RIVISTA INTERNAZIONALE DI FILOSOFIA E PSICOLOGIA
DOI: 10.4453/rifp.2020.0015
ISSN 2039-4667; E-ISSN 2239-2629
Vol. 11 (2020), n. 2, pp. 238-250

RICERCHE

Moral dilemmas in self-driving cars

Chiara Lucifora,[(α)] Giorgio Mario Grasso,[(α)] Pietro Perconti[(α)] & Alessio Plebe[(α)]

Ricevuto: 14 ottobre 2019; accettato: 6 maggio 2020

Abstract Autonomous driving systems promise important changes for future of transport, primarily through the reduction of road accidents. However, ethical concerns, in particular, two central issues, will be key to their successful development. First, situations of risk that involve inevitable harm to passengers and/or bystanders, in which some individuals must be sacrificed for the benefit of others. Secondly, and identification responsible parties and liabilities in the event of an accident. Our work addresses the first of these ethical problems. We are interested in investigating how humans respond to critical situations and what reactions they consider to be morally right or at least preferable to others. Our experimental approach relies on the trolley dilemma and knowledge gained from previous research on this. More specifically, our main purpose was to test the difference between what human drivers actually decide to do in an emergency situations whilst driving a realistic simulator and the moral choices they make when they pause to consider what they would do in the same situation and to better understand why these choices may differs.
KEYWORDS: Self-driving Cars; Trolley Problem; Moral Choices; Moral Responsibility; Virtual Reality

Riassunto *Dilemmi morali nelle automobili a guida autonoma* – I sistemi di guida autonomi promettono importanti cambiamenti per il futuro dei trasporti, principalmente attraverso la riduzione degli incidenti stradali. Tuttavia, vi sono preoccupazioni etiche, in particolare due questioni centrali, fondamentali per il loro sviluppo. In primo luogo, le situazioni di rischio che comportano inevitabili danni ai passeggeri e/o ai pedoni, ovvero situazioni in cui alcune persone devono essere sacrificate a beneficio di altri. In secondo luogo, l'identificazione delle parti responsabili in caso di incidente. Il nostro lavoro affronta il primo di questi problemi etici. Siamo interessati a studiare come gli umani rispondono a situazioni critiche e quali reazioni considerano moralmente giuste o almeno preferibili. Il nostro approccio sperimentale si basa sul *trolley problem* e sulle conoscenze acquisite da precedenti ricerche su questo ambito. Più specificamente, il nostro scopo principale è quello di testare la differenza tra ciò che i conducenti umani decidono effettivamente di fare in una situazione di emergenza, mentre guidano un simulatore realistico, e le scelte morali che compiono se posti nella stessa situazione e hanno la possibilità di decidere senza limiti di tempo. Lo scopo è inoltre comprendere come e perché queste scelte possono differire.
PAROLE CHIAVE: Automobili a guida autonoma; Trolley problem; Scelte morali; Responsabilità morale; Realtà virtuale

[(α)]Dipartimento di Scienze Cognitive, Psicologiche, Pedagogiche e Studi Culturali, Università degli Studi di Messina, via Concezione, 6 - 98121 Messina (I)

E-mail: clucifora@unime.it (✉); gmgrasso@unime.it; perconti@unime.it; aplebe@unime.it

1 Introduction

THE DREAM OF SELF-DRIVING CARS was presented for the first time in 1939 at the New York Expo, where engineer Norman Bel Geddes presented his project "Futurama" on radio-controlled vehicles.[1]

Autonomous driving systems can be defined as systems that are able to satisfy the requirements of driving a traditional car, in the absence of any responsibility on the part of real people. They are based on systems that are able to analyse "sensory" data acquired by devices such as video cameras, radar, lidar, and navigation systems; and algorithms – often based on deep neural networks – capable of recognizing relevant objects such as lanes, signs, and other vehicles. Based on this information, the control component of the system should predict the optimal trajectory for the car and issue the appropriate commands for lateral (steering) and longitudinal (cruise velocity) control.

Today, the companies with the largest experience in autonomous vehicles are Tesla, Google, Nissan, Ford, General Motors, BMW, Mercedes, Bosh, and Uber. It is possible to distinguish 6 levels of autonomous driving (SAE International, 2014): total absence of autonomy (*level 0*), driver assistance *(level 1*), partial automation (*level 2*), specific conditions for automation in normal driving situations (*level 3 / level 4*), and complete automation (*level 5*). Devices with the fifth level of automation may be available within a few years.[2]

The development of autonomous systems presents many important advantages but also faces several obstacles. Among the envisaged advantages is the reduction of road accidents. McKinsey & Company estimates that, by reducing traffic accidents, the introduction of self-driving cars would save the US government two hundred billion dollars a year. However, the available data is still insufficient to refute or corroborate this hypothesis. Among other advantages, we can include improved traffic flow, reduction of road pollution, and improved mobility for people with disabilities. The specific obstacle addressed here is the lack of generally accepted moral rules for handling unavoidable accidents.

This is an awkward issue for self-driving cars engineers, but a rare opportunity for philosophers to step away from armchair discussions. Not surprisingly, several moral philosophers and moral psychologists are now turning their attention to self-driving cars.

One of the most obvious prerequisites for designing a moral component for automated vehicles is a reliable knowledge of the diffuse morality behind the critical driving decisions humans make. It is reasonable to require autonomous vehicles to be better at avoiding accidents than humans; indeed, this is a widespread expectation. Including well-defined rules for moral behaviour and decisions in the control software of self-driving cars has a strong positive impact on public opinion.[3]

Recently, Bonnefon and colleagues[4] formulated three requirements for moral algorithms for self-driving cars: be consistent; do not cause public outrage; do not discourage buyers. However, before trying to make machines that attain a higher "moral" benchmark than humans, we need a better understanding of the morality that guides human decisions. Our study aims to address this critical issue. We have chosen to work with the trolley dilemma, for reasons explained in the next section. We made an effort to set up our trolley-like experiment in the most ecological way possible, using virtual reality. A specific objective of our study is to compare the behaviour of people in two different conditions: when they are driving using a simulator and make their moral choices on the spot in critical situations, and when they can pause to reflect on alternative choices and respond without the high arousal elicited by this dangerous situation.

2 Relevance of the trolley dilemma for driverless cars

First of all, let us introduce the original "trolley problem" by considering Philippa

Foot's seminal formulation of this dilemma: «you're standing by the side of a track when you see a runaway train hurtling toward you: clearly the brakes have failed. Ahead are five people, tied to the track. If you do nothing, the five will be run over and killed. Luckily, you are next to a signal switch: turning this switch will send the out-of-control train down a side track, a spur, just ahead of you. Alas, there's a snag: on the spur you spot one person tied to the track: changing direction will inevitably result in this person being killed. What should you do?».[5] The situation is illustrated in *Figure 1*.

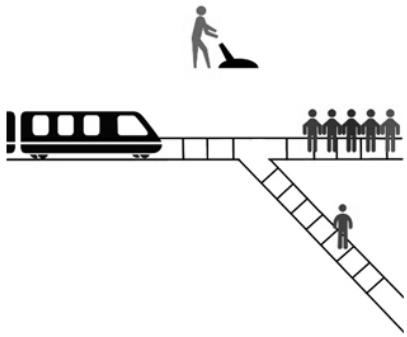

Figure 1. The trolley problem

Most individuals tend to favour quantity rather than quality, preferring the death of a single person to the death of more people. This choice can be understood as "utilitarian", because it is based on the economic principle that implies the duty to maximize expected utility.[6] The same choice might also be made for different reasons, that is because people implicitly embrace the so-called *Double Effect Doctrine* (DDE). This doctrine hypothesizes that a moral choice can be considered legitimate if:

- The act, considered independently of its harmful effects, is not in itself wrong;
- The agents intend good and not harm, even if they can foresee that some harm will happen;
- There is no causal link between the negative and positive effects;
- The harmful effects are not greater than the good sought.

And yet the principles behind utilitarianism are not shared by all people in all circumstances. In fact, cultural differences play an important role in our moral considerations. For example, it has been noted that the Chinese are less likely to make a utilitarian choice or to consider such a choice to be morally correct if a potential human intervention interferes with outcomes predetermined by fate.[7]

Moreover, the way in which a dilemma is presented can change the responses participants provide. For example, in the fat man's dilemma[8] the situation requires direct involvement by participants: «You're on a footbridge overlooking the railway track. You see the trolley hurtling along the track and, ahead of it, five people tied to the rails. Can these five be saved? [...] There's a very fat man leaning over the railing watching the trolley. If you were to push him over the footbridge, he would tumble down and smash on the track below. He's so obese that his bulk would bring the trolley to a shuddering halt. Sadly, the process would kill the fat man. But it would save the other five. Should you push the fat man?».[9] The situation is illustrated in *Figure 2*.

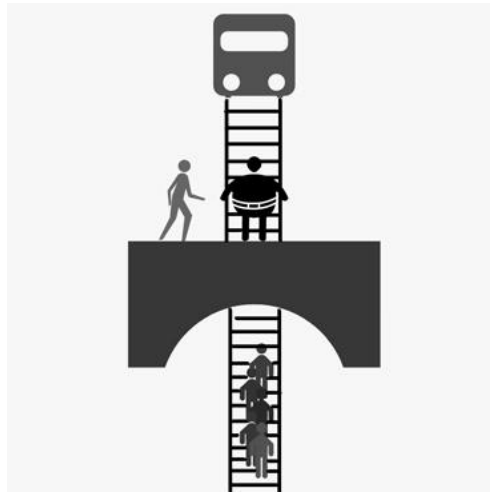

Figure 2. The fat man's dilemma

In this case, the percentage of utilitarian responses decreases because pulling a lever which results in killing one person is not the same as intentionally pushing someone and causing his death. This dilemma has been the foundation of moral philosophy experiments for decades but has become even more popular in the age of cognitive neuroscience, as a suitable stimulus for studying brain mechanisms engaged in moral decision-making.[10]

Self-driving cars appear to offer a new application for this rather old dilemma. There have, however, been some criticisms regarding its use in this context. For example, Johannes Himmelreich argues that this moral dilemma is not relevant to an autonomous car, because the hypothesized situation rarely occurs in the real world, and when it happens, the speed of the car is too great to make a choice in time. A second issue raised in the literature is that, while ethics assumes a top-down approach,[11] engineering control often uses a "bottom-up" approach, in which actions are not guided by explicit rules. Moreover, the moral dilemma ignores some properties of the real world, such as who is accountable for a road accident, and ignores information about special obligations that, in the real world, are very important in terms of understanding the moral permissibility of an act.[12] In addition, since there is no agreement on what moral principles we should all share, resolutions of this the moral dilemma are likely to display too much inter-individual variability to be of any help in devising a general rule for self-driving cars.[13]

Other authors,[14] including ourselves, feel that despite these criticisms and limitations, the trolley dilemma can help us understand moral issues in the context of driving. It is certainly not often that we find ourselves in an emergency situation where we must choose between killing one or five people. However, the trolley dilemma can be seen as a useful idealization of a range of less extreme situations that do occur in the real world, and that require us to make difficult choices.[15] The top-down versus bottom-up argument is also not so compelling, since there are only few studies that have suggested using a top-down approach to simulating morality in autonomous agents (examples include, i.e., Winfield and colleagues as well as Anderson and Anderson).[16]

The argument based on the lack of agreement on ethical standards is relevant, but we don't think it should lead to excluding trolley-like studies. First, a distinction can be made between implicit and explicit ethical agency. It is possible to build agents that act on implicit ethically relevant considerations, but this does not mean that the agent must explain or justify its choices.

More importantly, the lack of ethical standards for critical driving situations can – and should – be bridged by pragmatic standards of human morality. The lack of ethical standards has led many of us to want to know more about how people behave morally when they face driving emergencies.

Studying human morality in the context of driving is important for evaluating social expectations about ethical principles that should guide the behaviour of cars. In the *Moral Machine Experiment*, Awad and colleagues[17] conducted a survey on three groups of participants distinguished by their geographical origins: Western (from North America, European countries); Eastern (from Japan, Saudi Arabia, etc.); Southern (from South America, Central America, etc.). The analysis showed that while the Eastern group was less likely to sacrifice the elderly or make decision based on social status, the Southern group was the least inclined to sacrifice animals and women. Maxmen found important social and economic differences.[18] For example, she showed that people from countries with strong governmental institutions, such as Finland and Japan, choose to kill pedestrians who crossed the road illegally more often than nations with weaker institutions, such as Nigeria or Pakistan. Moreover, when participants had to choose between saving a homeless person on the edge of the road or a leader placed on the opposite side of the same road, choice preferences were correlat-

ed with the level of economic inequality in that society. In Finland, where the gap between rich and poor is relatively small, subjects showed little preference for killing either the homeless person or the manager; on the contrary, in Colombia, where there are important economic disparities, most of the subjects chose to kill the person with the lowest economic status.

3 Moral algorithms

The topic at hand is an instance of the more general problem of designing algorithms instantiating some kind of moral functionality. A minimal account of a moral algorithm must include evaluations regarding the consequences of the agent's behaviour and the ethical principles that the agent can derive from the values and rules encoded in the control program. It is possible to speak about "functional morality",[19] in which the system simply acts within acceptable standards. Floridi and Sanders[20] have identified three important characteristics that artificial agents should exhibit:

- *Interactivity*: the agent responds to the stimulus and changes his state;
- *Autonomy*: the agent changes his state without stimulation;
- *Adaptability*: the agent can learn through experience.

The very notion of moral algorithms is controversial. A key objection is that algorithms lack the kind of understanding, feelings, and emotions that define human relationships, including morality. A common reference for these sorts of objections is the famous argument by John Searle[21] against artificial intelligence. His *Chinese Room* thought experiment allegedly demonstrated that it is possible for a computer to pass the Turing Test, without having any true understanding or genuine intelligence. Based on Searle's mental experiment, several scholars have expressed scepticism about the possibility of building agents who truly understand their actions and can therefore exhibit morality. Those who are most skeptical believe that a human brain/mind is capable of observing proprieties that are critical for morality but cannot be replicated in a silicon machine. One example of such proprieties is responsibility, that is the ability to understand that actions can be harmful. In fact, robots can learn that stimuli belong to protected categories, e.g., military drones must avoid civilians and driverless cars must avoid pedestrians. Thus, robots can have some form of programmed responsibility. Often the list of mental properties alien to machines includes intentionality, a capacity to connect actions to the wishes and reasons that drive behaviour. According to Floridi and Sanders, intentional states are important concerns for the philosophy of mind, but are not necessary to establish moral algorithms. One more item in the list is free will, understood as the ability to do otherwise. Free will is a slippery topic in philosophy, but there are senses in which robots can be considered to exhibit free will. After all, robots are able to change actions plan independently, to learn through experience; these features are included in some accounts of free will.

A further element that traditionally characterizes human moral decisions is the emotions that people experience in relation to their choices.[22] Damasio talks about emotions as "somatic markers" that the brain uses to quickly understand the positive or negative consequences of a choice.[23] While we fully agree with the fundamental role of emotions in human morality, we think that there is no need for a machine to rely on artificial emotions to replicate moral behaviour in the driving context. In the end, what is important is the functional correspondence between the decisions taken by the autonomous car and the ones taken by humans, given that they find themselves in the same critical situation. It doesn't matter how the moral engine is constructed.

4 Our experiment

This section describes the methods used in our study. The participants were tested in a driving simulator, in 15-minute sessions. At the beginning of the simulation, they find themselves inside a virtual garage and stay there for about one minute (see *Figure 3*). This environment allows the participants to become familiar with the simulator, its commands, and the properties of the virtual world.

Figure 3. The garage environment

As anticipated in the Introduction, our experiment comprises two tasks that we define as "hot" and "cold". The hot condition is an ecological simulation of a driving scenario in which an emergency situation occurs. In the cold condition, the driving environment remains highly realistic, but the participant does not continue to drive throughout the emergency situation; instead, the simulation is put on pause, and the participant can take their time to make a choice. This "hot" *vs* "cold" distinction allows us to distinguish a genuine moral decision, driven by contingent emotion, from a deferred, premeditated moral judgment. While the hot decision manifests itself during the emergency situation and is supported by an emotional and unconscious evaluation of events, the cold decision involves a cognitive and conscious evaluation of the possible alternatives and related consequences, which requires time. In the hot task, the participant drives along an extra-urban road with the aim of reaching the city, under specific directions given by the experimenter. They are led along a straight road and have to brake when faced with an inevitable traffic accident when a child, unaware of any danger, suddenly crosses the street. At this point, the user is presented with three alternatives: (A) Hit the child; (B) Swerve to the right and kill three pedestrians on the sidewalk, who are unaware of the dangerous situation; (C) swerve to the left and kill two workers on the roadway (see *Figure 4*).

Figure 4. Critical hot scene

The choice must be made by the participant in a short time, thus constituting what we can call a genuine moral decision. In the deferred – cold – task, the participant is positioned directly inside the city, in a different location. Following specific directions given by the experimenter, the participant arrives at a point where road repair work is in progress, defining a line that crosses the road. Here, while otherwise maintaining the same variables present in the hot case, the simulation stops before the participant enters the accident site. The purpose of this manipulation is to give the participant the time they need to ponder their choice and justify their decision. In this second case, the alternative choices are visible to the participant, and clearly explained by the experimenter: "You found yourself involved in a car accident similar to the previous one. In the first case you reacted instinctively. Now, instead, I ask you to take all the time you need to decide what action to take. You could: (A) go ahead and kill 1 child; (B) swerve to your right and sacrifice 3 people on the sidewalk; (C) swerve to your left and kill 2 workers on the road. I ask you to tell me who you choose to sacrifice

and why" (see *Figure 5*).

Figure 5. Critical cold scene

5 Driving simulator

Our experimental setup consists of a virtual reality helmet, or Oculus Rift, made by Oculus VR, equipped with two Pentile OLED displays, 1080×1200 resolution per eye, 90 Hz refresh rate and 110° field of view. This device features rotation and position tracking as well as integrated headphones that provide 3D sound effects.[24] This peripheral is driven by a graphics workstation, equipped with a NVIDIA Titan X graphics card, used to run the simulation, ensuring uniform high-resolution rendering of the virtual environment is projected onto the VR headset. The driving interface is provided by a Logitech steering wheel with force feedback and has pedals and a gear shift that are realistic enough to provide a complete driving experience during the simulation. The interface is mounted on a setup consisting of a real driving seat positioned on a rigid frame, in which all the equipment is installed.

The driving simulator was developed using software called Unity 3D. A highly detailed city model was used, taken from the free Unity repository, namely Windridge City for AirSim on Unity. Package included (i) Urban roads surrounded by forests and extra-urban roads; (ii) Interconnected roads; (iii) Garden furniture, street signs and buildings.

The model includes a real urban environment accompanied by an extra-urban road, which allows the subject to become familiar with the simulation in the absence of traffic. In addition, we have worked particularly on pedestrians and vehicles to make their behaviour as close to reality as possible, thus respecting the laws of physics in their movement, as well as in collisions and interactions with other objects.

Virtual reality allows a participant to feel immersed in a temporally and spatially different place. It can be defined as a simulated environment, in which the participant experiences telepresence. Many researchers have, in fact, investigated the degree of presence an individual experience by measuring their physiological reactions to the virtual environment. For example, bringing a person into a stressful virtual situation can elicit bodily responses similar to those expected in the real world, such as an increase in heart rate, an increase in skin conductance and a reduction in peripheral temperature. For this reason, we considered the possible side effects of our virtual reality situation and adopted specific measures for participants potentially at risk.[25] Indeed, in our experiment, we found that some participants (11,25%) experienced intolerable motion sickness (due to non-correspondence between the visual stimuli of the movement and the real movement of the body, for example, during virtual acceleration). This disorder was felt to be tolerable by 74% of the sample with an average of M=1.73 on a Likert scale (from 0 "nothing" to 5 "a lot").

6 Preliminary results

To date, experiments have been carried out on a sample of 84 people (from the original sample of 90, 6 people were excluded because they were unable to complete the experiment due to excessive nausea, dizziness, and obvious pallor in the face, typical side effects of VR). The sample (N = 84) consisted of 10 males and 74 females with a mean age of M = 22,25. 71 subjects had a driving license and 33 participants had experience with video games.[26]

The first results show a significant difference in the choices made by the participants in the first (hot) and the second (cold) situation (see *Figure 6*).

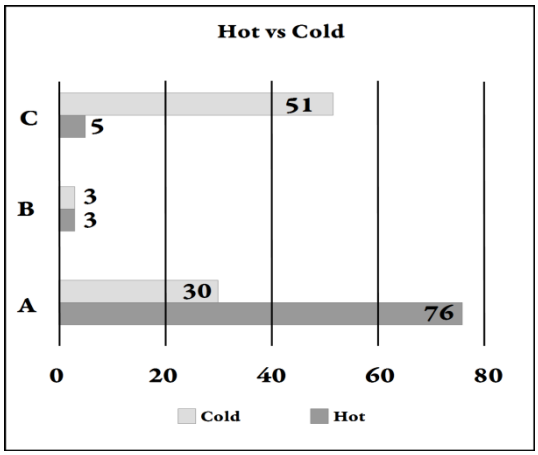

Figure 6. Distribution of motivations accumulated over choices

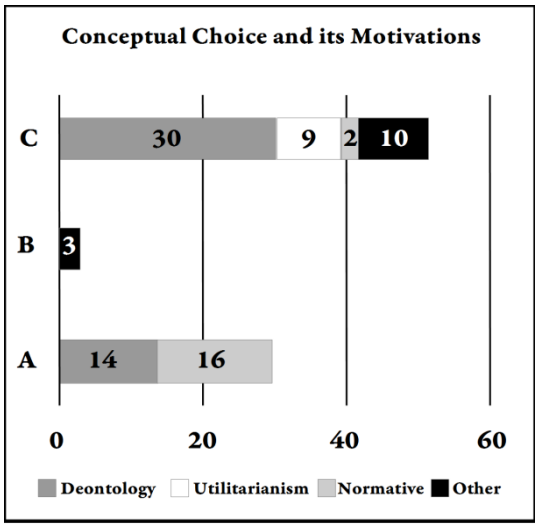

Figure 7: Distribution of motivations in the cold choice

In the realistic hot situation (see *Figure 6*), 92% of participants choose A; in the deferred cold choice, only 35% confirmed this decision, while 61% choose C and 4% choose B. The reported reasons for the meditated choice were classified as: (a) Deontological reasons; (b) Utilitarian reasons; (c) Normative reasons; (d) Other. *Figure 7* shows the distribution of these motivations, grouped by the choice made (A, B, C); in *Figure 8*, the distribution of motivations is shown grouped by type of choice.

Choice A (kill 1 child) was selected by 35% of the participants. For 46% of these participants, this choice was made for utilitarian reasons, "it's better to kill one person, rather than two or three people"; 54% cited normative reasons (relating to compliance with traffic laws). In this case, the participants said "it's the child's fault, because he crossed the street without visual contact with the driver and didn't use the pedestrian crossing.

Choice B (to kill 3 people) was made for other kinds of motivations that had no statistical significance (only 3 participants out of 84).

Choice C (kill 2 road workers) was made by 61% of the participants. 75% offered deontological reasons: they considered the three pedestrians and the child to be a family unit and defended the value of the family; 17% gave utilitarian reasons: they thought that it was preferable to kill two people than three; 3% gave normative reasons, saying that the workers were not allowed to work on a roadway without signalling their presence and taking the appropriate protective measures; and 5% provided other reasons, due to the social role of the workers.

7 Discussion

In general, the deferred choices of participants to date have mainly driven by the high value accorded to family (42,86%). Utilitarian reasons came second, and were motivated by quantity (number of peopled harmed) rather than quality (27, 38%). Next were motivations due to regulations related to compliance with traffic laws (21,43%), and last the reasons driven by other factors (8,33%).

Our results show that while the utilitarian choice had marginal prevalence in the cold task, it dominated in the hot task. This is an overall agreement with the findings of Bonnefon and colleagues.[27] In their study, when people had to choose between the death of 10 pedestrians and the death of a single passenger, they found the "morality of the sacrifice" of a passenger acceptable in 72% of cases. But when the choice balanced the life of one passenger or one pedestrian, only 23% would accept this sacrifice. So, in theory the people are prone to act

in a utilitarian manner when driving, but they prefer protective cars for themselves. Faulhaber and colleagues also found most people preferred to minimize the number of deaths.[28]

Our results provide new insights, showing that other factors also influence this moral choice. In the cold condition, where the deontological can be compared to the normative and utilitarianism choices (see *Figure 8*), our results show a preference for the deontological choice (42,86%), followed by the utilitarian (27,38%), normative (21,43%) and other choices (8,33%). It's possible that immersion in the virtual world during the first part of the task influenced decision-making, so that subjects preferred quality over quantity. Other choices were guided by socio-economic factors. This result could reflect significant discrimination against a specific social class. However, it is possible that the choice to kill two workers on the roadway was conditioned by the presence of the presumed family unit on the scene. In the "hot situation", our results reflect automatic and non-conceptual human behavior; in fact, choice A in the hot situation, seems in accordance with previous theoretical studies which proposed that in emergency driving situations, moral behaviour is combined with rapid risk analysis.[29]

In our virtual scenario, swerving to the right or to left could have caused the participant-driver to lose control of the car, while the less dangerous action would be to try to stop the car.

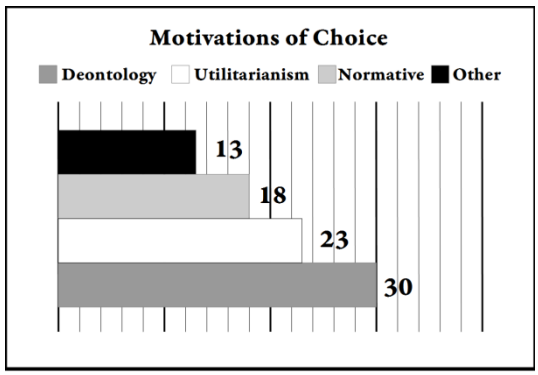

Figure 8. Ecological Choice "hot" *vs* Conceptual Choice "cold"

8 Conclusions

In this study, we described an experiment we ran using a driving simulator, to investigate human morality during driving. We implemented a trolley-like scenario in which the participants had to decide whether (A) to hold the course of the car and kill a child; (B) swerve to their right and sacrifice 3 people on the sidewalk; or (C) swerve to their left and kill 2 workers on the road. The participants took this decision under two different conditions, which we called "hot" and "cold". In the hot condition, they made the decision while driving the car in the simulator, by urgently performing one of the three possible actions. In the cold condition, they drove the car in the simulator only until they reached the site of the accident. Here the simulation stopped and participants were asked by the investigator what they would have done if they had continued driving and had to choose how to react a nearly identical trolley-like situation.

The comparison between the hot and the cold tasks showed that moral decisions taken rapidly during action and those taken slowly after reflection differ widely. In the hot condition, people react mainly according to utilitarian principles, while in the cold condition they chose on the basis of other criteria.

Considering the wide cultural differences pointed out by the literature in the ways in which people react to trolley-like scenarios,[30] we plan to replicate our research in other countries besides Italy in the future. The aim is to further investigate the social and geographical frontiers of human morality. Moreover, we would also like to delve more deeply into the reasons that motivate the different choices made in the hot and cold conditions. In particular, since the decision to hold the course of the car steady is the less risky for the driver, we would like to explore the relevance of self-risk analysis to moral choices we make when confronted with hot as compared to cold scenarios.

Notes

[1] Cf. R. MARCHAND, *The designers go to the fair II: Norman Bed Geddes, The General Motors Futurama, and the visit to the factory transformed*, in: «Design Issue», vol. VIII, n. 2, 1992, pp. 23-40.
[2] Cf. E. CANDELO, *Innovation and digital transformation in the automotive industry*, Springer, Cham 2019, pp. 155-173.
[3] Cf. J.F. BONNEFON, A. SHARIFF, I. RAHWAN, *The social dilemma of autonomous vehicles*, in: «Science», vol. CCCLII, n. 6293, 2016, pp. 1573-1576.
[4] Cf. J.F. BONNEFON, A. SHARIFF, I. RAHWAN, *Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars?*, 2015, arXiv preprint: arXiv:1510.03346.
[5] D. EDMONDS, *Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong*, Princeton University Press, Princeton 2014, p. 9.
[6] Cf. J. VON NEUMANN, O. MORGESTERN, *Theory of games and economic behavior*, Princeton University Press, Princeton 1944.
[7] Cf. H. AHLENIUS, T. TANNSJÖ, *Chinese and Westernes respond differently to the trolley dilemmas*, in: «Journal of Cognition & Culture», vol. XII, n. 3-4, 2012, pp. 195 -201.
[8] Cf. J.J. THOMSON, *The trolley problem*, in: «The Yale Law Journal», vol. XCIV, 1985, pp. 1395-1415.
[9] D. EDMONDS, *Would you kill the fat man?*, cit., p. 37.
[10] Cf. J.D. GREEN, F.A. CUSHMAN, L.E. STEWART, K. LOWENBERG, L.E. NYSTROM, J.D. COHEN, *Pushing moral buttons: The interaction between personal force and intention in moral judgment*, in: «Cognition», vol. CXI, n. 3, 2009, pp. 364-371.
[11] Cf. J. HIMMELREICH, *Never mind the trolley: The ethics of autonomous vehicles in mundane situations*, in: «Ethical Theory & Moral Practice», vol. XXI, n. 3, 2018, pp. 669-684 and F. ALAIERI, A. VELLINO, *Ethical decision making in robots: Autonomy, trust and responsibility*, in: A. AGAH, J.-J. CABIBIHAN, A.M. HOWARD, M.A. SALICHS, H. HE (eds.), *Social robotics*, Springer, Cham 2016, pp. 159-168; C. ALLEN, I. SMIT, W. WALLACH, *Artificial morality: Top-down, bottom-up, and hybrid approaches*, in: «Ethics & Information Technology», vol. VII, n. 3, 2005, pp. 149-155.
[12] Cf. S. NYHOLM, J. SMIDS, *The ethics of accident-algorithms for self-driving cars: An applied trolley problem?*, in: «Ethical Theory & Moral Practice», vol. XIX, n. 5, 2016, pp. 1275-1289.
[13] Cf. J. HIMMELREICH, *Never mind the trolley*, cit.
[14] Cf. G. KEELING, *Why trolley problems mater for the ethics of automated vehicles*, in: «Science & Engineering Ethics», vol. XXVI, n. 1, 2020, pp. 293-307.
[15] Cf. W. WALLACH, C. ALLEN, *Moral machines: Teaching robots right from wrong*, Oxford University Press Oxford 2008.
[16] Cf. A.F. WINFIELD, C. BLUM, W. LIU, *Towards an ethical robot: internal models, consequences and ethical action selection*, in: M. MISTRY, A. LEONARDIS, M. WITKOWSKI, C. MELHUISH (eds.), *Advances in autonomous robotic systems*, *TAROS 2014, Lecture Notes in Computer Science*, Springer, Cham 2014, pp. 85-96; M. ANDERSON, S.L. ANDERSON, *Machine ethics: Creating an ethical intelligent agent*, in: «Artificial Intelligence Magazine», vol. XXVIII, n. 4, 2007, p. 15.
[17] Cf. E. AWAD, S. DSOUZA, R. KIM, J. SCHULZ, J. HENRICH, A. SHARIFF, J.F. BONNEFON, I. RAHWAN, *The moral machine experiment*, in: «Nature», vol. DLXIII, n. 7729, 2018, pp. 59-64.
[18] Cf. A. MAXMEN, *Self-driving car dilemmas reveal that moral choices are not universal*, in: «Nature», vol. DLXII, n. 7728, 2018, pp. 469-470.
[19] Cf. W. WALLACH, C. ALLEN, *Moral machines*, cit.
[20] Cf. L. FLORIDI, J.W. SANDERS, *On the morality of artificial agents*, in: «Mind & Machines», vol. XIV, n. 3, 2004, pp. pp. 349-379.
[21] Cf. J.R. SEARLE, *Minds, Brains, and Programs*, in: «Behavioral & Brain Sciences», vol. III, n. 3, 1980, pp. 417-458.
[22] Cf. D. HUME, *A Treatise of Human Nature* (1739), edited by L.A. SELBY-BIGGE, John Noon, London 1951.
[23] Cf. A.R. DAMASIO, *Descartes' error and the future of human life*, in: «Scientific American», vol. CCLXXI, n. 4, 1994, pp. 144.
[24] Cf. G. GRASSO, C. LUCIFORA, P. PERCONTI, A. PLEBE, *Evaluating mentalization during driving*, in: O. GUSIKHIN, M. HELFERT (ed.), *Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2019)*, Scitepress, Prague 2019, pp. 536-541.
[25] Cf. G. RIVA, B.K. WIEDERHOLD, E. MOLINARI, *Virtual environments in clinical psychology and neuroscience: Methods and techniques in advanced patient-therapist interaction*, IOS press, Amsterdam 1998.
[26] The sample is made up of university students from the Department of Cognitive Science, University of Messina (Italy). The size of our sample

is too small to draw definite conclusions. Unfortunately, the second series of tests scheduled at the beginning of the year, could not be carried out due to the COVID-19 containment restrictions imposed by the Italian Government.

[27] Cf. J.F. BONNEFON, A. SHARIFF, I. RAHWAN, *The social dilemma of autonomous vehicles*, cit.

[28] Cf. A.K. FAULHABER, A. DITTMER, F. BLIND, M.A. WÄCHTER, S. TIMM, L.R. SÜTFELD, P. KÖNIG, *Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles*, in: «Science & Engineering Ethics», vol. XXV, n. 2, 2019, pp. 399-418.

[29] Cf. R. DAVNALL, *Solving the single-vehicle self-driving car trolley problem using risk theory and vehicle dynamics*, in: «Science & Engineering Ethics», vol. XXVI, n. 1, 2020, pp. 431-449; B. MEDER, N. FLEISCHHUT, N.C. KRUMNAU, M.R. WALDMANN, *How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty*, in: «Risk Analysis», vol. XXXIX, n. 2, 2019, pp. 295-314.

[30] Cf. H. AHLENIUS, T. TANNSJÖ, *Chinese and Westerners respond differently to the trolley dilemmas*, cit.

References

AHLENIUS, H., TANNSJÖ, T. (2012). *Chinese and Westernes respond differently to the trolley dilemmas*. In: «Journal of Cognition and Culture», vol. XII, n. 3-4, pp. 195 -201.

ALAIERI, F., VELLINO, A. (2016). *Ethical decision making in robots: Autonomy, trust and responsibility*, in: A. AGAH, J.-J. CABIBIHAN, A.M. HOWARD, M.A. SALICHS, H. HE (eds.), *Social robotics*, Springer, Cham, pp. 159-168.

ALLEN, C., SMIT, I., WALLACH, W. (2005). *Artificial morality: Top-down, bottom-up, and hybrid approaches*. In: «Ethics & Information Technology», vol. VII, n. 3, pp. 149-155.

ANDERSON, M., ANDERSON, S.L. (2007). *Machine ethics: Creating an ethical intelligent agent*, in: «Artificial Intelligence Magazine», vol. XXVIII, n. 4, p. 15.

AWAD, E., DSOUZA, S., KIM, R., SCHULZ, J., HENRICH, J., SHARIFF, A., BONNEFON, J.F., RAHWAN, I. (2018). *The moral machine experiment*. In: «Nature», vol. DLXIII, n. 7729, pp. 59-64.

BONNEFON, J.F., SHARIFF, A., RAHWAN, I. (2015). *Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars?*, arXiv preprint: arXiv:1510.03346.

BONNEFON, J.F., SHARIFF, A., RAHWAN, I. (2016). *The social dilemma of autonomous vehicles*, in: «Science», vol. CCCLII, n. 6293, pp. 1573-1576.

CANDELO, E. (2019). *Innovation and digital transformation in the automotive industry*, Springer, Cham, pp. 155-173.

DAMASIO, A.R. (1994). *Descartes' error and the future of human life*. In: «Scientific American», vol. CCLXXI, n. 4, pp.144.

DAVNALL, R. (2020). *Solving the single-vehicle self-driving car trolley problem using risk theory and vehicle dynamics*, in: «Science & Engineering Ethics», vol. XXVI, n. 1, pp. 431-449.

EDMONDS, D. (2014). *Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong*, Princeton University Press, Princeton.

FAULHABER, A.K., DITTMER, A., BLIND, F., WÄCHTER, M.A., TIMM, S., SÜTFELD, L.R., KÖNIG, P. (2019). *Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles*, in: «Science & Engineering Ethics», vol. XXV, n. 2, pp. 399-418.

FLORIDI, L., SANDERS, J.W. (2004). *On the morality of artificial agents.* In: «Mind & Machines», vol. XIV, n. 3, pp. 349-379.

GRASSO, G., LUCIFORA, C., PERCONTI, P., PLEBE, A. (2019). *Evaluating mentalization during driving*. In: O. GUSIKHIN, M. HELFERT (ed.), *Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2019)*, Scitepress, Prague, pp. 536-541 - doi: 10.5220/0007756505360541.

GREEN, J.D., CUSHMAN, F.A., STEWART, L.E., LOWENBERG, K., NYSTROM, L.E., COHEN, J.D. (2009). *Pushing moral buttons: The interaction between personal force and intention in moral judgment*. In: «Cognition», vol. CXI, n. 3, pp. 364-371.

HIMMELREICH, J. (2018). *Never mind the trolley: The ethics of autonomous vehicles in mundane situations*. In: «Ethical Theory & Moral Practice», vol. XXI, n. 3, pp. 669-684.

HUME, D. (1951). *A Treatise of Human Nature* (1739), edited by L.A. SELBY-BIGGE, John Noon, London.

KEELING, G. (2020). *Why trolley problems mater for the ethics of automated vehicles*. In: «Science and Engineering Ethics», vol. XXVI, n. 1, pp. 293-307.

MARCHAND, R. (1992). *The designers go to the fair II: Norman Bed Geddes, The General Motors Futurama, and the visit to the factory transformed*. In: «Design Issue», vol. VIII, n. 2, pp. 23-40.

MAXMEN, A. (2018). *Self-driving car dilemmas reveal that moral choices are not universal*. In: «Nature», vol. DLXII, n. 7728, pp. 469-470.

MEDER, B., FLEISCHHUT, N., KRUMNAU, N.C., WALDMANN, M.R. (2019). *How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty*. In: «Risk Analysis», vol. XXXIX, n. 2, pp. 295-314.

NYHOLM, S., SMIDS, J. (2016). *The ethics of accident-algorithms for self-driving cars:*

An applied trolley problem?. In: «Ethical Theory & Moral Practice», vol. XIX, n. 5, pp. 1275-1289.

RIVA, G., WIEDERHOLD, B.K., MOLINARI, E. (1998). *Virtual environments in clinical psychology and neuroscience: Methods and techniques in advanced patient-therapist interaction*, IOS press, Amsterdam.

SEARLE, J.R. (1980). *Minds, Brains, and Programs*. In: «Behavioral & Brain Sciences», vol. III, n. 3, pp. 417-458.

THOMSON, J.J. (1985). *The trolley problem*. In: «The Yale Law Journal», vol. XCIV, pp. 1395-1415.

VON NEUMANN, J., MORGESTERN, O. (1944). *Theory of games and economic behavior*, Princeton University Press, Princeton.

WALLACH, W., ALLEN, C. (2008). *Moral machines: Teaching robots right from wrong*, Oxford University Press Oxford.

WINFIELD, A.F., BLUM, C., LIU, W. (2014). *Towards an ethical robot: internal models, consequences and ethical action selection*. In: M. MISTRY, A. LEONARDIS, M. WITKOWSKI, C. MELHUISH (eds.), *Advances in autonomous robotic systems*, *TAROS 2014, Lecture Notes in Computer Science*, Springer, Cham, pp. 85-96 - doi: 10.1007/978-3-319-10401-08.