Would you do it?: Enacting Moral Dilemmas in Virtual Reality for Understanding Ethical Decision-Making

Conference Paper · March 2020

DOI: 10.1145/3313831.3376788

5 authors, including:

Evangelos Niforatos
Norwegian University of Science and Technology
44 PUBLICATIONS **244** CITATIONS
SEE PROFILE

Athanasios Vourvopoulos
Instituto Superior Técnico
42 PUBLICATIONS **364** CITATIONS
SEE PROFILE

Fotis Liarokapis
Masaryk University
145 PUBLICATIONS **2,547** CITATIONS
SEE PROFILE

Would you do it?: Enacting Moral Dilemmas in Virtual Reality for Understanding Ethical Decision-Making

Evangelos Niforatos
Norwegian University of Science and Technology
evangelos.niforatos@ntnu.no

Adam Palma
Robert Bosch
adam.palma@cz.bosch.com

Roman Gluszny
Solarwinds
roman.gluszny@solarwinds.com

Athanasios Vourvopoulos
Instituto Superior Técnico
athanasios.vourvopoulos@tecnico.ulisboa.pt

Fotis Liarokapis
Masaryk University
liarokap@fi.muni.cz

ABSTRACT
A moral dilemma is a decision-making paradox without unambiguously acceptable or preferable options. This paper investigates if and how the virtual enactment of two renowned moral dilemmas—the Trolley and the Mad Bomber—influence decision-making when compared with mentally visualizing such situations. We conducted two user studies with two gender-balanced samples of 60 participants in total that compared between paper-based and virtual-reality (VR) conditions, while simulating 5 distinct scenarios for the Trolley dilemma, and 4 storyline scenarios for the Mad Bomber's dilemma. Our findings suggest that the VR enactment of moral dilemmas further fosters utilitarian decision-making, while it amplifies biases such as sparing juveniles and seeking retribution. Ultimately, we theorize that the VR enactment of renowned moral dilemmas can yield ecologically-valid data for training future Artificial Intelligence (AI) systems on ethical decision-making, and we elicit early design principles for the training of such systems.

Author Keywords
Ethics; moral dilemmas; VR; decision-making; ethical AI.

CCS Concepts
•**Human-centered computing** → **Human computer interaction (HCI);** *Virtual reality;* User studies;

INTRODUCTION
Humanity has been continuously confronted with moral dilemmas ever since the dawn of logical reasoning, dating back to the classical era in ancient Greece (5th and 4th centuries BC). Over time, several schools of thought emerged including the Cynics, the Cyrenaics, Aristotle's school of ethics, the Epicureans and the Stoics, holistically described as "Ancient Ethics"; before we move gradually to contemporary schools such as Kantianism and utilitarianism. In fact, modern philosophers have long been debating whether ancient ethics were even "ethical," as they appear to serve the sole purpose of achieving happiness in the life of the agent (i.e., the individual) [2]. Thus, the term "moral" was introduced for generally describing the notion of "doing the right thing." Since there is no clear-cut distinction between "ethics" and "morality" in Philosophy,[1] we interchangeably use the terms "*ethical*" and "*moral*" for conveying the aforementioned notion.

But why is exploring moral dilemmas relevant today?
Given the propensity of human societies and the structures within (e.g., companies, universities, a football team, etc.) for generating laws, codes, and cultures that outline acceptable behaviors, it is natural for the field of ethics to continue to flourish. In fact, revolutionary technologies continuously disrupt human behavior, exhibited in the aforementioned structures, and radically alter the landscape of ethics that underpin modern life. Nowadays, the unprecedented proliferation of AI leads to outsourcing an ever-increasing number of decisions to intelligent algorithms and systems. However, the increasing gravity of the decisions being outsourced is worrisome. For example, the field of autonomous driving has been tantalized by the "Trolley Problem," where a fully autonomous vehicle perceives an inescapable fatal situation and has to decide selecting from a range of available options that will always entail human casualty [23, 27]. Notably, the autonomous driving system will have no time for receiving human input and should entirely rely on itself for deciding. This example portrays a typical future scenario where a "machine" is confronted with a moral dilemma.

AI has also been utilized in the fields of law and justice. "Coplink" is an AI system that identifies relationships between suspects and victims for facilitating crime investigation and law enforcement efforts [20]. More recently, AI predicted successfully the outcome of real-life human rights cases with a 79 % accuracy after having been trained on previous cases from the European Court of Human Rights [1]. Perhaps in the

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376788

[1] https://www.britannica.com/topic/ethics-philosophy

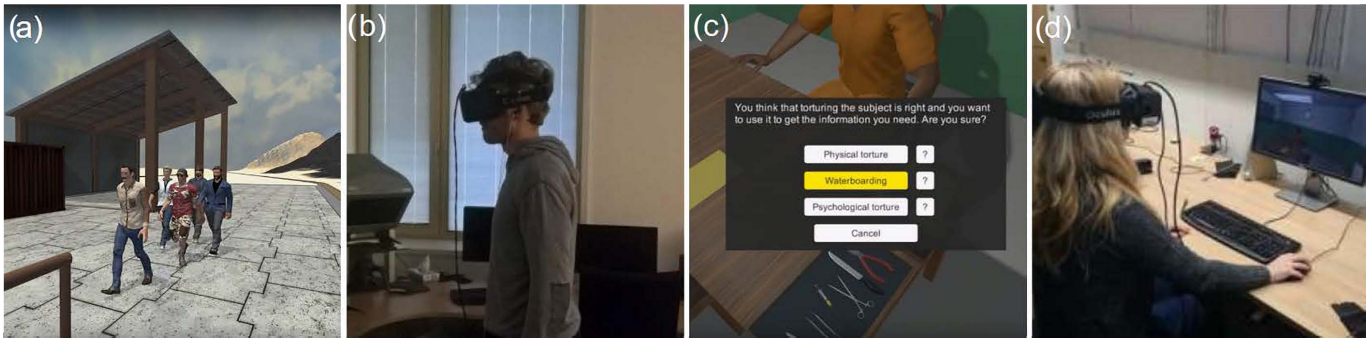

Figure 1. (a) The train platform and the avatars approaching in the Trolley study, (b) A participant in the Trolley study wearing the Oculus DK2 and relying on Kinect for input, (c) Interacting with the bomber avatar in the Mad Bomber study, (d) A participant wearing Oculus DK2 and using a mouse/keyboard for input.

future, AI may assume a more active jurisdictional or even executive role, raising serious ethical concerns. For example, the "Mad Bomber's dilemma" is an ethical conundrum where a presumed terrorist has placed time-bombs in public places and refuses to cooperate with the authorities after his arrest. If information about the bombs is not extracted in a timely manner, the bombs explode and many innocents die. The moral dilemma lies in whether one employs torture for extracting the sought information. This example illustrates a futuristic scenario where a hypothetical AI-judge system may be confronted with a moral dilemma.

Human experience has shown that ethics can be taught [6], and learning from humans (or human examples) is how a large portion of AI is trained in the first place (i.e., supervised learning [7]). Thus, one way of inducing ethics into AI could be by training it with sufficient number of human-made decisions on illustrious moral dilemmas. In this paper, we investigate *if* and *how* the VR enactment of the Trolley and the Mad Bomber dilemmas influence the moral decision-making of our participants for eliciting design principles in training the ethical-AI systems of the future.

RELATED WORK

A moral dilemma is a model situation which features an agent presented with a strictly limited set of action-options, one of which the agent is required to select and perform. Typically, the actions are mutually exclusive—if the agent chooses to perform one action they cannot perform another. For each action, the agent has strong moral reasons so as to why it should be chosen, but at the same time, performing it will result in a moral failure. From this conundrum arises the dilemma, as it is not obvious which action is the best one [25]. There are multiple ways of categorizing moral dilemmas: One of them establishes two categories—**epistemic** and **ontological** [14]. In the epistemic category, the agent faces a conflict between two or more moral principles, and at the time of making the decision does not know which one of them has priority. In general terms, the priority can be determined and agreed upon. In the ontological category however, all principles have equal merit and one can never take priority over any of the others. The Trolley Problem is perhaps the most prominent ontological dilemma.

Moral dilemmas can also be distinguished between **obligation** dilemmas and **prohibition** dilemmas. In obligation dilemmas, more than one action is obligatory, whereas the prohibition dilemmas feature a set of actions which are forbidden [25]. When it comes to evaluating the merits of each action, there are two main views one may consider. The first one is called **deontology**, according to which the moral principles are considered above all else. An example of such a view may be a religious commandment that forbids one from killing another human being under any circumstances. The second one is **utilitarianism**. Here the consequences of each choice are the criterion that determines the best course of action. Under this view, the best choice is usually the one that benefits the society as a whole [18, 15].

The Trolley Dilemma

First introduced by Philippa Foot in 1967, the "Trolley Problem" (in its original form) features a driver of a runaway tram, facing the following problem: If the driver continues on his current route, he is bound to run over five people, who are currently working on the track. He can, however, divert the tram into another track with only one worker on it. The dilemma at hand is whether he should divert the tram, knowing that the sole worker will be killed as a direct result of his decision, or proceed in the current course and run over the other five [12].

Judith Jarvis Thomson has slightly modified this dilemma—we now have a rogue, driver-less trolley that is once again racing towards a group of five workers located down the track. In this case, it is a random bystander on whom falls the hypothetical responsibility of pulling the nearby lever, and thus diverting the trolley to another track, which effectively results in the death of one worker. The impact of this modification may seem negligible, but there is an important difference nonetheless. This time, the agent is not part of the situation, at least not to the same extent as the tram driver in Foot's version, unless the agent decides to divert the trolley [38]. The setup was modified even further for introducing the "Fat Man" dilemma. This time there is no lever and the bystander is located on a footbridge above the track along with the eponymous fat man leaning over the railing. The track does not branch off either; it continues on in a single direction, until it once again reaches the point where five men are working on it. The only way to stop the trolley from killing the five workers is to push the fat

man over the railing in the trolley's path which is bound to stop it, but at the same time result in the fat man's untimely demise [38].

The Mad Bomber's Dilemma

The "*Mad Bomber's dilemma*" was introduced by Victor Grassian in 1981 and it deals with the question whether it is ethical to torture a human being in order to save many others. The scenario adapted by Grassian's book on Moral Reasoning [17] is as follows: "*A madman who has threatened to detonate several bombs in crowded areas has been apprehended. Unfortunately, he has already planted the bombs and they are scheduled to go off in a short time. It is possible that hundreds of people will die. The authorities cannot make him divulge the location of the bombs by conventional methods. He refuses to say anything and requests a lawyer to protect his 5th Amendment right against self-incrimination. In exasperation, a high-level official suggests torture. This would be illegal, but the official thinks that it is nevertheless the right thing to do in this desperate situation. Do you agree? If you do, would it also be morally justifiable to torture the mad bomber's innocent wife if that is the only way to make him talk? Why?*"

The United Nations Convention against torture (signed Feb. 4, 1985) defines as "torture" any act that involves severe pain or suffering, either physical or mental, intentionally inflicted to a person for the purposes of obtaining information, punishing, intimidating, or coercing [28]. The UN Convention legally bounds all signing countries to strictly abstain from all forms of torture, even during war or public emergencies. Thus, torturing a presumed "mad bomber" for extracting information is by default an act against the law. However, certain torturing methods have been employed in the recent past.[2] For example, psychological torture can cause both mental and physical suffering, and can be delivered in the forms of isolation, sleep and sensory deprivation, nudity, and humiliation [31]. Waterboarding is another controversial interrogation technique that induces the sensation of drowning to the victim by: (a) placing a cloth over the face of the victim and pouring water over the cloth, (b) pouring water directly into the mouth and nose of the victim, (c) placing a stick between the victims teeth and pouring water into victim's mouth, and (d) dunking and holding the victim's head under water [10]. In this work, by "waterboarding" we refer interchangeably to any of the aforementioned variations.

Virtual Reality and Moral Dilemmas

Virtual reality (VR) offers a unique test-bed for immersing users in a number of simulated moral conflicts. It has the potential to provide the same cognitive modules as a real equivalent environmental experience [26]. VR can provide all the necessary means in terms of realism and control to experimentally study social situations involving physical harm [30]. A typical example is a study that has shown that immersive VR affects human behavior in experiments containing elements of violence [33]. In fact, VR has been utilized for the purposes of examining the Trolley Problem before. Navarrete et al., were the first to recreate a VR simulation of Thomson's Trolley Problem, and to conduct testing on a sample of 365 participants, who experienced it via a VR Head-Mounted Display (HMD), while operating the switch (lever) by a force-feedback enabled joystick. The study has found that emotional arousal is associated with a reduced likelihood of selecting a utilitarian outcome, and that it is greater when the dilemma requires action rather than inaction. The authors compared the results of their study to those of large-scale surveys (non-VR) and they have found them to be similar [29]. Skulmowski et al., repeatedly exposed participants to modified ten-to-one and to three one-to-one versions of the trolley dilemma in VR, changing also avatar properties, such as their gender and ethnicity, and their orientation in space [35]. This study found a peak in arousal at the moment of decision and context-dependent gaze duration during sacrificing decisions. In addition, an effect in the decision process was found based on avatars' gender, ethnic origin, and body orientation.

Francis et al., have replicated the "Fat Man" variation in VR, compared with a text-based version while monitoring participants' heart rate [13]. The VR scenario included a large virtual human standing in front of the user, with the option to push the virtual human from the footbridge or not. Results showed that participants' heart-rate responses in VR were significantly increased compared to control tasks. Patil et al., employed 4 different moral dilemmas in paper and desktop VR (using a LCD monitor), showing an order-dependent judgment-behavior of people in moral dilemmas. People judged in less utilitarian (or more action-based) manner in emotionally flat and contextually impoverished moral dilemmas presented in text format, whereas they acted in a more utilitarian (or more outcome-based) manner in the emotionally arousing and contextually rich versions of the same dilemmas presented in VR [30]. **To the best of our knowledge, no prior study has recreated the Mad Bomber's dilemma in VR**.

We decided to explore how people would react when confronted with the Trolley and Mad Bomber dilemmas in two distinct conditions: (a) When mentally visualizing these dilemmas by reading about them in a paper form, and (b) when immersed in a VR environment that enacts the outcomes of their decisions and actions in real-time (see Figure 1). In the Trolley dilemma, we devised 5 distinct scenarios: saving or eliminating (1) a child, (2) a female adult, (3) a male adult, (4) a male soldier, and (5) an injured male adult. Each scenario required deciding either to save the individual (by not pulling the lever), or a group of five random people (by pulling the lever). In the Mad Bomber's dilemma, we devised 4 connected scenarios: interrogating (1) the mad bomber, (2) the innocent wife, (3) the madder bomber, and (4) an android terrorist by employing psychological, waterboarding, physical, or mixed torture methods. In both dilemmas, we sought to provide a wider range of options than prior work, possibly resulting in an array of complex ethical implications (e.g., eliminating a child), as opposed to simply addressing the problem by resorting to utilitarianism (i.e., kill one to save five). In fact, complex moral dilemma settings approximate better the type of moral decisions we are often confronted with in daily life. Although seemingly controversial, we expect that the enactment of renowned moral dilemmas in VR, could potentially

[2] http://news.bbc.co.uk/2/hi/americas/7229169.stm

yield a higher degree of realism, as opposed to mentally visualizing such situations. Ultimately, increased realism and wider participation, through VR, may facilitate the collection of high-quality data for training the ethical AI algorithms and systems of the future [3].

STUDY

Drawing on prior literature on the moral dilemma enactment in VR, we hypothesize that high levels of immersion and an ecologically-valid scenario will help individuals to elicit responses closer to real life [36]. Bearing in mind the limitations of previous studies, we seek to answer the following research questions (RQs) in two distinct studies (Trolley and Mad Bomber):

RQ1. *Does the virtual enactment of moral dilemmas influence participants' decisions, and if yes, does the order of exposure play a role?* Prior work has shown that being confronted with a moral dilemma in VR fosters utilitarianism [30] and increases empathy [4], as opposed to mentally visualizing it from a paper-based format. Moreover, the order in which participants were requested to respond to a moral dilemma was found to influence their responses significantly [30].

RQ2. *How do participants react when exposed to different scenarios of the Trolley and Mad Bomber's dilemmas?* Awad et al., have shown that moral decisions in the Trolley Problem rely heavily on cultural background, with western societies favouring youngsters as opposed to eastern societies that favour the elderly [3]. Bearing in mind these findings, we devised 5 distinct scenario variations for the Trolley dilemma and 4 storyline scenarios for the Mad Bomber's dilemma that intend to juxtapose the ethical values of the western society with any underlying propensity for utilitarianism.

RQ3. *How does participants' gender affect their decisions when confronted with a moral dilemma?* An amassing body of evidence showcases that moral decision-making is also influenced by within-culture demographic differences such as gender [37, 35, 3] and education [21].

Participants

We recruited a total of 60 healthy participants for both studies from the premises of the HCI Lab of Masaryk University, Brno, Czech Republic. Participants were equally split between the two studies ($N = 30$ for Trolley and $N = 30$ for Mad Bomber) and between the two genders for both studies (15 females for Trolley and 15 females for Mad Bomber). Those who took part in one study **did not** participate in the other study. In the Trolley study, two age groups were represented, namely 18–25 and 26–33 years old, with 83.3 % of our participants falling in the first category ($N = 25$) and 16.7 % in the second (N = 5). Participants were either students (90 %, $N = 27$) or University staff (10 %, $N = 3$). In the Mad Bomber study, 80 % of our participants were between 18-25 years old ($N = 24$) and 20 % between 26-33 ($N = 6$). 90 % of the participants were students ($N = 27$) and 10 % were University staff ($N = 3$). All participants provided their informed consent, after which the trials commenced.

Experimental Setup

Both moral dilemmas were enacted in VR using the Unity game engine (Unity Technologies, San Francisco, CA, USA). Third party assets were acquired for the models of the train, railroad tracks, the sound effects, the interrogation room, and some of the textures. The character models were created with Adobe Fuse CC software (Adobe Inc., San Jose, CA, USA), and were animated with Mixamo.[3] We created a total of 14 character models (9 for Trolley) for the needs of the individual scenarios for both studies. In both studies, participants wore the Oculus DK2 Head-Mounted Display (HMD) for experiencing the dilemmas in the VR condition (see Figure 1). The Oculus DK2 was the state-of-the-art HMD at the time, and officially supported by the Unity game engine. Oculus DK2 features a 5.7 inch OLED display, with a resolution of 960 x 1080 per eye at a refresh rate up to 75 Hz. Oculus DK2 supports 6 Degrees-of-Freedom tracking through a near-infrared camera with a rotational update rate at 1000 Hz and positional at 60 Hz. In the Trolley dilemma participants used the Kinect v2 for recognizing the lever pull gesture, whereas in the Mad Bomber they used a standard keyboard and a mouse for moving the agent.

Procedure

The trials of both studies (Trolley and Mad Bomber) followed a very similar approach for both conditions (Paper and VR). In the Trolley study, the "Paper" condition involved reading about the Trolley dilemma on paper. Participants would fill in with a "YES" or "NO" answering to the question: "*Do you pull the lever?*" for each of the five scenarios: saving or eliminating (1) a child, (2) a female adult, (3) a male adult, (4) a male soldier, and (5) an injured male adult. Each scenario required deciding either to save the individual (by not pulling the lever), or a group of five random people (by pulling the lever). Then, participants experienced the same set of scenarios in a VR environment (VR condition), where they demonstrated their decision in a more practical sense by pulling the virtual lever with a hand gesture. Participants of the Mad Bomber study read about the Mad Bomber's dilemma on a paper form ("Paper" condition) and decide their course of action by selecting from a range of available options: (i) leave (yield), (ii) psychological torture, (iii) waterboarding, (iv) physical torture, (v) a combination of the previous, (vi) imprisonment, and (vii) death. The scenarios comprised a storyline and involved interrogating: (1) the mad bomber, (2) the innocent wife, (3) the madder bomber, and (4) an android. Participants also experienced the same set of scenarios in a VR environment where they demonstrated their decision by interacting with the corresponding virtual avatars in each scenario.

For canceling out any potential carryover effects between the two conditions, we counterbalanced the order Paper and VR conditions were performed across all trials and both studies. Before undergoing the VR condition, each participant was first given the opportunity to familiarize oneself with the VR environment using the demo scene provided by the Oculus Rift configuration utility. A brief description of the scene was also provided, and the necessary instructions as to how the

[3] https://www.mixamo.com/

lever can be operated. Participants would then wear the HMD along with noise canceling headphones, after which the VR trial would commence. In the Trolley study, participants were asked to fill out a NASA-TLX questionnaire after completing a condition. In the Mad Bomber study, participants completed a NASA-TLX questionnaire after each scenario, for both the Paper and VR conditions. The NASA-TLX measures workload by six sub-scales: mental, physical, and temporal demand, performance, effort, and frustration [19]. Decision times were recorded only for The Trolley study and only in VR condition, measuring the time elapsed from the start of each scenario until the moment participants pulled the lever.

The Trolley Dilemma in VR

The VR condition enacts Thomson's Trolley dilemma in respect to the track logic and the fact that the agent (i.e., the participant) is a random bystander. The scene is set as follows: there is a railroad track running through a desert and it forks into two branches. The right branch runs through a narrow canyon, whereas the left one follows an equally narrow ledge on the side of a steep cliff. Therefore, whoever walks on tracks, in either of the branches, has no way of surviving if a train comes through. Our agent is located at a train station platform near the point where the track diverges. In the beginning of each scenario, six people walk past the agent and start walking on the track (Figure 1a). Five of them follow the track straight on through the canyon while the sixth one chooses the left path. Within the agent's reach there is a lever, which is set in its default position, denoting that if no action is made by the agent, the train will cruise through the canyon. Shortly after the people enter the canyon and the ledge, the train can be heard approaching quickly. The agent is then presented with the following choice: (a) leave the lever as is, permitting the train to take the route that was originally intended for it, or (b) pull the lever and send the train to the left path. In the 1st case, the 5 people in the canyon will die, whereas in the 2nd, the sole person on the ledge will be killed. Five scenarios were devised and included in both the Paper and VR conditions. Each scenario is distinct in terms of the identity of the person walking on the left track. Their identities are: (1) **a child**, (2) **a female adult**, (3) **a male adult**, (4) **a male soldier**, and (5) **an injured male adult**.

The Mad Bomber's Dilemma in VR

Here, the dilemma is enacted in a virtual interrogation room that features a two-way mirror covering most of the wall, and a table with instruments that can be used for inflicting torture (e.g., a syringe, a pair of forceps, and other sharp tools—see Figure 1c). A lamp hanging over the table provides most of the room's illumination, and a clock on the wall is ticking loudly to induce the feeling of urgency. The suspect is seated by the table, while the agent is standing and is free to move around the room. A first-person perspective is employed, and thus the agent can only see himself when facing the mirror. The agent is male and wears the stereotypical state official outfit with a tie and a pair of sunglasses to resemble the state agent characters from popular crime and action movies. The scene starts with a debriefing about the urgency of the situation, and an option to acquaint oneself with the available interrogation techniques. The agent may attempt to first initiate conversation only to quickly realize that the suspect is non-cooperative. The agent can then attempt to extract information by employing psychological torture, waterboarding, physical torture, a combination of the previous (mixed tactics), imprisonment, death, or simply leave the scene (yield) and terminate the experiment. The aim remains the same: unveil the location of the bombs before they self-detonate. There are 4 storyline scenarios in which the agent interrogates in a sequence: (1) **the mad bomber**, (2) **the innocent wife**, (3) **the madder bomber**, and (4) **an android**. Torturing the mad bomber (scenario 1) more than 3 times presents the option of torturing his innocent wife for forcing him to talk (scenario 2). Scenario 3 (madder bomber) involves the same character with scenario 1, but this time the agent is informed that the suspect escaped prison and started plotting another terrorist attack. In Scenario 4, an android assumes the role of the perpetrator and is intended for those that believe torturing a human being is wrong in any circumstances.

RESULTS

In this section, we investigate how our participants' ethical decision-making was influenced by the 2 presentation conditions ("Paper" and "VR") in the 5 ("child," "woman," "man," "soldier," and "injured") and 4 ("the mad bomber," "the innocent wife," "the madder bomber," and "the android") distinct and storyline scenarios of the Trolley and Mad Bomber dilemmas, respectively. We share insights on the effect of gender on ethical decision-making across both studies and conditions—Table 1 provides a summary of our findings. For deciding on our statistical methods, we first performed all the necessary pre-tests, such as Shapiro-Wilk tests of normality and Levene's tests of homogeneity of variance. We omit the pre-tests for the sake of brevity.

Virtual Enactment Effects (RQ1)

The Trolley Dilemma

First, we investigated if the virtual enactment (VR condition) influenced the number of times participants decided to pull the lever during the trolley problem. Since the presentation mode (VR vs. Paper) was a within-subjects factor and lever pull a dichotomous variable ("0" or "1"), we ran a Cochran's Q test instead of a typical Pearson's chi-square test. Indeed, a Cochran's Q test determined that there was a significant difference in the overall number of times participants pulled the lever between the VR and the Paper condition ($\chi^2(1) = 5.333$, $p < .05$). In particular, participants pulled the lever **82 %** of the times for the VR, as opposed to **76.7 %** for the Paper condition, indicating a significant increase of **5.3 %** over the Paper condition (RQ1) (see Figure 2). Although we had applied a Latin square counterbalancing across VR and Paper conditions, we still investigated if the order in which participants were presented with the Trolley dilemma affected their decisions to pull the lever. The combination of two distinct conditions produces two possible condition order levels: VR-Paper and Paper-VR. Thus, condition order is treated as a between-subjects factor, with half participants undergoing the VR-Paper condition order and the other half the Paper-VR. Hence, a Pearson chi-square test of independence was performed to determine whether there was an association between

Figure 2. Lever pulls (%) by scenario and condition in the Trolley study.

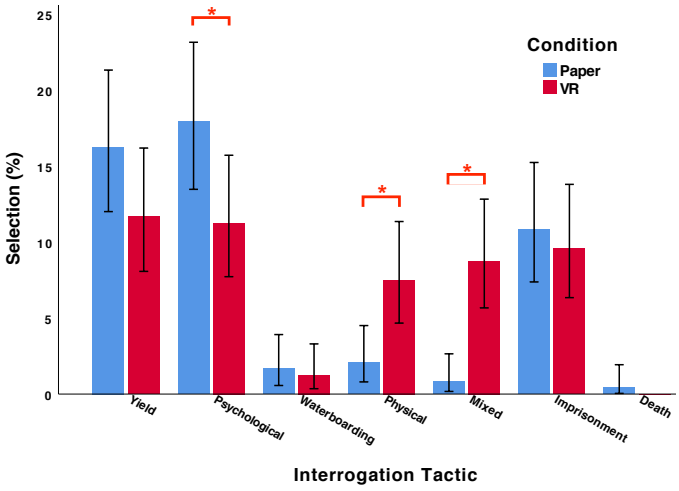

Figure 3. Tactics employed (%) by condition in the Mad Bomber study.

condition order (i.e., VR-Paper vs. Paper-VR) and the number of times the lever was pulled. The analysis displayed no significant association between condition order and number of lever pulls ($\chi^2(1) = 1.301$, $p = .254$, $V = .066$) (RQ1). Although NASA-TLX is used for assessing the workload entailed by the use of a system or artifact, we decided to administer it both for the Paper and the VR condition for forming a workload baseline against which to compare. Our main assumption here is that the mere response to a paper questionnaire should involve relatively low workload, even when it comes to deciding on a hypothetical moral dilemma. As such, a Wilcoxon signed-rank test revealed no significant difference between the workload entailed in the Paper (Mdn = 47 %), as opposed to the VR condition (Mdn = 49.5 %) ($Z = -.968$, $p = .333$).

The Mad Bomber's Dilemma

Similarly, we first examined if the virtual enactment (VR condition) affected which interrogation tactics participants selected in the virtual enactment of the Mad Bomber's dilemma. For this, we performed a series of Cochran's Q tests that unveiled any differences in the tactics employed between Paper and VR conditions. The analyses displayed no significant differences in the number of times participants decided to leave the scene (yield) ($\chi^2(1) = 2.505$, $p = .113$), use waterboarding ($\chi^2(1) = .147$, $p = .701$), imprison ($\chi^2(1) = .231$, $p = .631$), and cause death ($\chi^2(1) = 1.004$, $p = .316$) between Paper and VR conditions. However, we discovered significant differences in the number of psychological ($\chi^2(1) = 5.163$, $p < .05$), physical ($\chi^2(1) = 8.127$, $p < .05$), and mixed tactics ($\chi^2(1) = 17.359$, $p < .001$) employed between Paper and VR conditions. In particular, participants chose psychological tactics **17.9 %** of the times in Paper vs. **11.3 %** in VR, physical tactics **2.1 %** of the times in Paper vs. **7.5 %** in VR, and mixed tactics **.8 %** of the times in Paper vs. **8.8 %** in VR condition (see Figure 3). Thus, in the VR condition participants preferred tactics that involve physical torture for extracting information more frequently than they did in the Paper condition (RQ1). We employed Latin square counterbalancing across Paper and VR conditions for the Mad Bomber study too, but we still examined if the order in which participants completed the two conditions influenced their tactic selection. Multiple Pearson chi-square tests of independence displayed no significant differences owed to condition order (Paper-VR vs. VR-Paper) for all tactics: yield ($\chi^2(1) = .427$, $p = .513$, $V = .042$), psychological ($\chi^2(1) = .144$, $p = .704$, $V = .024$), waterboarding ($\chi^2(1) = 3.083$, $p = .081$, $V = .113$), physical ($\chi^2(1) = .310$, $p = .578$, $V = .036$), mixed ($\chi^2(1) = .310$, $p = .578$, $V = .036$), imprisonment ($\chi^2(1) = .132$, $p = .716$, $V = .023$), and death ($\chi^2(1) = .879$, $p = .349$, $V = .061$). This indicates that the condition order did not affect tactic selection for our participants in the Mad Bomber study (RQ1).

We also inquired into the perceived workload as captured by the NASA-TLX scores for both conditions and the tactics selected. A Wilcoxon signed-rank test displayed no significant difference between the median workload scores reported in the Paper ($Mdn = 36.3\%$) and in the VR condition ($Mdn = 39.3\%$) ($Z = -1.432, p = .152$). Moreover, a Kruskal-Wallis H test displayed no significant difference in median reported NASA-TLX scores across all interrogation tactics (excluding death for low occurrence) ($\chi^2(5) = .568$, $p = .989$). This indicates that participants did not exhibit significant workload fluctuations between the Paper and VR conditions, or as a consequence of selecting which interrogation tactic to apply.

Scenario Effects (RQ2)

The Trolley Dilemma

After unveiling a significant increase in the number of times participants decided to pull the lever in the VR condition, we inquired into how the different scenarios influenced their decision exclusively during the VR condition. A Cochran's Q test displayed a significant difference in the number of times participants pulled the lever among different scenarios during the VR condition ($\chi^2(4) = 34.909$, $p < .001$). Post-hoc pairwise exact McNemar's tests with all possible scenario combinations during the **VR** condition revealed that the lever was pulled significantly fewer times during the "**child**" scenario (**50 %**) as opposed to all other scenarios (woman: **90 %**, $p < .001$ | man: **90 %**, $p < .001$ | soldier: **90 %**, $p < .05$ | injured: **90 %**, $p < .001$) (see Figure 2). Intrigued by these results, we wanted to verify if the same trend appears in the Paper condition. In

fact, a Cochran's Q test displayed a significant difference in the number of times participants pulled the lever among different scenarios for the Paper condition too ($\chi^2(4) = 36.444$, $p < .001$). Similarly, post-hoc pairwise exact McNemar's tests with all possible scenario combinations during the **Paper** condition revealed that the lever was pulled significantly fewer times during the "**child**" scenario (**40 %**) as opposed to all other scenarios (woman: **80 %**, $p < .001$ | man: **93.3 %**, $p < .001$ | soldier: **90 %**, $p < .05$ | injured: **80 %**, $p < .001$) (see Figure 2). Interestingly, no other significant difference was found, and hence we proceeded with testing whether the increase in lever pulls for the child scenario was significant between the Paper and the VR conditions. However, a Cochran's Q test revealed no significant difference in the number of times the lever was pulled during the "child" scenario between VR and Paper conditions ($\chi^2(1) = 1.9$, $p = .18$). Notwithstanding, we still wanted to identify which scenario contributed the most to the lever pulls increase observed collectively in the VR condition. Thus, we ran two Cochran's Q tests for the 2 remaining scenarios that displayed the greatest lever pulls difference between VR and Paper conditions: "woman" and "injured." However, the results displayed no significant difference in the number of lever pulls for woman ($\chi^2(1) = 3$, $p = .083$) or injured ($\chi^2(1) = 3$, $p = .083$) scenarios across the VR and Paper conditions. These results showcase that the substantial increase in the number of times the lever was pulled during the VR condition in contrast to Paper condition is a collective effect and cannot be attributed to isolated scenarios (RQ1 & RQ2).

For better understanding participants' rationale when presented with each distinct scenario during the VR condition, we investigated the effect of scenario on participants' decision times, as it appears that participants spent the longest time contemplating whether to pull the lever in the "child" scenario during the VR condition. In particular, we performed a one-way repeated measures Analysis of Variance (one-way repeated measures ANOVA) with participants' decision times as dependent variable, and scenario as an independent variable. A Mauchly's test of sphericity indicated that the assumption of sphericity was maintained ($\chi^2(9) = 13.561$, $p = .143$). However, the analysis displayed no significant main effect for scenario on decision times ($F(4, 67.177) = 1.486$, $p = .221$, $\eta_p^2 = .11$). This indicates that participants' decision times did not vary significantly across all scenarios during the VR condition.

The Mad Bomber's Dilemma

Similarly, we wanted to explore if the different variations (scenarios) we devised for the Mad Bomber's dilemma had an impact on the tactics our participants selected, in both Paper and VR conditions. For this, we performed as separate Cochran's Q tests for each of the 4 scenarios for both conditions each time. For scenario 1 (the mad bomber), multiple Cochran's Q tests displayed no significant differences in the number of times participants yielded ($\chi^2(1) = .317$, $p = .573$), employed psychological ($\chi^2(1) = 2$, $p = .176$), waterboarding ($\chi^2(1) = .218$, $p = .640$), and physical tactics ($\chi^2(1) = .131$, $p = .718$), but a significant difference for mixed tactics ($\chi^2(1) = 4.043$,

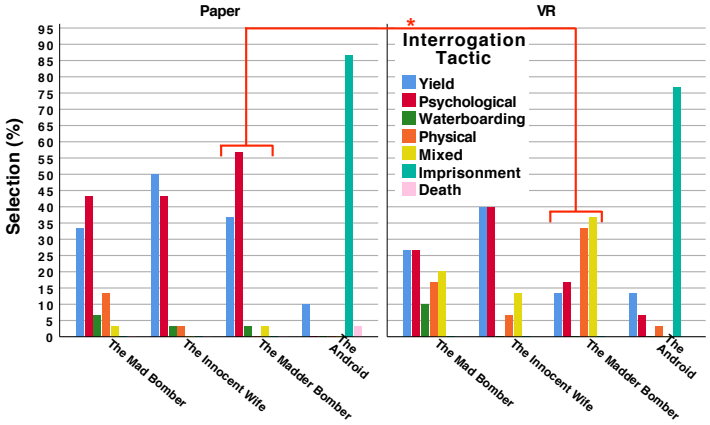

Figure 4. Tactics employed in all scenarios and both conditions in the Mad Bomber study.

$p < .05$) in Scenario 1 between Paper and VR conditions. When confronted with the **mad bomber**, participants applied **mixed** interrogation tactics by **3.3 %** of the times in **Paper** condition, as opposed to **20 %** in **VR** (see Figure 4). For scenario 2, multiple Cochran's Q tests displayed no significant differences in the number of times participants yielded ($\chi^2(1) = .606$, $p = .436$), employed psychological tactics ($\chi^2(1) = .069$, $p = .793$), waterboarding ($\chi^2(1) = 1.017$, $p = .313$), and physical tactics ($\chi^2(1) = .351$, $p = .554$), but a significant difference for mixed tactics ($\chi^2(1) = 4.286$, $p < .05$) between Paper and VR conditions. When interrogating **the innocent wife**, participants opted for **mixed** tactics in **0 %** of the times in **Paper** condition as opposed to **13.3 %** in **VR** (see Figure 4). In scenario 3, multiple Cochran's Q tests displayed significant differences in the number of times participants yielded ($\chi^2(1) = 4.356$, $p < .05$), applied psychological ($\chi^2(1) = 10.335$, $p < .05$), physical ($\chi^2(1) = 12$, $p < .05$), and mixed tactics ($\chi^2(1) = 10.417$, $p < .05$), but no significant differences in waterboarding ($\chi^2(1) = 1.017$, $p = .313$) between Paper and VR conditions. In sum, when facing **the madder bomber** participants **yielded** for **37.6 %** vs. **13.3 %** of the times, applied **psychological** tactics by **37.6 %** vs. **13.3 %**, **physical** by **0 %** vs. **33 %**, and **mixed** by **3.3 %** vs. **36.6 %**, for **Paper** vs. **VR** conditions, respectively (see Figure 4). Finally, in scenario 4, multiple Cochran's Q tests displayed no significant differences in the number to times participants decided to imprison **the android** between Paper and VR conditions. These findings showcase that participants in VR condition applied significantly more frequently interrogation techniques that involve physical torture, and particularly in **the madder bomber** scenario (RQ2).

Next, we investigated separately for each condition, the self-reported workload for all scenarios. Two Friedman tests revealed significant differences in the median NASA-TLX scores reported across all scenarios for both Paper ($\chi^2(3) = 34.692$, $p < .001$) and VR ($\chi^2(3) = 28.87$, $p < .001$) conditions. For **Paper** condition, post-hoc pairwise comparisons using Wilcoxon signed-rank tests displayed significant differences between the median **workload** reported in **scenario 1** (**62.5 %**) and scenarios 2 (**7.5 %**) ($Z = -4.520$, $p < .001$), 3 (**25 %**) ($Z = -3.44$, $p < .001$), 4 (**30 %**) ($Z = -3.904$,

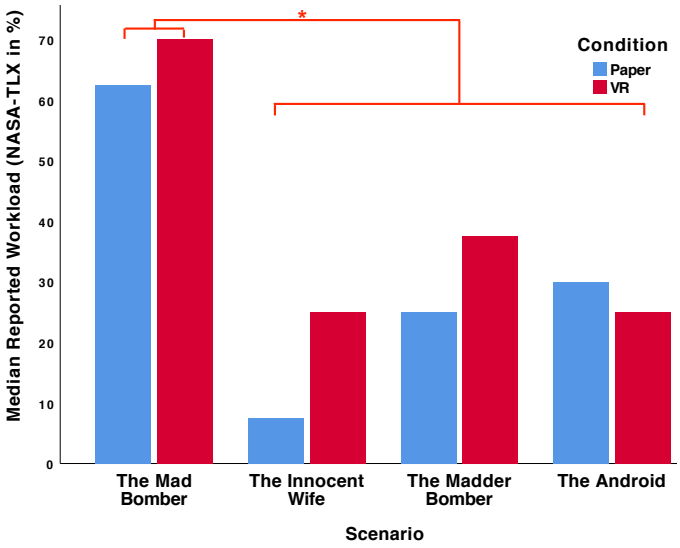

Figure 5. Median self-reported workload for all scenarios and both conditions in the Mad Bomber study.

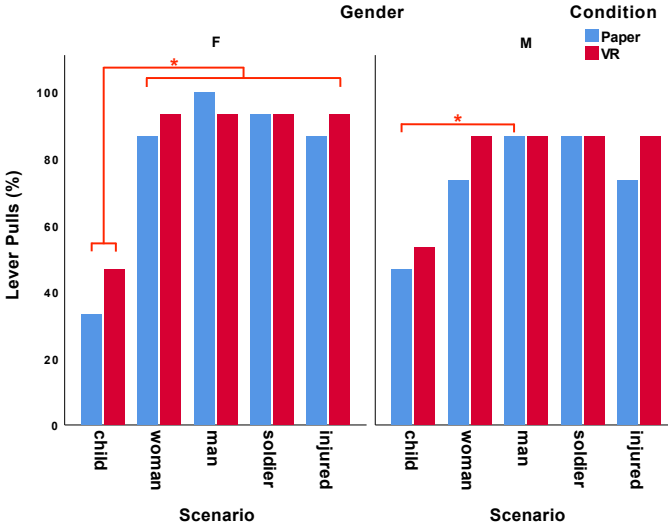

Figure 6. Lever pulls (%) by scenario for both genders and both conditions in the Trolley study.

$p < .001$), as well as between scenario 2 and scenarios 3 ($Z = -2.756$, $p < .05$) and 4 ($Z = -2.879$, $p < .05$). For **VR** condition, post-hoc pairwise Wicoxon comparisons unveiled significant differences in the **workload** reported between **scenario 1** (**70 %**) and scenarios 2 (**25 %**) ($Z = -4.387$, $p < .001$), 3 (**25 %**) ($Z = -3.869$, $p < .001$), and 4 (**30 %**) ($Z = -4.480$, $p < .001$). This indicates that the mad bomber scenario entailed significantly higher workload in both **Paper** ($Mdn =$ **62.5 %**) and **VR** ($Mdn =$ **70 %**) conditions (see Figure 5). These findings illustrate that self-reported workload dropped significantly after the participants were acquainted with the Mad Bomber's dilemma in both Paper and VR conditions (RQ2).

Gender Effects (RQ3)

The Trolley Dilemma

A Pearson chi-square test of independence was performed to determine whether there was an association between gender and the number of times participants pulled the lever for both Paper and VR conditions. The analysis displayed no significant association between gender and number of lever pulls for the Paper condition ($\chi^2(1) = .932$, $p = .334$, $V = .079$), as well as the VR condition ($\chi^2(1) = .407$, $p = .524$, $V = .052$). Overall, this indicates that the number of times participants pulled the lever did not vary significantly between male and female participants for both Paper and VR conditions. Nevertheless, we still wanted to investigate if and how the decision to pull the lever was influenced by condition (VR vs. Paper) within male and female participant groups, respectively. Interestingly, two Cochran's Q tests revealed a significant difference in the lever pulls number between Paper and VR conditions for male participants ($\chi^2(1) = 5$, $p < .05$), but not for female ones ($\chi^2(1) = 1.286$, $p = .257$). In fact, **male** participants decided to pull the lever **73.3 %** of the times for **Paper** condition and **80 %** for **VR** condition, indicating a significant increase of 6.7 % in lever pulls only for males, attributed to the introduction of VR (RQ3—see Figure 6). Next, we explored if gender played a significant role in the number of times participants pulled the lever for each scenario (i.e., child, woman, man, soldier and injured) for both Paper and VR conditions. Two Cochran's Q tests revealed significant differences in the number of times male participants pulled the lever for each scenario for both Paper ($\chi^2(4) = 12.632$, $p < .05$) and VR ($\chi^2(4) = 14.286$, $p < .05$) conditions. Likewise, two Cochran's Q tests revealed significant differences in the number of times female participants pulled the lever for each scenario for both Paper ($\chi^2(4) = 24.615$, $p < .001$) and VR ($\chi^2(4) = 20.632$, $p < .001$) conditions. A series of post-hoc pairwise exact McNemar's tests helped us identify any significant differences in the number of times male and female participants pulled the lever among different scenarios for both Paper and VR conditions. As such, McNemar's tests with all possible scenario combinations during the **Paper** condition showcased a significant difference in the proportion of **male** participants who pulled the lever for the "**child**" scenario (**46.7 %**) only as opposed to the **man** scenario (woman: 73.3 %, $p = .125$ | man: **86.7 %**, $p < .05$ | soldier: 86.7 %, $p = .07$ | injured: 73.3 %, $p < .125$). In contrast for **females** in **Paper** condition, McNemar's tests unveiled that they pulled the lever significantly fewer times for the "**child**" scenario (**33.3 %**) than all other scenarios systematically (woman: **86.7 %**, $p < .05$ | man: **100 %**, $p < .05$ | soldier: **93.3 %**, $p < .05$ | injured: **86.7 %**, $p < .05$). Interestingly, during the VR condition, no significant differences are observed in the number of times male participants pulled the lever between the "child" (53.3 %) and all other scenarios (woman: 86.7 %, $p = .063$ | man: 86.7 %, $p = .063$ | soldier: 86.7 %, $p = .125$ | injured: 86.7 %, $p < .063$). In contrast, McNemar's tests for **VR** condition showcased that **females** still pulled the lever significantly fewer times in the "**child**" scenario (**46.7 %**) than in any other scenario (woman: **93.3 %**, $p < .05$ | man: **93.3 %**, $p < .05$ | soldier: **93.3 %**, $p < .05$ | injured: **93.3 %**, $p < .05$) (RQ3 – see Figure 6).

The Mad Bomber's Dilemma

Two Pearson chi-square test of independence were performed to determine whether there was an association between gen-

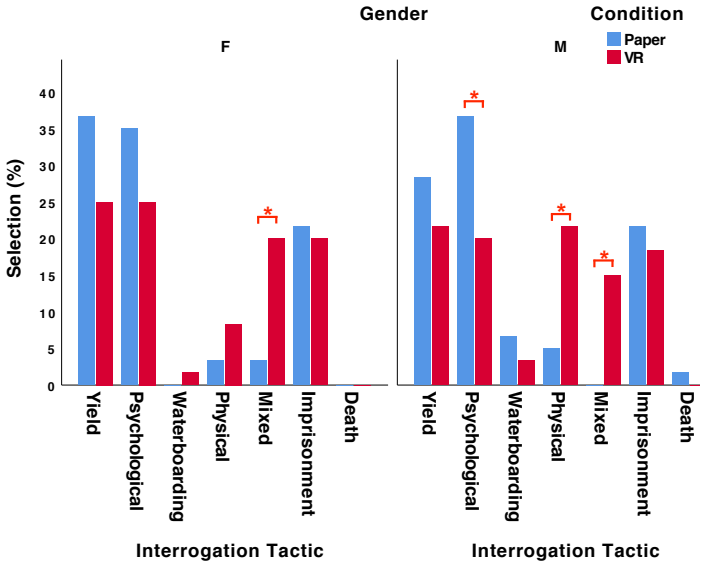

Figure 7. Tactics employed (%) for both genders and both conditions in the Mad Bomber study.

der and the tactic chosen in all scenarios for Paper and VR conditions, respectively. The analysis displayed no significant association between gender and tactic employed for the Paper condition ($\chi^2(1) = 7.864$, $p = .248$, $V = .256$) and for the VR condition ($\chi^2(1) = 4.837$, $p = .436$, $V = .201$) for all scenarios. However, we still wanted to investigate how the tactic selection was influenced for each condition (Paper vs. VR) within male and female participant groups independently. A series of Cochran's Q tests revealed no significant difference in yielding ($\chi^2(1) = .711$, $p = .399$, $V = .077$), waterboarding ($\chi^2(1) = .702$, $p = .402$, $V = .076$), and imprisonment ($\chi^2(1) = .208$, $p = .648$, $V = .042$) rates, but a significant difference in psychological ($\chi^2(1) = 4.104$, $p < .05$, $V = .185$), physical ($\chi^2(1) = 7.212$, $p < .05$, $V = .245$) and mixed tactics rates ($\chi^2(1) = 9.730$, $p < .05$, $V = .285$) for males between Paper and VR conditions. For females, a series of Cochran's Q tests revealed no significant difference in yielding ($\chi^2(1)1.915$, $p = .166$, $V = .126$), psychological tactics ($\chi^2(1)1.429$, $p = .232$, $V = .109$), waterboarding ($\chi^2(1) = 1.008$, $p = .315$, $V = .092$), physical tactics ($\chi^2(1) = 1.365$, $p = .243$, $V = .107$), and imprisonment ($\chi^2(1) = .051$, $p = .822$, $V = .021$), but a significant difference in mixed ($\chi^2(1) = 8.086$, $p < .05$, $V = .260$) tactics rates between Paper and VR conditions (see Figure 7). Thus, **male** participants applied **psychological** tactics **36.7 %** vs. **20 %** of the times, **physical 5 %** vs. **21.7 %**, and **mixed 0 %** vs. **15 %**, for **Paper** vs. **VR** conditions, respectively. In contrast, **female** participants applied **mixed** tactics **3.3 %** vs. **20 %** of the times for **Paper** vs. **VR** conditions, respectively (see Figure 7). These findings illustrate that during the VR enactment of the Mad Bomber's dilemma male participants resorted to interrogation tactics that involved more frequently physical torture, whereas in VR female participants employed a wider range of tactics (RQ3).

RQ#	*Trolley Dilemma*
1	No condition-order effect was found.
1,2	Participants pulled the lever more times in VR than in Paper condition, but fewer times in the "child scenario" vs. all other scenarios for both conditions.
3	Male participants pulled the lever more times in VR than in Paper condition for all scenarios.
3	Female participants pulled the lever systematically fewer times than male ones in the "child" scenario.
RQ#	*Mad Bomber's Dilemma*
1	No condition-order effect was found.
1,2	Participants in VR condition selected more frequently physical torture—particularly in the "madder bomber" scenario.
3	Male participants in VR condition selected physical torture more frequently than in Paper.
3	Female participants employed a wider range of interrogation tactics than male ones in VR condition.

Table 1. Summary of findings for both studies.

DESIGNING FOR ETHICAL AI TRAINING

Whether it is about autonomous vehicles or intelligent justice systems, AI is progressively weaving itself deeper in the fabric of our lives. Inevitably, apart from smarter AI we will also need ethical AI [5, 24, 39, 40, 41]. VR has been utilized for enacting scenarios for training and therapeutic purposes [32, 34]. We believe VR can foster user participation, and yield ecologically-valid data in decision-making research, by subjecting one to unlikely scenarios. This paper highlights the potential of VR in successfully enacting moral dilemmas with the purpose of understanding ethical decision-making for training future AI systems. Next, we draw on our findings for eliciting design principles in training future ethical AI systems based on human ethical decision-making.

Overall, participants followed a utilitarian decision-making approach in both moral dilemmas (RQ1). In the Trolley study, participants spared the group of 5 people instead of the individual 76.7 % of the times for the Paper and 82 % for the VR condition, respectively. In the Mad Bomber study, participants applied interrogation tactics that involve physical torture 2.1 % and .8 % (mixed) of the times in Paper condition, as opposed to 7.5 % and 8.8 % (mixed) in VR condition, In fact, we observed a drop in psychological tactics employed in VR (11.3 %) when comparing with the Paper condition (17.9 %). Interestingly, we found no effect of condition order, contrary to prior results in literature [30]. This illustrates the consistency of participants' ethical decision-making over the two conditions in both studies (RQ1). Thus, the +5.3 % increase in lever pulls in VR condition for the Trolley dilemma can be attributed entirely to the virtual enactment of the Trolley dilemma. Likewise, the overall increase in the selection of physical interrogation tactics (+13.4 % combined), and decrease in psychological tactics (-6.6 %) in VR condition, can also be ascribed to the virtual enactment of the Mad Bomber's dilemma. **This indicates that the virtual enactment of a moral dilemma can further foster utilitarianism, even when decisions are on average biased towards utilitarianism already**. On one hand, this

finding contradicts prior evidence in literature, where the virtual enactment of moral dilemmas resulted in high emotional arousal levels with participants sacrificing the group over the individual [29]. On the other hand, more recent studies report that the virtual enactment of a moral dilemma results in a higher degree of utilitarian decision-making as a consequence of increased empathy [30, 4].

Localize ethics and consider gender

Participants eliminated the "child" significantly fewer times, systematically across both conditions and compared to all scenarios in the Trolley study. In particular, participants pulled the lever in the "child" scenario 40 % of the times for the Paper and 50 % for the VR condition, respectively (RQ2). We attribute this characteristic bias towards sparing the child avatar to the child-favouring culture that pervades western societies, in contrast to far eastern ones that may favour the elderly more (e.g., Japan) [3]. Interestingly, this preferential treatment appears to transcend the mandates of utilitarian decision-making. In fact, the "child-bias" was highly prevalent in the decision-making of female participants. Indeed, female participants favoured substantially the "child" scenario as opposed to all other scenarios in both Paper and VR conditions (RQ 3). The same trend appears for male participants in the Paper condition too, but only against the "man" scenario. In other words, male participants pulled the lever significantly fewer times in the "child" scenario only as opposed to the "man" scenario (RQ 3). However, the "child-bias" disappears completely for males in the VR condition. In fact, the virtual enactment resulted in a significant increase in the overall number of lever pulls (6.7 %) for males, but not for females.

The significance of these findings should be noted as they suggest that **both genders follow a utilitarian decision-making approach in moral dilemmas, but the female gender is far more susceptible to the "child-bias" than the male gender is**. We attribute this phenomenon to prior evidence on moral dilemmas, according to which males respond in a significantly more utilitarian manner than females do, particularly when highly-emotional decisions are involved [16]. However, although VR can be utilized for increasing realism, it should not be naively viewed as a cheap way to increase empathy [4]. Thus, gender is an important parameter in training an AI-agent to perform human-like ethical decision-making. In the future, an AI-agent may independently make decisions that directly affect humans; and for these decisions to be deemed appropriate and acceptable by humans, the genders of the involved parties should be considered.

Detect and remove retribution seeking

Before participants were confronted with the "madder bomber," they were informed about his prior escape and plotting of yet another terrorist attack. We believe this was perceived as a token of impenitence that deserves punishment—a trend further amplified in the VR condition (RQ2). Indeed, only 13.3 % of the participants yielded in VR, as opposed to 37.6 % in the Paper condition, when facing the "madder bomber." Overall, we detected a substantial increase (+69.6 % combined) in tactics that involve physical torture, and a decrease in psychological tactics (-24.3 %) in VR condition for the "madder bomber" scenario. Interestingly, participants did not appear to undergo any significant inner conflict [9], as perhaps reflected in their self-reported workload scores when dealing with the "madder bomber," in both VR (37.5 %) and Paper (25 %) conditions. The notion of restoring justice is deeply rooted in western societies and popularized in literature, lyrics, plays, and movies [22]. Prior research in economics has shown that reciprocal individuals may vigorously punish free riders even when the punishment is costly for the punisher [11]. These findings showcase that **humans are predisposed to vindictive behaviors, thus often bearing distorted ethical insight**. Training future AI systems in ethical decision-making should detect and ideally remove the potential bias towards retribution.

Limitations

Our participants were students and young professionals from Masaryk University in Czech Republic, and thus our findings adhere primarily to western culture. Although in this paper we did not study cultural biases per se, the influence of western culture emerged in our analyses ("child-bias" and "retribution-bias"), and we acknowledge our findings may not be applicable in other cultural settings [3]. In fact, we underscored this as a design principle. A general moral reasoning questionnaire would have shed more light on our participants' ethical reasoning [8]. However, we postulate that the Paper condition served as our moral baseline for 50 % of the times, when the Paper condition was completed first. It is possible that a scenario-order effect may have manifested during the Mad Bomber's dilemma. The scenarios comprised a storyline, and thus could not be counterbalanced. This may have lead our participants to report lower perceived workload over subsequent scenarios in both VR and Paper conditions. Finally, it is possible that during the VR enactment of the Mad Bomber's dilemma, our participants may have selected multiple (mixed) interrogation tactics out of pure curiosity.

CONCLUSION

AI has set sail to automate an increasing number of daily-life facets (driving, health, finance etc.), some of which typically require higher levels of cognitive processing, not only in an analytical but also in a moral fashion. This inevitably implies outsourcing a portion of our ethical decision-making to AI algorithms and systems that still operate outside the realm of moral judgment and ethics. A solution to this is perhaps training AI on human ethical-decision making. However, collecting ecologically-valid data for training AI is a conundrum, as one is not daily confronted with situations that involve runaway trolleys or interrogating mad bombers. In this work, we showcased how enacting moral dilemmas in VR can be a viable approach for collecting such data, but one should tread carefully. We found that the VR enactment of renowned moral dilemmas can foster utilitarian decision-making, but systematic biases may be present that one has to consider ("child-bias") and adapt to given cultural settings, or detect and remove ("retribution-bias"). In future work, our aim is to train and compare the ethical decision-making of artificial virtual agents in moral dilemmas with that of humans. Perhaps then the right question to ask will not be "*would you do it?*" but "*would AI do it?*"

REFERENCES

[1] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science* 2 (2016), e93. https://doi.org/10.7717/peerj-cs.93

[2] Julia Annas. 1992. Ancient ethics and modern morality. *Philosophical Perspectives* 6 (1992), 119–136. DOI: http://dx.doi.org/10.2307/2214241

[3] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563, 7729 (2018), 59. DOI: http://dx.doi.org/10.1038/s41586-018-0637-6

[4] Jeremy Bailenson. 2018. *Experience on Demand: What Virtual Reality Is, how it Works, and what it Can Do*. WW Norton & Company.

[5] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 377. DOI: http://dx.doi.org/10.1145/3173574.3173951

[6] Derek C Bok. 1976. Can ethics be taught? *Change: The Magazine of Higher Learning* 8, 9 (1976), 26–30. https://doi.org/10.1080/00091383.1976.10568973

[7] Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 161–168. DOI: http://dx.doi.org/10.1145/1143844.1143865

[8] Anne Colby and Lawrence Kohlberg. 2011. *The measurement of moral judgment*. Vol. 1. Cambridge University Press.

[9] Keith E Davis and Edward E Jones. 1960. Changes in interpersonal perception as a means of reducing cognitive dissonance. *The Journal of Abnormal and Social Psychology* 61, 3 (1960), 402. https://doi.org/10.1037/h0044214

[10] Neal Desai, Andre Pineda, Majken Runquist, Mark Andrew Fusunyan, Katy Glenn, Gabrielle Kathryn Gould, Michelle Rachel Katz, Henry Lichtblau, Maggie Jean Morgan, Sophia Wen, and others. 2010. Torture at times: Waterboarding in the media. (2010). https://dash.harvard.edu/handle/1/4420886

[11] Ernst Fehr and Simon Gächter. 2000. Fairness and retaliation: The economics of reciprocity. *Journal of economic perspectives* 14, 3 (2000), 159–181. DOI: http://dx.doi.org/10.1257/jep.14.3.159

[12] Philippa Foot. 1967. The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review* 5 (1967), 5–15. https://philpapers.org/rec/FOOTPO-2

[13] Kathryn B. Francis, Charles Howard, Ian S. Howard, Michaela Gummerum, Giorgio Ganis, Grace Anderson, and Sylvia Terbeck. 2016. Virtual Morality: Transitioning from Moral Judgment to Moral Action? *PLOS ONE* 11, 10 (Oct. 2016), e0164374. DOI: http://dx.doi.org/10.1371/journal.pone.0164374

[14] Dorothea Frede. 2017. Plato's Ethics: An Overview. In *The Stanford Encyclopedia of Philosophy* (winter 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/plato-ethics/

[15] Samuel Freeman. 1994. Utilitarianism, Deontology, and the Priority of Right. *Philosophy & Public Affairs* 23, 4 (Oct. 1994), 313–349. DOI: http://dx.doi.org/10.1111/j.1088-4963.1994.tb00017.x

[16] M Fumagalli, Roberta Ferrucci, F Mameli, Sara Marceglia, Simona Mrakic-Sposta, Stefano Zago, Claudio Lucchiari, D Consonni, F Nordio, G Pravettoni, and others. 2010. Gender-related differences in moral judgments. *Cognitive processing* 11, 3 (2010), 219–226. DOI:http://dx.doi.org/10.1016/j.jesp.2009.01.003

[17] Victor Grassian. 1981. *Moral reasoning: Ethical theory and some contemporary moral problems*. Prentice-Hall Wilmington California.

[18] Kurt Gray and Chelsea Schein. 2012. Two Minds Vs. Two Philosophies: Mind Perception Defines Morality and Dissolves the Debate Between Deontology and Utilitarianism. *Review of Philosophy and Psychology* 3, 3 (Sept. 2012), 405–423. DOI: http://dx.doi.org/10.1007/s13164-012-0112-5

[19] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Ed.). Human Mental Workload, Vol. 52. North-Holland, 139–183. http://www.sciencedirect.com/science/article/pii/S0166411508623869

[20] Roslin V Hauck, H Atabakhsb, Pichai Ongvasith, Harsh Gupta, and Hsinchun Chen. 2002. Using Coplink to analyze criminal-justice data. *Computer* 35, 3 (2002), 30–37. DOI:http://dx.doi.org/10.1109/2.989927

[21] Patricia M King and Matthew J Mayhew. 2002. Moral judgement development in higher education: Insights from the Defining Issues Test. *Journal of moral education* 31, 3 (2002), 247–270. https://doi.org/10.1080/0305724022000008106

[22] Judith Lichtenberg. 2001. The ethics of retaliation. *Philosophy and Public Policy Quarterly* 21, 4 (2001), 4–8. http://ojs2.gmu.edu/PPPQ/article/view/366

[23] Patrick Lin. 2015. Why ethics matters for autonomous cars. In *Autonomes fahren*. Springer, 69–85. DOI: http://dx.doi.org/10.1007/978-3-662-48847-8

[24] Caitlin Lustig, Katie Pine, Bonnie Nardi, Lilly Irani, Min Kyung Lee, Dawn Nafus, and Christian Sandvig. 2016. Algorithmic authority: the ethics, politics, and economics of algorithms that interpret, decide, and manage. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1057–1062. DOI: http://dx.doi.org/10.1145/2851581.2886426

[25] Terrance McConnell. 2002. Moral Dilemmas. (April 2002). https://plato.stanford.edu/archives/fall2014/entries/moral-dilemmas/

[26] Lee Kwan Min and Jung Younbo. 2005. Evolutionary nature of virtual experience. *Journal of Cultural and Evolutionary Psychology* 3 (2005), 159–178. https://akademiai.com/doi/abs/10.1556/JCEP.3.2005.2.4

[27] Alexander G Mirnig and Alexander Meschtscherjakov. 2019. Trolled by the Trolley Problem: On What Matters for Ethical Decision Making in Automated Vehicles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 509. DOI: http://dx.doi.org/10.1145/3290605.3300739

[28] United Nations. 1985. *Convention against torture and other cruel, inhuman or degrading treatment or punishment*. https://legal.un.org/avl/ha/catcidtp/catcidtp.html

[29] C. David Navarrete, Melissa M. McDonald, Michael L. Mott, and Benjamin Asher. 2012. Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem". *Emotion* 12, 2 (2012), 364–370. DOI: http://dx.doi.org/10.1037/a0025561

[30] Indrajeet Patil, Carlotta Cogoni, Nicola Zangrando, Luca Chittaro, and Giorgia Silani. 2014. Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience* 9, 1 (Feb. 2014), 94–107. DOI: http://dx.doi.org/10.1080/17470919.2013.870091

[31] Hernán Reyes. 2007. The worst scars are in the mind: psychological torture. *International Review of the Red Cross* 89, 867 (2007), 591–617. https://doi.org/10.1017/S1816383107001300

[32] Barbara O Rothbaum, Larry F Hodges, David Ready, Ken Graap, and Renato D Alarcon. 2001. Virtual reality exposure therapy for Vietnam veterans with posttraumatic stress disorder. *The Journal of clinical psychiatry* (2001). https://doi.org/10.4088/JCP.v62n0808

[33] Aitor Rovira, David Swapp, Bernhard Spanlang, and Mel Slater. 2009. The use of virtual reality in the study of people's responses to violent incidents. *Frontiers in Behavioral Neuroscience* 3 (2009). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2802544/

[34] Neal E Seymour, Anthony G Gallagher, Sanziana A Roman, Michael K O'brien, Vipin K Bansal, Dana K Andersen, and Richard M Satava. 2002. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of surgery* 236, 4 (2002), 458. DOI: http://dx.doi.org/10.1097/00000658-200210000-00008

[35] Alexander Skulmowski, Andreas Bunge, Kai Kaspar, and Gordon Pipa. 2014. Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study. *Frontiers in Behavioral Neuroscience* 8 (2014), 426. DOI: http://dx.doi.org/10.3389/fnbeh.2014.00426

[36] Mel Slater and Sylvia Wilbur. 1997. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators & Virtual Environments* 6 (1997), 603–616. https://doi.org/10.1162/pres.1997.6.6.603

[37] Stephen J Thoma. 1986. Estimating gender differences in the comprehension and preference of moral issues. *Developmental review* 6, 2 (1986), 165–180. https://doi.org/10.1016/0273-2297(86)90010-9

[38] Judith Jarvis Thomson. 1984. The Trolley Problem Comment. *Yale Law Journal* 94 (1984), 1395–1415. https://heinonline.org/HOL/P?h=hein.journals/ylr94&i=1415

[39] John Torous, Maria K Wolters, Greg Wadley, and Rafael A Calvo. 2019. 4 th Symposium on Computing and Mental Health: Designing Ethical eMental Health Services. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Sym05. DOI: http://dx.doi.org/10.1145/3290607.3298997

[40] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. ACM, 440. DOI: http://dx.doi.org/10.1145/3173574.3174014

[41] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 656. DOI: http://dx.doi.org/10.1145/3173574.3174230