestrians).

Veil Scenario, Pedestrian perspective:

An autonomous self-driving vehicle is traveling at the speed limit down an urban road, with a single passenger on board. YOU and two other pedestrians are now crossing the road. YOU and Pe2 are now in the middle of the road, whereas Pe3 is just behind you. Because of a traffic lights malfunction, the pedestrians are now in the direct path of the car. There is no more time to brake. Facing this event, the autonomous vehicle may be programmed to implement three different emergency maneuvers, resulting in different risks for the passenger and the pedestrians.



The image of the AV trilemma in the pedestrian perspective's version of the No Veil Scenario (TU = Pedestrian and moral agent; Pa = Passenger; Pe2/Pe3 = Pedestrians).

Additional tables

Study 3

Table A3.1: Beta estimates e p-values from M$_1$ to M$_4$.

	N	M$_1$ Moral Decision	M$_2$ Decision Time	M$_3$ Moral Evaluation	M$_4$ Willingness to Buy
Time condition					
Control	67	-	-	-	-
Time constraint	69	2.81	-0.28*	-	-
Time delay	70	3.57	-0.56***	-	-
Dilemma order					
D1	206	-	-	-	-
D2	206	2.25***	-.69***	-	-
D3	206	2.07**	-.79***	-	
SVO profiles					
Proself	184	-	-	-	-
Prosocial	22	4.16*	-1.25***	6.81	-
Gender					
Female	103	-	-	-	-
Male	103	-1.48*	-	-8.10*	-
Moral profile					
Utilitarian	79	-	-	-	-
Nonutilitarian	71	-	0.12	-31.10***	12.03*
Switcher	56	-	0.21*	-12.39***	14.22*
AV driving type					
Self-protective	-	-	-	-	-
Utilitarian	-	-	-	-	12.56**
R^2 marg (R^2cond)*		.11 (.81)	.15 (.39)	.23 (.23)	.05 (.33)

*Notes: *p < .05; **p < .01; ***p < .001; NUT = nonutilitarian, UT = utilitarian; UT-Sac = Utilitarian Sacrifice framing, NUT-Sac = Nonutilitarian Sacrifice framing. Beta estimates and p-values of the interaction effects are retrievable in the correspondent R script in the OSF project folder.*

Table A3.2: Mean and Standard Deviation (in brackets) of the five Moral Foundation Theory subscales (min = 0, max = 30), divided by SVO profiles (proself: prosocial: individualistic and competitive: altruistic and cooperative), Moral consistency profiles (fully utilitarian, fully nonutilitarian, switcher), and overall.

	Care/	**Fairness/**	**Ingroup/**	**Authority/**	**Purity/**
	Harm	**Justice**	**Loyalty**	**Respect**	**Sanctity**
SVO profiles					
Proself (N = 22)	24.4 (5.6)	26.4 (4.4)	19.4 (6.2)	17.4 (4.5)	15.1 (4.9)
Prosocial (184)	27.7 (4.3)	28.9 (4.1)	20.4 (5.2)	17.6 (5.1)	16.2 (5.4)
Moral profiles					
Utilitarian (N = 79)	28.2 (3.9)	29.5 (3.2)	20.6 (5.2)	17.6 (5.0)	16.0 (5.5)
Nonutilitarian (71)	26.5 (4.9)	28.3 (4.3)	20.6 (5.1)	17.3 (4.9)	16.2 (5.7)
Switcher (56)	27.1 (4.8)	27.9 (5.2)	19.4 (5.8)	17.8 (5.4)	16.1 (4.9)
Overall (N = 206)	27.3 (4.6)	28.6 (4.2)	20.3 (5.3)	17.5 (5.0)	16.1 (5.4)

Table A3.3: Mean and Standard Deviation (in brackets) of individuals' social interactions in life and at work (min = 0, max = 5), divided by SVO profiles (proself: prosocial: individualistic and competitive: altruistic and cooperative), Moral consistency profiles (fully utilitarian, fully nonutilitarian, switcher), and overall.

	Social activities (yes, %)	**Iterated interactions**	**Competition with colleagues/peers**	**Rely on rewards in case of poor results**	**Risk of sanctions in case of good results**
SVO profiles					
Proself (N = 22)	54.5	2.91 (1.57)	2.36 (1.47)	2.68 (1.43)	3.27 (1.24)
Prosocial (184)	61.9	3.27 (1.43)	2.07 (1.27)	3.14 (1.36)	3.05 (1.49)
Moral profiles					
Utilitarian (N = 79)	65.8	3.52 (1.35)	1.97 (1.29)	3.08 (1.36)	2.96 (1.48)
Nonutilitarian (71)	53.5	2.93 (1.54)	2.17 (1.31)	2.79 (1.44)	3.06 (1.45)
Switcher (56)	64.3	3.20 (1.41)	2.18 (1.29)	3.50 (1.19)	3.25 (1.48)
Overall (N = 206)	61.1	3.23 (1.45)	2.10 (1.30)	3.09 (1.37)	3.07 (1.47)

Table A4.1: Beta estimates e p-values from M_1 to M_4.

		M_1	**M_2**	**M_3**	**M_4**	**M_4**
	N	**Nonutilitarian**	**Utilitarian**	**Maximin**	**Moral**	**Willingness to**
		Decision	**Decision**	**Decision**	**Evaluation**	**Buy**
VOI type						
No veil	239	-	-	-	-	-
Thick veil	239	-6.40***	2.80***	-1.10**	-	-
Thin veil	239	-4.88***	1.57***	0.01	-	-
*Partial PT*VOI type*						
> No veil						
Pass Vs. Ped	-	2.43	-3.78***	2.30***	-	-
> Thick veil						
Pass Vs. Ped	-	2.41	-.031	0.11	-	-
> Thin veil						
Pass Vs. Ped	-	-1.46	-.94	0.59	-	-
Gender						
Female	120	-	-	-	-	-
Male	119	-	1.23**	-1.96***	-	-
AV behavior						
Minimize casualties	-	-	-	-	-	-
Prioritize AV passenger	-	-	-	-	-53.95***	-16.49***
Distribute risk	-	-	-	-	-39.13***	-29.26***
Moral profile						
Utilitarian	83	-	-	-	-	-
Maximin	45	-	-	-	-27.17***	-21.49***
Switcher	108	-	-	-	-8.15*	-11.00**
R^2 marg (R^2theo)*		.05	.17	.12	.37	0.09
		(.69)	(.69)	(.75)	(.39)	(0.33)

*Notes: *p < .05; **p < .01; ***p < .001; NUT = nonutilitarian, UT = utilitarian; UT-Sac = Utilitarian Sacrifice framing, NUT-Sac = Nonutilitarian Sacrifice framing. Beta estimates and p-values of the interaction effects are retrievable in the correspondent R script in the OSF project folder.*

Table A4.2: Mean and Standard Deviation (in brackets) of Interpersonal Reactivity Index (IRI) total sub scores (Fantasy, Emphatic concern, Perspective taking, Personal distress), divided by gender (Female, Male) moral consistency profiles (fully utilitarian, fully maximin, switchers), and overall.

	Fantasy	**Emphatic concern**	**Perspective taking**	**Personal distress**
	(max = 20)	**(max = 24)**	**(max = 28)**	**(max = 28)**
Gender				
Female (N = 120)	18.6 (3.3)	23.9 (3.8)	26.0 (4.6)	20.9 (5.2)
Male (119)	16.2 (4.0)	21.3 (4.1)	25.4 (3.9)	19.1 (4.8)
Moral profiles				
Utilitarian (N = 83)	16.9 (3.9)	22.3 (4.1)	25.4 (4.3)	18.8 (4.9)
Maximin (45)	18.3 (4.1)	23.8 (4.2)	25.6 (3.6)	21.0 (5.1)
Switcher (108)	17.4 (3.7)	22.4 (4.1)	26.0 (4.3)	20.4 (5.1)
Overall (N = 239)	17.4 (3.9)	22.6 (4.2)	25.7 (4.2)	20.0 (5.1)

Glossary

Autonomous Vehicle (AV): vehicles in which *"operation of the vehicle occurs without direct driver input to control the steering, acceleration, and braking and are designed so that the driver is not expected to constantly monitor the roadway while operating in self-driving mode*" (NHTSA, 2013). SAE (2021) distinguished six levels of automation depending on the on-board driver assistance system. This classification moves between No-automation (Level 0), Assistive driving systems (Level 1 and 2) and Automated driving systems (Level 3, 4 and 5). In the present dissertation, the Autonomous Vehicle label has been always referred to an automated driving system at Level 5.

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

Emotional activation: In the dissertation, this term has been used to describe the level of involvement of arousal and valence in response to particular moral stimuli. A higher emotional activation correspond to a higher self-referred level of arousal and a worst valence of the experienced stimulus.

Incidental dilemma: Definition based on the Doctrine of the Double Effect (DDE; Aquinas, 1247/1952), according to which it is not permissible to intentionally cause harm for a greater good, although it is permissible as a foreseen but unintended side effect. Specifically, the death of one (or few) person(s) is a foreseen but unintended consequence of the action aimed at saving more people. (Lotto et al., 2014)

Instrumental dilemma: Definition based on the Doctrine of the Double Effect (DDE; Aquinas, 1247/1952), according to which it is not permissible to intentionally cause harm for a greater good, although it is permissible as a foreseen but unintended side effect. Specifically, the instrumental dilemmas is a particular case in which the death of one (or a few) person(s) is a means to save more people. (Lotto et al., 2014)

Moral agent: a human being who has the ability to discern right from wrong, being accountable for his/her own actions.

(Moral) dilemma: In the most general sense, a dilemma is a situation that requires a choice between two options that are (or seem to be) equally undesirable or unsatisfactory. Specifically, a moral dilemma is a situation in which a moral agent has to prioritize one moral value over another, assuming that whichever action is take, it will offend an opposite moral value. A dilemma is defined as *nonmoral* when no moral dimensions are involved in the required decision. (Kvalnes, 2019; Maclagan, 2003).

Moral emotions: Those thought to relate to the capacity for human morality. Examples of such emotion types include disgust, shame, pride, anger, guilt, compassion, and gratitude. (Haidt, 2003b; Walsh, 2021)

Moral evaluation: the evaluation of moral acceptability of a given option.

Sacrificial dilemma: A moral dilemma in which the moral agent has to decide between two unsatisfying actions with outcomes involving the loss of at least one human life.

Social Value Orientation (SVO): it refers to individuals' preference for competition or cooperation in interpersonal exchanges.

Time Constraint: Time constraint is defined as an externally imposed deadline for the assumption of a particular decision (Ordóñez et al., 2015). This definition is typically used interchangeably with time pressure. Nonetheless, time pressure is better interpreted as the subjective feeling of having not enough time to finalize a certain task, while the time constraint is derived from external superimpositions. (Chu and Spires, 2001)

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

Veil of Ignorance (VOI): Hypothetical setting originally described by John Rawls. The author claims that when we are thinking about justice, we should imagine to not knowing many of the facts – both about ourselves and the society we currently live in – that typically influence our thinking in biased ways. By ignoring these facts, Rawls hoped that we would be able to avoid the biases that might otherwise come into a group decision. John Harsanyi also described this decisional setting, leaving a little amount of more information at disposal of the rational agent. (Harsanyi, 1975; 1978; Rawls, 1971/2009).

Willingness to buy: the availability to purchase a given item.

References

Abraham, H., Lee, C., Brady, S., Fitzgerald, C., Mehler, B., Reimer, B., and Coughlin, J. F. (2017). Autonomous vehicles and alternatives to driving: trust, preferences, and effects of age. In Proceedings of the transportation research board 96th annual meeting (pp. 8-12). Washington, DC: Transportation Research Board.

Ahituv, N., Igbaria, M., and Sella, A. V. (1998). The effects of time pressure and completeness of information on decision making. Journal of management information systems, 15(2), 153-172.

Albiero, P., Ingoglia, S. and Lo Coco, A. (2006). Contributo all'adattamento italiano dell'Interpersonal Reactivity Index. Testing Psicometria Metodologia, 13(2), 107-125.

Alicke, M. D. (2000). Culpable control and the psychology of blame. Psychological bulletin, 126(4), 556.

Anderson, J. M., Kalra, N., Stanley, K. D., Sorensen, P., Samaras, C., and Oluwatola, O. A. (2014). Brief History and Current State of Autonomous Vehicles. Autonomous Vehicle Technology: A Guide for Policymakers; RAND Corporation: Santa Monica, CA, USA, 55-74.

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

Aridağ, N. C., and Yuksel, A. (2010). Analysis of the Relationship between Moral Judgment Competences and Empathic Skills of University Students. Educational Sciences: Theory and Practice, 10(2), 707-724.

Atiyeh, C. (2012). Predicting traffic patterns, one Honda at a time. MSN Auto, 25, 106-136.

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

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

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., ... and Rahwan, I. (2018b). Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation. arXiv preprint arXiv:1803.07170.

Banks, V. A., Plant, K. L., and Stanton, N. A. (2018). Driver error or designer error: Using the perceptual cycle model to explore the circumstances surrounding the fatal tesla crash on 7th May 2016. Safety Science, 108, 278–285.

Bansal, P., Kockelman, K. M., and Singh, A. (2016). Assessing public opinions of and interest in new vehicle technologies: An Austin perspective. Transportation Research Part C: Emerging Technologies, 67, 1-14.

Barsalou, L. (2003). Situated simulation in the human conceptual system. Language and cognitive processes, 18(5-6), 513-562.

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

Barth, M., and Boriboonsomsin, K. (2009). Energy and emissions impacts of a freeway-based dynamic eco-driving system. Transportation Research Part D: Transport and Environment, 14(6), 400-410.

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

Bazilinskyy, P., Kyriakidis, M., and de Winter, J. (2015). An international crowdsourcing study into people's statements on fully automated driving. Procedia Manufacturing, 3, 2534-2542.

Beanland, V., Fitzharris, M., Young, K. L., and Lenné, M. G. (2013). Driver inattention and driver distraction in serious casualty crashes: Data from the Australian National Crash In-depth Study. Accident Analysis and Prevention, 54, 99-107.

Belliveau, J., Soucy, K. I., and Yakovenko, I. (2022). The validity of qualtrics panel data for research on video gaming and gaming disorder. Experimental and Clinical Psychopharmacology.

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

Benvegnù, G., Pluchino, P., and Gamberini, L. (2021, March). Virtual morality: Using virtual reality to study moral behavior in extreme accident situations. In 2021 IEEE Virtual Reality and 3D User Interfaces (VR) (pp. 316-325). IEEE.

Björklund, F. (2003). Differences in the justification of choices in moral dilemmas: Effects of gender, time pressure and dilemma seriousness. Scandinavian Journal of Psychology, 44(5), 459–466.

Bloom, P. (2011). Family, community, trolley problems, and the crisis in moral psychology. The Yale Review, 99(2), 26-43.

Bobbio, A., Nencini, A., and Sarrica, M. (2011). Il Moral Foundation Questionnaire: Analisi della struttur a fattoriale della versione italiana. Giornale di Psicologia, 5(1), 7-18.

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

Bonner, C., Trevena, L. J., Gaissmaier, W., Han, P. K., Okan, Y., Ozanne, E., ... and Zikmund-Fisher, B. J. (2021). Current best practice for presenting probabilities in patient decision aids: fundamental principles. Medical Decision Making, 41(7), 821-833.

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

Borg, J. S., Hynes, C., Van Horn, J., Grafton, S., and Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. Journal of Cognitive Neuroscience, 18, 803–817.

Bostyn, D. H., and Roets, A. (2022). Sequential decision-making impacts moral judgment: How iterative dilemmas can expand our perspective on sacrificial harm. Journal of Experimental Social Psychology, 98, 104244.

Bradley, M. M., and Lang, P. J. (1994). Measuring emotion: the self- assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry, 25,49–59.

Brandts, J., and Charness, G. (2000). Hot vs. cold: Sequential responses and preference stability in experimental games. Experimental Economics, 2(3), 227-238.

Bridges, D., Pitiot, A., MacAskill, M. R., and Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. PeerJ, 8, e9414.

Briggs, R. A. (2014, August). Normative theories of rational choice: Expected utility. Retrieved August 18, 2022, from https://plato.stanford.edu/entries/rationality-normative-utility/.

Bruers, S., and Braeckman, J. (2014). A review and systematization of the trolley problem. Philosophia, 42(2), 251-269.

Buehler, M., Iagnemma, K., and Singh, S. (Eds.). (2007). The 2005 DARPA grand challenge: the great robot race (Vol. 36). Springer.

Buehler, M., Iagnemma, K., and Singh, S. (Eds.). (2009). The DARPA urban challenge: autonomous vehicles in city traffic (Vol. 56). springer.

Buso, I. M., Di Cagno, D., Ferrari, L., Larocca, V., Lorè, L., Marazzi, F., ... and Spadoni, L. (2021). Lab-like findings from online experiments. Journal of the Economic Science Association, 7(2), 184-193.

Callaway, F., Griffiths, T. L., and Karreskog, G. (2022). Rational Heuristics for One-Shot Games.

Campbell, R. (2017). Learning from moral inconsistency. Cognition, 167, 46-57.

Campbell, M., Egerstedt, M., How, J. P., and Murray, R. M. (2010). Autonomous driving in urban environments: approaches, lessons and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368(1928), 4649-4672.

Carlo, G., Mestre, M. V., Samper, P., Tur, A., and Armenta, B. E. (2010). Feelings or cognitions? Moral cognitions and emotions as longitudinal predictors of prosocial and aggressive behaviors. Personality and Individual Differences, 48(8), 872-877.

Chaiken, S. (1989). Heuristic and systematic information processing within and beyond the persuasion context. Unintended thought, 212-252.

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., and Gomila, A. (2014). Moral judgment reloaded: a moral dilemma validation study. Frontiers in psychology, 5, 607.

Christensen, J. F., and Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. Neuroscience and Biobehavioral Reviews, 36(4), 1249-1264.

Chu, P. C., and Spires, E. E. (2001). Does time constraint on users negate the efficacy of decision support systems? Organizational Behavior and Human Decision Processes, 85, 226–249.

Colman, A. M., Pulford, B. D., and Krockow, E. M. (2018). Persistent cooperation and gender differences in repeated Prisoner's dilemma games: some things never change. Acta psychologica, 187, 1-8.

Cone, J., and Rand, D. G. (2014). Time pressure increases cooperation in competitively framed social dilemmas. PLoS one, 9(12), e115756.

Conway, P., and Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. Journal of personality and social psychology, 104(2), 216.

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

Corey, J. D., Hayakawa, S., Foucart, A., Aparici, M., Botella, J., Costa, A., and Keysar, B. (2017). Our moral choices are foreign to us. Journal of experimental psychology: Learning, Memory, and Cognition, 43(7), 1109.

Côté, S., Piff, P. K., and Willer, R. (2013). For whom do the ends justify the means? Social class and utilitarian moral judgment. Journal of personality and social psychology, 104(3), 490.

Crain, W. C. (1985). Th eories of development. Upper Saddle River, NJ: Prentice-Hall.

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

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. Cognition, 108(2), 353-380.

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

Cusimano, C. J., Magar, S. T., and Malle, B. F. (2017). Judgment Before Emotion: People Access Moral Evaluations Faster than Affective States. In Proceedings of the 39th Annual Conference of the Cognitive Science Society,ed. G Gunzelmann, AHowes, T Tenbrink, EJ Davelaar, pp. 1848–53. Austin, TX: Cogn. Sci. Soc.

Damasio, A. (1994). Descartes' error: Emotion, rationality and the human brain. New York: Putnam, 352.

Danaher, J., Skaug Sætra, H.S. (2022). Technology and moral change: the transformation of truth and trust. Ethics and Information Technology, 24, 35.

Dandurand, F., Shultz, T. R., and Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. Behavior research methods, 40(2), 428-434.

Darwen, P. J., and Yao, X. (1993). On evolving robust strategies for iterated prisoner's dilemma. In Progress in evolutionary computation (pp. 276-292). Springer, Berlin, Heidelberg.

Davidson, P., and Spinoulas, A. (2015). Autonomous vehicles: what could this mean for the future of transport. In Australian Institute of Traffic Planning and Management (AITPM) National Conference, Brisbane, Queensland.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. Journal of Personality and Social Psychology, 44, 113–126.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. JSAS Catalog of Personality and Social Psychology, 36, 752-766.

Davis, F.D., Bagozzi, R.P., Warshaw, P.R., (1989). User acceptance of computer technology: a comparison of two theoretical models. Manage. Sci. 35 (8), 982–1003.

De Freitas, J., Anthony, S. E., Censi, A., and Alvarez, G. A. (2020). Doubting driverless dilemmas. Perspectives on psychological science, 15(5), 1284-1288.

De Freitas, J., Censi, A., Walker Smith, B., Di Lillo, L., Anthony, S. E., and Frazzoli, E. (2021). From driverless dilemmas to more practical commonsense tests for automated vehicles. Proceedings of the national academy of sciences, 118(11), e2010202118.

Dellenback, S. (2013). Director, intelligent systems department, automation and data systems division, southwest research institute. Communication by email, 26(05), 2013.

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

De Winter, J. C., Happee, R., Martens, M. H., and Stanton, N. A. (2014). Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. Transportation research part F: traffic psychology and behaviour, 27, 196-217.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114.

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

Djeriouat, H., and Trémolière, B. (2014). The Dark Triad of personality and utilitarian moral judgment: The mediating role of Honesty/Humility and Harm/Care. Personality and Individual Differences, 67, 11-16.

Dovidio, J. F., Piliavin, J. A., Schroeder, D. A., and Penner, L. A. (2006). The social psychology of prosocial behavior. Mahwah, NJ: Lawrence Erlbaum Associates.

Edland, A., and Svenson, O. (1993). Judgment and decision making under time pressure. In Time pressure and stress in human judgment and decision making (pp. 27-40). Springer, Boston, MA.

Eisenberg, N., Guthrie, I. K., Cumberland, A., Murphy, B. C., Shepard, S. A., Zhou, Q., and Carlo, G. (2002). Prosocial development in early adulthood: a longitudinal study. Journal of personality and social psychology, 82(6), 993.

Eisenberg, N., and Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. Psychological bulletin, 101(1), 91.

Ekman, P. (1992). Are there basic emotions? Psychological Review, 99, 550-553

Ekman, P. E., and Davidson, R. J. (1994). The nature of emotion: Fundamental questions. Oxford University Press.

Eriksson, A., and Stanton, N. A. (2017). Takeover time in highly automated vehicles: noncritical transitions to and from manual control. Human factors, 59(4), 689-705.

European Transport Safety Council (2022, June). Ranking EU progress on road safety 16th. Road Safety Performance Index Report. Retrieved June 22, 2022, from https://etsc.eu/16th- annual-road-safety-performance-index-pin-report/

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

Fagnant, D. J., and Kockelman, K. M. (2018). Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. Transportation, 45(1), 143-158.

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

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., ... and König, P. (2019). Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. Science and engineering ethics, 25(2), 399-418.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. Oxford review, 5.

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

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

Fritz, A., Brandt, W., Gimpel, H., and Bayer, S. (2020). Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI). De Ethica, 6(1), 3-22.

Fumagalli, M., Ferrucci, R., Mameli, F., Marceglia, S., Mrakic-Sposta, S., Zago, S., ... and Priori, A. (2010). Gender-related differences in moral judgments. Cognitive processing, 11(3), 219-226.

Geistfeld, M. A. (2017). A roadmap for autonomous vehicles: State tort liability, automobile insurance, and federal safety regulation. Calif. L. Rev., 105, 1611.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. Psychological science in the public interest, 8(2), 53-96.

Gigerenzer, G., and Galesic, M. (2012). Why do single event probabilities confuse patients? Bmj, 344.

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

Gilligan, C. (1982). In a different voice: Psychological theory and women’ s development. Cambridge, MA: Harvard University Press.

Giner-Sorolla, R., Kupfer, T., and Sabo, J. (2018). What makes moral disgust special? An integrative functional review. In Advances in experimental social psychology (Vol. 57, pp. 223-289). Academic Press.

Gkartzonikas, C., and Gkritza, K. (2019). What have we learned? A review of stated preference and choice studies on autonomous vehicles. Transportation Research Part C: Emerging Technologies, 98, 323-337.

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

Goeschl, T., and Lohse, J. (2018). Cooperation in public good games. Calculated or confused? European Economic Review, 107, 185–203.

Gogoll, J., and Müller, J. F. (2017). Autonomous cars: in favor of a mandatory ethics setting. Science and engineering ethics, 23(3), 681-700.

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

Goodman, L. A. (1961). Snowball sampling. The annals of mathematical statistics, 148-170.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In Advances in experimental social psychology (Vol. 47, pp. 55-130). Academic Press.

Graham, J., Haidt, J., and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of personality and social psychology, 96(5), 1029.

Graham, J., Meindl, P., Beall, E., Johnson, K. M., and Zhang, L. (2016). Cultural differences in moral judgment and behavior, across and within societies. Current Opinion in Psychology, 8, 125-130.

Graham, J., Nosek, B., Haidt, J., Iyer, R., Koleva, S., Ditto, P. (2011). Mapping the moral domain. Journal of Personality and Social Psychology, 101, 366-385.

Grant, A. M., and Shandell, M. S. (2022). Social motivation at work: the organizational psychology of effort for, against, and with others. Annual review of psychology, 73, 301-326.

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

Greene, J. D. (2009). The cognitive neuroscience of moral judgment.

Greene, J. (2014). Moral tribes: Emotion, reason, and the gap between us and them. Penguin.

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

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

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

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

Guériau, M., Billot, R., El Faouzi, N. E., Monteil, J., Armetta, F., and Hassas, S. (2016). How to assess the benefits of connected vehicles? A simulation framework for the design of cooperative traffic management strategies. Transportation research part C: emerging technologies, 67, 266- 279.

Guglielmo, S. (2015). Moral judgment as information processing: an integrative review. Frontiers in psychology, 6, 1637.

Gui, D. Y., Gan, T., and Liu, C. (2016). Neural evidence for moral intuition and the temporal dynamics of interactions between emotional processes and moral cognition. Social neuroscience, 11(4), 380-394.

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

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

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

Haidt, J. (2001). Th e emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review, 108, 814–834.

Haidt, J. (2003a). The emotional dog does learn new tricks: A reply to Pizarro and Bloom (2003).

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

Haidt, J., and Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. Daedalus, 133(4), 55-66.

Haidt, J., and Joseph, C. (2007). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. The innate mind, 3, 367-391.

Hänggi, Y. (2004). Stress and Emotion Recognition: An Internet Experiment Using Stress Induction. Swiss Journal of Psychology/Schweizerische Zeitschrift für Psychologie/Revue Suisse de Psychologie, 63(2), 113.

Harari, Y. N. (2016). Homo Deus: A brief history of tomorrow. random house.

Harman, G., Mason, K., and Sinnott-Armstrong, W. (2010). Moral reasoning. In J. M. Doris and The Moral Psychology Research Group (Eds.), Th e moral psychology handbook (pp. 206–245). Oxford, England: Oxford University Press.

Harper, C. D., Hendrickson, C. T., Mangones, S., and Samaras, C. (2016). Estimating potential increases in travel with autonomous vehicles for the non-driving, elderly and people with travel-restrictive medical conditions. Transportation research part C: emerging technologies, 72, 1-9.

Hauser, M. (2006). Moral minds: How nature designed our universal sense of right and wrong. Ecco/HarperCollins Publishers.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., and Mikhail, J. (2007). A dissociation between moral judgments and justifications. Mind and language, 22(1), 1-21.

Hauser, M. D., Tonnaer, F., and Cima, M. (2009). When moral intuitions are immune to the law: A case study of euthanasia and the act-omission distinction in the Netherlands. Journal of Cognition and Culture, 9, 149–169.

Heaslip, K., Goodall, N. J., Kim, B., and Aad, M. A. (2020). Assessment of Capacity Changes Due to Automated Vehicles on Interstate Corridors (No. FHWA/VTRC 21-R1). Virginia. Dept. of Transportation.

Helion, C., and Ochsner, K. N. (2018). The role of emotion regulation in moral judgment. Neuroethics, 11(3), 297-308.

Hill, W. T., and Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. Quarterly journal of experimental psychology, 65(12), 2343-2368.

Hoeffler, S., and Ariely, D. (1999). Constructing stable preferences: A look into dimensions of experience and their impact on preference stability. Journal of consumer psychology, 8(2), 113-139.

Hoff, K. A., and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407-434.

Holleman, B. (1999). Wording effects in survey research using meta-analysis to explain the forbid/allow asymmetry. Journal of Quantitative Linguistics, 6(1), 29-40.

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

Huber, O., and Kunz, U. (2007). Time pressure in risky decision-making: effect on risk defusing Psychology Science, 49(4), 415.

Huebner, B., Dwyer, S., and Hauser, M. (2009). The role of emotion in moral psychology. Trends in cognitive sciences, 13(1), 1-6.

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

Hulse, L. M., Xie, H., and Galea, E. R. (2018). Perceptions of autonomous vehicles: Relationships with road users, risk, gender and age. Safety science, 102, 1-13.

Hutcherson, C. A., and Gross, J. J. (2011). The moral emotions: A social–functionalist account of anger, disgust, and contempt. *Journal of personality and social psychology*, *100*(4), 719.

Irwin, T. H. (2013). Sympathy and the Basis of Morality. A Companion to George Eliot, 279-293.

Istat (2020), Incidenti stradali in Italia. Retrieved June 22, 2022, from https://www.istat.it/it/archivio/245757

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., and Capiola, A. (2019, July). The measurement of the propensity to trust automation. In International Conference on Human-Computer Interaction (pp. 476-489). Springer, Cham.

Jian, J. Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. International journal of cognitive ergonomics, 4(1), 53-71.

Kahan, D. M., Braman, D., Gastil, J., Slovic, P., and Mertz, C. K. (2007). Culture and identity-protective cognition: Explaining the white-male effect in risk perception. Journal of Empirical Legal Studies, 4(3), 465-505.

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

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

Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

Kallioinen, N., Pershina, M., Zeiser, J., NosratNezami, F., Pipa, G., Ste- phan, A., and König, P. (2019). Moral Judgements on the Actions of Self-Driving Cars and Human Drivers in Dilemma Situations From Different Perspectives. Frontiers in Psychology, 10, 1–15.

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

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

Khemiri, L., Guterstam, J., Franck, J., and Jayaram-Lindström, N. (2012). Alcohol dependence associated with increased utilitarian moral judgment: a case control study. PLoS one, 7(6), e39882.

Kilduff, G. J., Galinsky, A. D., Gallo, E., and Reade, J. J. (2016). Whatever it takes to win: Rivalry increases unethical behavior. Academy of Management Journal, 59(5), 1508-1534.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., and Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In D. C. Noelle, et al. (Eds.), Proceedings of the 37th annual conference of the cognitive science society (pp. 1123–1128). Austin, TX: Cognitive Science Society.

Klenk, M., et al. (2022). Recent work on moral revolutions. Analysis, 82(2), 354–366.

Klenk, M. (2021). The influence of situational factors in sacrificial dilemmas on utilitarian moral judgments. Review of Philosophy and Psychology, 1-33.

Kocher, M. G., and Sutter, M. (2006). Time is money—Time pressure, incentives, and the quality of decision-making. Journal of Economic Behavior and Organization, 61(3), 375-392.

Kohlberg, L. (1964). Development of moral character and moral ideology. Review of child development research, 1, 383-431.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. Handbook of socialization theory and research, 347, 480.

Kohlberg, L. (1981). Th e philosophy of moral development. San Francisco, CA: Harper.

Koopman, P., and Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. IEEE Intelligent Transportation Systems Magazine, 9(1), 90-96.

Kopelias, P., Demiridi, E., Vogiatzis, K., Skabardonis, A., and Zafiropoulou, V. (2020). Connected and autonomous vehicles–Environmental impacts–A review. Science of the total environment, 712, 135237.

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

Korosec, K. (2022, January 5). GM aims to sell personal autonomous vehicles by mid-decade. *TechCrunch.* Retrieved June 21, 2022.

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

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

Krügel, S., and Uhl, M. (2022). Autonomous vehicles and moral judgments under risk. Transportation research part A: policy and practice, 155, 1-10.

Kumar, D.M. (2021, October 11). Standardized Regulatory Framework and Rapid Technological Advancement Set to Propel Autonomous Vehicles Globally. Frost and Sullivan, retrieved 24 June 2022.

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

Kvalnes, Ø. (2015). Moral dilemmas. In Moral reasoning at work: Rethinking ethics in organizations (pp. 9-17). Palgrave Macmillan, London.

Kyriakidis, M., Happee, R., and de Winter, J. C. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. Transportation research part F: traffic psychology and behavior, 32, 127-140.

Lerche, V., Voss, A., and Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. Behavior research methods, 49(2), 513-537.

Leuthold, H., Kunkel, A., Mackenzie, I. G., and Filik, R. (2015). Online processing of moral transgressions: ERP evidence for spontaneous evaluation. Social cognitive and affective neuroscience, 10(8), 1021-1029.

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

Li, S., Zhang, J., Li, P., Wang, Y., and Wang, Q. (2019). Influencing factors of driving decision-making under the moral dilemma. IEEE Access, 7, 104132-104142.

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

Litman, T. (2013). Online Transportation Demand Management Encyclopedia. Victoria Transport Policy Institute retrieved 23 June 2022, from: http://www.vtpi.org/tdm/index.php>.

Litman, T. (2022). Autonomous vehicle implementation predictions. Victoria, BC, Canada: Victoria Transport Policy Institute.

Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. Journal of Computer-Mediated Communication, 26(6), 384-402.

Liu, F., Zhao, F., Liu, Z., and Hao, H. (2019). Can autonomous vehicle reduce greenhouse gas emissions? A country-level evaluation. Energy Policy, 132, 462-473.

Long, D., and Magerko, B. (2020, April). What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-16).

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

Lovibond, S. H., and Lovibond, P. F. (1996). Manual for the depression anxiety stress scales. Psychology Foundation of Australia.

Luccioni, A., and Bengio, Y. (2020). On the morality of artificial intelligence [Commentary]. IEEE Technology and Society Magazine, 39(1), 16-25.

Machery, E., and Mallon, R. (2010). Evolution of morality. In J. M. Doris (Ed.) and Moral Psychology Research Group, The moral psychology handbook (pp. 3–46). Oxford University Press.

Maclagan, P. (2003). Varieties of moral issue and dilemma: A framework for the analysis of case material in business ethics education. Journal of Business Ethics, 48(1), 21–32.

MacKenzie, A. (2022, May). Mercedes-Benz Drive Pilot Level 3 Autonomous First "Drive": We Try a World's First Driverless System. Motortrend. retrieved June 22, 2022.

Maibom, H. L. (2009). Feeling for others: Empathy, sympathy, and morality. Inquiry, 52(5), 483-499.

Malle, B. F. (2021). Moral judgments. Annual Review of Psychology, 72(1), 293-318.

Malle, B. F., Guglielmo, S., and Monroe, A. E. (2014). A theory of blame. Psychological Inquiry, 25(2), 147-186.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 117-124). IEEE.

Mallon R, Nichols S (2011). Dual processes and moral rules. Emotion Review 3: 284–285

Marchant, G. M., and Lindor, R. A. (2012). The coming collision between autonomous vehicles and the liability system. Santa Clara Law Review, 52, 1321.

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

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

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

Massar, M., Reza, I., Rahman, S. M., Abdullah, S. M. H., Jamal, A., and Al-Ismail, F. S. (2021). Impacts of autonomous vehicles on greenhouse gas emissions—positive or negative?. International Journal of Environmental Research and Public Health, 18(11), 5567.

Maxcy, S. J. (2002). Ethical school leadership. RandL Education.

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

McConnell, T. (2002). Moral dilemmas. The Stanford Encyclopedia of Philosophy (2018 Edition), Edward N. Zalta (ed.), retrieved on 29 June 2022 from: https://plato.stanford.edu/entries/moral-dilemmas/

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

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

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

Merat, N., Jamson, A. H., Lai, F. C., Daly, M., and Carsten, O. M. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. Transportation research part F: traffic psychology and behaviour, 27, 274-282.

Mercedes-Benz (2022, December 9). First internationally valid system approval for conditionally automated driving: Mercedes-Benz Group. Mercedes. Retrieved June 22, 2022, from https://group.mercedes-benz.com/innovation/product-innovation/autonomous-driving/system-approval-for-conditionally-automated-driving.html

Messick, D. M., and McClintock, C. G. (1968). Motivational bases of choice in experimental games. Journal of experimental social psychology, 4(1), 1-25.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. Trends in cognitive sciences, 11(4), 143-152.

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

Moberg, D. J. (2000). Time pressure and ethical decision making: The case for moral readiness. Business and Professional Ethics Journal, 19, 41–67.

Moehler, M. (2018). The Rawls–Harsanyi dispute: A moral point of view. Pacific Philosophical Quarterly, 99(1), 82-99.

Montealegre, A., and Jimenez-Leal, W. (2019). The role of trust in the social heuristics hypothesis. PloS one, 14(5), e0216329.

Moody, J., Bailey, N., and Zhao, J. (2020). Public perceptions of autonomous vehicle safety: An international comparison. Safety science, 121, 634-650.

Moore, A. B., Clark, B. A., and Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. Psychological science, 19(6), 549-557.

Morgenstern, O., and Von Neumann, J. (1944). Theory of games and economic behavior. Princeton university, 18-625

Muda, R., Niszczota, P., Białek, M., and Conway, P. (2018). Reading dilemmas in a foreign language reduces both deontological and utilitarian response tendencies. Journal of Experimental Psychology: Learning, Memory, and Cognition, 44(2), 321.

Mukhopadhyay, M., Ghosh, K., Chakraborty, A., and Goswami, M. (2020). Disanalogical discourse on Trolley Problem for Autonomous vehicles. Available at SSRN 3563378.

Mullen, E., and Monin, B. (2016). Consistency versus licensing effects of past moral behavior. Annual review of psychology, 67(1), 363-385.

Munster, G. (2017). Here’s when having a self-driving car will be a normal thing. Fortune Magazine, 13.

Murphy, R. O., and Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. Personality and Social Psychology Review, 18(1), 13-41.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. (2011). Measuring social value orientation. Judgment and Decision making, 6(8), 771-781.

National Highway Traffic Safety Administration (2008). National Motor Vehicle Crash Causation Survey. U.S. Department of Transportation, Report DOTHS 811 059.

National Highway Traffic Safety Administration (2017). Automated driving systems 2.0 a vision for safety. National Highway Traffic Safety Administration, U.S. Department of Transportation.

National Highway Traffic Safety Administration (2020) Traffic Safety Facts: Overview of Motor Vehicle Crashes in 2019. U.S. Department of Transportation, Report DOTHS 813 060.

Navarrete, C. D., McDonald, M. M., Mott, M. L., and Asher, B. (2012). Virtual morality: Emotion and action in a simulated three-dimensional “trolley problem”. Emotion, 12(2), 364.

Nejati, V., Majdi, R., Salehinejad, M. A., and Nitsche, M. A. (2021). The role of dorsolateral and ventromedial prefrontal cortex in the processing of emotional dimensions. Scientific Reports, 11(1), 1-12.

Newman, A., Bavik, Y. L., Mount, M., and Shao, B. (2021). Data collection via online platforms: Challenges and recommendations for future research. Applied Psychology, 70(3), 1380-1402.

Nichols, S., and Mallon, R. (2006). Moral dilemmas and moral rules. Cognition, 100(3), 530-542

Olinga, L. (2022, January 27). Elon Musk promises full self-driving Teslas in 2022. *The Street.* Retrieved June 22, 2022.

Ordóñez, L. D., Benson III, L., and Pittarello, A. (2015). Time-pressure perception and decision making. The Wiley Blackwell handbook of judgment and decision making, 2, 517-542.

Othman, K. (2022). Exploring the implications of autonomous vehicles: A comprehensive review. Innovative Infrastructure Solutions, 7(2), 1-32.

Palan, S., and Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance, 17, 22-27.

Palmiotti, G. P., Del Popolo Cristaldi, F., Cellini, N., Lotto, L., and Sarlo, M. (2020). Framing the outcome of moral dilemmas: effects of emotional information. Ethics and Behavior, 30(3), 213-229.

Panagiotopoulos, I., and Dimitrakopoulos, G. (2018). An empirical investigation on consumers' intentions towards autonomous driving. Transportation research part C: emerging technologies, 95, 773-784.

Parker, C., Scott, S., and Geddes, A. (2019). Snowball sampling. SAGE research methods foundations.

Pastötter, B., Gleixner, S., Neuhauser, T., and Bäuml, K. H. T. (2013). To push or not to push? Affective influences on moral judgment depend on decision frame. Cognition, 126(3), 373-377.

Patil, I. (2015). Trait psychopathy and utilitarian moral judgement: The mediating role of action aversion. Journal of Cognitive Psychology, 27(3), 349-366.

Patil, I., and Silani, G. (2014). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. Frontiers in psychology, 5, 501.

Paxton, J. M., and Greene, J. D. (2010). Moral reasoning: Hints and allegations. Topics in Cognitive Science, 2, 511–527.

Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. Journal of Experimental Social Psychology, 70, 153-163.

Pérez-Manrique, A., and Gomila, A. (2018). The comparative study of empathy: sympathetic concern and empathic perspective-taking in non-human animals. Biological Reviews, 93(1), 248-269.

Peters, E., Hart, P. S., and Fraenkel, L. (2011). Informing patients: the influence of numeracy, framing, and format of side effect information on risk perceptions. Medical Decision Making, 31(3), 432-436.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., and Dickert, S. (2006). Numeracy and decision making. Psychological Science, 17, 407–413.

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

Petrović, Đ., Mijailović, R., and Pešić, D. (2020). Traffic accidents with autonomous vehicles: type of collisions, manoeuvres and errors of conventional vehicles' drivers. Transportation research procedia, 45, 161-168.

Pfattheicher, S., Nielsen, Y. A., and Thielmann, I. (2022). Prosocial behavior and altruism: A review of concepts and definitions. Current opinion in psychology, 44, 124-129.

Pfleging, B., and Schmidt, A. (2015). driving-related activities in the car: Defining driver activities for manual and automated driving. In Workshop on experiencing autonomous vehicles: Crossing the boundaries between a drive and a ride at CHI (Vol. 15).

Piao, J., McDonald, M., Hounsell, N., Graindorge, M., Graindorge, T., and Malhene, N. (2016). Public views towards implementation of automated vehicles in urban areas. Transportation research procedia, 14, 2168-2177.

Pingol, E. (2021, August 20). Level 4 Autonomous Cars Allowed on German Roads. Trend Micro. Retrieved 24 June 2022.

Pizarro, D. A., and Bloom, P. (2003). The intelligence of the moral intuitions: A reply to Haidt (2001). Psychological Review, 110, 193–196.

Pleskac, T. J., and Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. Psychological review, 117(3), 864.

Pletti, C., Lotto, L., Tasso, A., and Sarlo, M. (2016). Will I regret it? Anticipated negative emotions modulate choices in moral dilemmas. Frontiers in psychology, 7, 1918.

Prelec, D., and Loewenstein, G. (1991). Decision making over time and under uncertainty: A common approach. Management science, 37(7), 770-786.

Pretus, C., Hamid, N., Sheikh, H., Gómez, Á., Ginges, J., Tobeña, A., ... and Atran, S. (2019). Ventromedial and dorsolateral prefrontal interactions underlie will to fight and die for a cause. Social cognitive and affective neuroscience, 14(6), 569-577.

Quinn, W., 1989. Actions, intentions and the doctrine of doing and allowing. Philosophical Review 98, 287–312

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Radlmayr, J., Gold, C., Lorenz, L., Farid, M., and Bengler, K. (2014, September). How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving. In Proceedings of the human factors and ergonomics society annual meeting (Vol.58, No. 1, pp. 2063-2067). Sage CA: Los Angeles, CA: Sage Publications.

Raihani, N. J., and Bshary, R. (2011). Resolving the iterated prisoner's dilemma: theory and reality. Journal of Evolutionary Biology, 24(8), 1628-1639.

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous giving and calculated greed. Nature, 489(7416), 427-430.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., and Greene, J. D. (2014). Social heuristics shape intuitive cooperation. Nature communications, 5(1), 1-12

Randsazzo, R. (2019, March 19). Family of woman killed in crash with self-driving Uber sues Arizona, Tempe. Arizona Republican, Retrieved 8 August 2022

Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: Theory and data for two- choice decision tasks. Neural Computation, 20(4), 873–922.

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

Rhim, J., Lee, J. H., Chen, M., and Lim, A. (2021). A deeper look at autonomous vehicle ethics: an integrative ethical decision-making framework to explain moral pluralism. Frontiers in Robotics and AI, 8, 632394.

Richardson, E., and Davies, P. (2018). The changing public's perception of self-driving cars.

Rieskamp, J., and Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. Acta Psychologica, 127(2), 258–276

Riordan, C. A., Marlin, N. A., and Kellogg, R. T. (1983). The effectiveness of accounts following transgression. Social Psychology Quarterly, 213-219.

Riva, P., Manfrinati, A., Sacchi, S., Pisoni, A., and Romero Lauro, L. J. (2019). Selective changes in moral judgment by noninvasive brain stimulation of the medial prefrontal cortex. Cognitive, Affective, and Behavioral Neuroscience, 19(4), 797-810.

Rodier, C. J. (2018). Travel effects and associated greenhouse gas emissions of automated vehicles. UC Davis: National Center for Sustainable Transportation. Retrieved from https://escholarship.org/uc/item/9g12v6r0

Rokeach, M. (1973). The nature of human values. New York: Free Press.

Rosas, A., and Aguilar-Pardo, D. (2020). Extreme time-pressure reveals utilitarian intuitions in sacrificial dilemmas. Thinking and Reasoning, 26(4), 534-551.

Roseman, I. J. (1996). Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. Cognition and Emotion, 10(3), 241-278.

Roulin, N. (2015). Don't throw the baby out with the bathwater: Comparing data quality of crowdsourcing, online panels, and student samples. Industrial and Organizational Psychology, 8(2), 190-196.

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

Russell. J. A. (1991). Culture and the categorization of emotions. Psychological Bulletin. 110.426-450.

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

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

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

SAE International (2021). SAE: J3016_202104: Taxonomy and Definitions for Terms Related to Driving Automation System for On-Road Motor Vehicles.

Samuel, S., Yahoodik, S., Yamani, Y., Valluru, K., and Fisher, D. L. (2020). Ethical decision making behind the wheel–a driving simulator study. Transportation research interdisciplinary perspectives, 5, 100147.

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

Schapira, M. M., Davids, S. L., McAuliffe, T. L., and Nattinger, A. B. (2004). Agreement between scales in the measurement of breast cancer risk perceptions. Risk Analysis, 24, 665–673. 8, 108.

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

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

Schoettle, B., and Sivak, M. (2014). A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia. University of Michigan, Ann Arbor, Transportation Research Institute.

Schoettle, B., and Sivak, M. (2015). Motorists' preferences for different levels of vehicle automation. University of Michigan, Ann Arbor, Transportation Research Institute.

Schroeder, D. A., and Graziano, W. G. (2015). The field of prosocial behavior: An introduction and overview.

Sentinel, M. (1926). Phantom Auto'will tour city. The Milwaukee Sentinel, 4.

Shalvi, S., Eldar, O., and Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). Psychological Science, 23(10), 1264–1270.

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

Shaw, A., DeScioli, P., Barakzai, A., and Kurzban, R. (2017). Whoever is not with me is against me: The costs of neutrality among friends. Journal of Experimental Social Psychology, 71, 96–104.

Shchetko, N. (2014). Laser eyes pose price hurdle for driverless cars. The wall street journal, 21.

Shladover, S. E., Su, D., and Lu, X. Y. (2012). Impacts of cooperative adaptive cruise control on freeway traffic flow. Transportation Research Record, 2324(1), 63-70.

Simpson, E. (1974). Moral development research: A case study of scientific cultural bias. Human Development, 17, 81–106.

Simpson, A., Laham, S. M., and Fiske, A. P. (2016). Wrongness in different relationships: Relational context effects on moral judgment. The Journal of Social Psychology, 156, 594–609. doi:10.1080/00224545.2016.1140118

Singh, S., and Saini, B. S. (2021). Autonomous cars: Recent developments, challenges, and possible solutions. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012028). IOP Publishing.

Sinnott-Armstrong, W. (2016). The disunity of morality. Moral brains: The neuroscience of morality, 331-354.

Skulmowski, A., Bunge, A., Kaspar, K., and Pipa, G. (2014). Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study. Frontiers in behavioral neuroscience, 8, 426.

Smith, B. W. (2013). Automated vehicles are probably legal in the United States. Tex. AandM L. Rev., 1, 411.

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

Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., and Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. Cognition, 217, 104890.

Stanchev, P., and Geske, J. (2015). Autonomous cars. History. State of art. Research problems. In International Conference on Distributed Computer and Communication Networks (pp. 1-10). Springer, Cham.

Strohminger, N., Lewis, R. L., and Meyer, D. E. (2011). Divergent effects of different positive emotions on moral judgment. Cognition, 119(2), 295-300.

Sun, H., Jing, P., Zhao, M., Chen, Y., Zhan, F., and Shi, Y. (2020). Research on the Mode Choice Intention of the Elderly for Autonomous Vehicles Based on the Extended Ecological Model. Sustainability, 12(24), 10661.

Suter, R. S., and Hertwig, R. (2011). Time and moral judgment. Cognition, 119(3), 454-458.

Sütfeld, L. R., Gast, R., König, P., and 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. Frontiers in behavioral neuroscience, 11, 122.

Sütfeld, L. R., Ehinger, B. V., König, P., and 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, 14(10), 1–14

Tangney, J. P., Stuewig, J., and Mashek, D. J. (2007). Moral emotions and moral behavior. Annual review of psychology, 58, 345.

Terraciano, A., McCrae, R. R., and Costa Jr, P. T. (2003). Factorial and construct validity of the Italian Positive and Negative Affect Schedule (PANAS). European journal of psychological assessment, 19(2), 131.

Thielmann, I., Spadaro, G., and Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. Psychological Bulletin, 146(1), 30.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. The Monist, 59,204–217.

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

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

Thomson, R., Yuki, M., Talhelm, T., Schug, J., Kito, M., Ayanian, A. H., ... and Visserman, M. L. (2018). Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat. Proceedings of the National Academy of Sciences, 115(29), 7521-7526.

Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., ... and Johannesson, M. (2013). Intuition and cooperation reconsidered. Nature, 498(7452), E1-E2.

Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., and Västfjäll, D. (2016). Intuition and moral decision-making–the effect of time pressure and cognitive load on moral judgment and altruistic behavior. PloS one, 11(10), e0164012.

Trevena, L. J., BPsych, H. M. D., Barratt, A., Butow, P., and Caldwell, P. (2006). A systematic review on communicating with patients about evidence. Journal of evaluation in clinical practice, 12(1), 13-23.

Trope, Y., and Liberman, N. (2010). Construal-level theory of psychological distance. Psychological review, 117(2), 440.

Ugazio, G., Lamm, C., and Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion*, 12(3), 579.

Uhlemann, E. (2022). Legislation Supports Autonomous Vehicles But Not Connected Ones [Connected and Automated Vehicles]. IEEE Vehicular Technology Magazine, 17(2), 112-115.

Useche, S. A., Ortiz, V. G., and Cendales, B. E. (2017). Stress-related psychosocial factors at work, fatigue, and risky driving behavior in bus rapid transport (BRT) drivers. Accident Analysis and Prevention, 104, 106-114.

van de Poel, I., and Kudina, O. (2022). Understanding technology-induced value change: A pragmatist proposal. Philosophy and Technology 35, 40 (2022). https://

Vianello, M., Galliani, E. M., and Haidt, J. (2010). Elevation at work: The effects of leaders' moral excellence. The Journal of Positive Psychology, 5(5), 390-411.

Vivoli, R., Bergomi, M., Rovesti, S., Bussetti, P., and Guaitoli, G. M. (2006). Biological and Behavioral Factors Affecting Driving Safety. Journal of Preventive Medicine and Hygiene 2006, 47, (pp. 69–73).

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

Waldmann, M.R., Dieterich, J.H., 2007. Throwing a bomb on a person versus throwing a person on a bomb – intervention myopia in moral intuitions. Psychological Science 18 (3), 247–253.

Waldmann, M. R., Nagel, J., and Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak and R. G. Morrison (Eds.), The Oxford handbook of thinking and reasoning (pp. 364–389). Oxford University Press.

Walsh, E. (2021). Moral Emotions. Encyclopedia of Evolutionary Psychological Science, 5209-5216.

Wang, B., Rau, P. L. P., and Yuan, T. (2022). Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. Behaviour and Information Technology, 1-14.

Waters, E. A., Weinstein, N. D., Colditz, G. A., and Emmons, K. (2006). Formats for improving risk communication in medical tradeoff decisions. Journal of health communication, 11(2), 167-182.

Watkins, H. M. (2020). The morality of war: A review and research agenda. Perspectives on Psychological Science, 15(2), 231-249.

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

Weber, M. (2014). Where to? A History of Autonomous Vehicles. Computer History Museum. Computerhistory. org.

Weinstein, N. D., Atwood, K., Puleo, E., Fletcher, R., Colditz, G., and Emmons, K. M. (2004). Colon cancer: risk perceptions and risk communication. Journal of health communication, 9(1), 53-65.

White, L., Jr. (1962). Medieval technology and social change. OUP

Williams, E., Das, V., and Fisher, A. (2020). Assessing the sustainability implications of autonomous vehicles: Recommendations for research community practice. Sustainability, 12(5), 1902.

Wilson, D. S., Near, D., and Miller, R. R. (1996). Machiavellianism: A synthesis of the evolutionary and psychological literatures. Psychological Bulletin, 119, 285–299.

Wintersberger, P., Riener, A., and Frison, A. K. (2016, October). Automated driving system, male, or female driver: Who'd you prefer? comparative analysis of passengers' mental conditions, emotional states and qualitative feedback. In Proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications (pp. 51-58).

Woollard, F. (2012). The doctrine of doing and allowing II: The moral relevance of the doing/allowing distinction. Philosophy Compass, 7(7), 459-469.

Zabat, M., Stabile, N., Farascaroli, S., and Browand, F. (1995). The aerodynamic performance of platoons: A final report. California PATH Research Report, UCB-ITS-PRR-95-35.

Zavala, A. M., Day, G. E., Plummer, D., and Bamford-Wade, A. (2017). Decision-making under pressure: medical errors in uncertain and dynamic environments. Australian Health Review, 42(4), 395- 402.

Zelazo, P.D., Helwig, Ch.C., Lau, A., 1996. Intention, act, and outcome in behavioral prediction and moral judgment. Child Development 67, 2478–2492.

Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R., and Zhang, W. (2019). The roles of initial trust and perceived risk in public's acceptance of automated vehicles. Transportation research part C: emerging technologies, 98, 207-220.

Zhao, L., and Li, W. (2020). "Choose for No Choose"—Random-Selecting Option for the Trolley Problem in Autonomous Driving. In LISS2019 (pp. 665-672). Springer, Singapore.