Moral thinking across the world: Exploring the influence of personal force and intention in moral dilemma judgements

Bence Bago[1], Balazs Aczel[2], Zoltan Kekecs[2], John Protzko[3], Marton Kovacs[2], Tamas Nagy[2], Rink Hoekstra[4], Bence Palfi[5,6], Alain Quiamzade[7], Alan D. A. Mattiassi[8], Alasdair R. Gordon-Finlayson[9], Alberto Mirisola[10], Alexandra Fleischmann[11], Alexios Arvanitis[12], Anabel Belaus[13], Andree Hartanto[14], Andrej Findor[15], Andrew G. Thomas[16], Anna Wlodarczyk[17], Anni Sternisko[18], Anthony Lantian[19], Anton Kühberger[20], Aurélien Graton[21], Bana Bashour[22], Barnaby J. W. Dixson[23], Bastiaan T. Rutjens[24], Bastian Jaeger[25], Bastien Trémoliere[26], Benjamin Becker[27], Benjamin T. Brown[28], Biljana Gjoneska[29], Biljana Jokić[30], Bill Jiaxin Shi[31], Carlota Batres[32], Carmel A Levitan[33], Cecilia Reyna[13], Celine-Justine Matibag[34], Charlotte A. Hudson[35], Christina Shane-Simpson[36], Christoph Schild[37], Christopher Castille[38], Cubillas C. Perez[39], Daniel Storage[40], Danielle P. Ochoa[41], Danka Purić[30], Darius-Aurel Frank[42], David Moreau[43], Diego A. Reinero[18], Don C. Zhang[44], Dustin P. Calvillo[45], Elif Gizem Demirag Burak[46], Elisa Puvia[47], Elizabeth Harris[18], Elke B. Lange[48], Ennio Bilancinio[49], Erica D. Musser[50], Erin M. Buchanan[51], Ernest Baskin[53], Evgeniya Hristova[54], Fanny Lalot[55], Farzan Karimi-Malekabadi[56], Fatih Bayrak[57], Fei Xin[27], Fernando Ruiz-Dodobara[58], Francesca Dumančić[59], Gabriela M. Marcu[60], Gilad Feldman[31], Halil Emre Kocalar[61], Ian D. Stephen[62], Iris Žeželj[30], Ivy Y-M. Lau[14], Jackson G. Lu[63], Jan Philipp Röer[64], Jared B. Celniker[65], Jasper Hidding[4], Jennifer Nicole Lenz[66], Jerome Olsen[67], Jin X. Goh[68], Jordan Rose Wagge[69], Jordane Boudesseul[58], Joris Lammers[11], Kaminski Gwenaël[70], Karolina Aleksandra Ścigała[37], Katarzyna Rybus[71], Kathleen Schmidt[72], Keith M. Kendrick[27], Keith M. Kendrick[27], Kevin E. Tiede[73], Kim Peters[23], Kimberley M. Hill[9], Krystian Barzykowski[74], Kyle J. Susa[75], Laura Jane Stevens-Wilson[35], Lauren Victoria Tunstead[35], Leigh Ann Vaughn[76], Leigh S. Wilton[77], Ljiljana B. Lazarevic[30], Luis Eudave[78], Luisa Batalha[79], Macyjane T. Iones[35], Maja Becker[70], Maliki E. Ghossainy[22], Manyu Li[80], Marietta Papadatou-Pastou[81], Marijke C. Leliveld[4], Mario Gollwitzer[82], Mariola Paruzel-Czachura[83], Martin Voracek[67], Mathias Twardawski[82], Matus Adamkovic[84], Maurice Grinberg[54], Mehrad Moeini-Jazani[4], Michael C. Mensink[36], Michał Białek[85], Michal Parzuchowski[71], Miguel A. Vadillo[39], Mirjana Tonković[59], Miroslav Sirota[86], Nada Pop-Jordanova[29], Nate C. Carnes[45], Neil McLatchie[34], Nicholas Owsley[87], Niklas K. Steffens[23], Onurcan Yilmaz[88], Panagiotis Mitkidis[42], Pavol Kačmár[89], Philip Pärnamets[18], Raluca Diana Szekely-Copîndean[90], Ribeiro M. F. F.[91], Richard Huskey[92], Rizqy Amelia Zein[93], Rockwell F. Clancy[94], Samuel Lins[95], Sangsuk Yoon[96], Sara Álvarez Solas[97], Sarah E. Wood[36], Sean P. Mackinnon[98], Shan Gao[27], Sinan Alper[99], Skye-loren West[35], Stefan Czoschke[100], Stefan Stieger[66], Steve M. J. Janssen[101], Sumaya Albalooshi[4], Susann Fiedler[102], Sylwia Adamus[74], Taciano L. Milfont[103], Thomas Rhys Evans[104], Tiago Jessé Souza de Lima[105], Trent M Lynds[98], Tripat Gill[106], Ulf-Dietrich Reips[73], William J. Chopik[107], Xiaoxiao Zheng[27], Xin Liu[108], Yansong Li[109,110], Zahir Vally[111], Zoe Magraw-Mickelson[82], Robert M. Ross[62], Maximilian A. Primbs[112], Miriam Panning[113], Pia K. Andresen[112], Natalia B. Dutra[114], Christopher R. Chartier[52]

[1]IAST, Toulouse School of Economics, [2]Institute of Psychology, ELTE Eötvös Loránd University, [3]University of California, Santa Barbara, United States, [4]University of Groningen, Netherlands, [5]School of Psychology, University of Sussex, Brighton, UK, [6]Sackler Centre for Consciousness

Science, University of Sussex, Brighton, UK, [7]University of Geneva, Switzerland, [8]University of Modena and Reggio Emilia, Italy, [9]University of Northampton, United Kingdom, [10]Department of Psychology, Educational Science and Human Movement, University of Palermo, [11]University of Cologne, Germany, [12]University of Crete, Greece, [13]National University of Córdoba, Argentina, [14]Singapore Management University, Singapore, [15]Comenius University, Slovakia, [16]Swansea University, United Kingdom, [17]Universidad Católica del Norte, Antofagasta, Chile, [18]New York University, United States, [19]Paris Nanterre Universitye, France, [20]University of Salzburg, Austria, [21]Université Savoie Mont-Blanc, France, [22]American University of Beirut, Lebanon, [23]University of Queensland, Australia, [24]University of Amsterdam, Netherlands, [25]Tilburg University, The Netherlands, [26]University of Nîmes, Nîmes, France, [27]University of Electronic Science and Technology of China, China, [28]Georgia Gwinnett College, United States, [29]Academy of Sciences and Arts of North Macedonia, [30]University of Belgrade, Serbia, [31]University of Hong Kong, China, [32]Franklin and Marshall College, United States, [33]Occidental College, United States, [34]Lancaster University, United Kingdom, [35]University of Portsmouth, United Kingdom, [36]University of Wisconsin–Stout, United States, [37]University of Copenhagen, Denmark, [38]Nicholls State University, United States, [39]Autonomous University of Madrid, Spain, [40]University of Denver, United States, [41]University of the Philippines Diliman, Philippines, [42]Aarhus University, Denmark, [43]The University of Auckland, New Zealand, [44]Louisiana State University, United States, [45]California State University, San Marcos, United States, [46]Koç University, Turkey, [47]University of Palermo, Italy, [48]Max Planck Institute for Empirical Aesthetics, Germany, [49]IMT School of Advanced Studies, Italy, [50]Florida International University, United States, [51]Harrisburg University of Science and Technology, United States, [52]Ashland University, United States, [53]Saint Joseph's University, United States, [54]New Bulgarian University, Bulgaria, [55]University of Kent, United Kingdom, [56]Tarbiat Modares University, Iran, [57]Başkent University, Turkey, [58]Instituto de Investigación Científica, Universidad de Lima, [59]University of Zagreb, Croatia, [60]Lucian Blaga University of Sibiu, Romania, [61]Mugla Sitki Koçman University, Turkey, [62]Macquarie University, Australia, [63]Massachusetts Institute of Technology, United States, [64]Witten/Herdecke University, Germany, [65]University of California, Irvine, United States, [66]Karl Landsteiner University of Health Sciences, Austria, [67]University of Vienna, Austria, [68]Colby College, USA, [69]Avila University, United States, [70]University of Toulouse II - Le Mirail, France, [71]University of Social Sciences and Humanities, Poland, [72]Southern Illinois University Carbondale, USA, [73]University of Konstanz, Germany, [74]Jagiellonian University, Poland, [75]California State University, Bakersfield, USA, [76]Ithaca College, United States, [77]Skidmore College, United States, [78]Neuroimaging Laboratory, Division of Neurosciences, Center for Applied Medical Research (CIMA), University of Navarra, Pamplona, Spain, [79]Australian Catholic University, Australia, [80]University of Louisiana at Lafayette, United States, [81]National and Kapodistrian University of Athens, Greece, [82]Ludwig Maximilian University of Munich, Germany, [83]University of Silesia, Poland, [84]University of Prešov, Slovakia, [85]Kozminski University, Poland, [86]University of Essex, United Kingdom, [87]Busara Center for Behavioral Economics, Kenya, [88]Kadir Has University, Turkey, [89]University of Pavol Jozef Šafárik, Slovakia, [90]Babeş-Bolyai University, Romania, [91]University of Brasilia, [92]University of California, Davis, [93]Airlangga University, Indonesia, [94]Shanghai Jiao Tong University, China, [95]University of Porto, Portugal, [96]University of Dayton, United States, [97]Universidad Regional Amazónica Ikiam, Ecuador, [98]Dalhousie University, Canada, [99]Yaşar University, Turkey, [100]Goethe University Frankfurt, Germany, [101]University of Nottingham Malaysia, Malaysia, [102]Max Planck Institute for Research on Collective Goods, Germany, [103]Victoria University of Wellington, New Zealand, [104]Coventry University, United Kingdom, [105]University of Fortaleza, Brazil,, [106]Wilfrid Laurier University, Canada, [107]Michigan State University, United States, [108]Peking

University, China, [109]Department of Psychology, School of Social and Behavioral Sciences, Nanjing University, Nanjing, China, [110]Institute for Brain Sciences, Nanjing University, Nanjing, China, [111]United Arab Emirates University, United Arab Emirates, [112]Radboud University, The Netherlands, [113]Maastricht University, The Netherlands, [114]Universidade do Rio Grande do Norte, Brazil

Corresponding author: Bence Bago, bencebagok@gmail.com, Institute for Advanced Study in Toulouse - Toulouse School of Economics, 21 Allée de Brienne, 31015 Toulouse, France

Abstract

The study of moral judgements is often centered on moral dilemmas in which options consistent with deontological perspectives (i.e., emphasizing rules, individual rights and duties) are in conflict with options consistent with utilitarian judgements (i.e., following the greater good defined through consequences). In a seminal study of this field, Greene et al. (2009) showed that psychological and situational factors (e.g., the intent of the agent or the presence of physical contact between the agent and the victim) can play an important role in moral dilemma judgements. As their study was conducted with US samples, our knowledge is limited concerning the universality of this effect, in general, and the impact of culture on the situational and psychological factors of moral judgements, in particular. Here, we empirically test the universality of intent and personal force on moral dilemma judgements by testing the replicability of the experiments of Greene et al. on a large (N = X,XXX) and diverse sample across the world. We hypothesize that intent and personal force universally increase the unacceptability of outcome-maximizing harm in these dilemmas, and that the effect is stronger in collectivistic than in individualistic cultures due to cultural differences in emotional processing (guilt, shame, anxiety). The relevance of this exploration to a broad range of policy-making problems is discussed.

Keywords: moral thinking, cultural differences, trolley problem, doctrine of double effect, personal force,WEIRD samples, non-WEIRD samples, replication, decision-making

Introduction

Moral dilemmas can be portrayed as decisions between two main conflicting moral principles: utilitarian and deontological. Utilitarian (also referred to as consequentialist) philosophies[1] hold that an action is morally acceptable if it maximizes well-being for the greatest number of people (in terms of saved lives, for example). On the other hand, deontological philosophy[2] evaluates the morality of the action based on the intrinsic nature of the action (i.e., the deontological option often reflects greater concern for the individual rights and duties[3]). The dilemma between these two principles plays a prominent role in law and policy-making decisions ranging from decisions of health budget allocations[4] to the dilemma of self-driving vehicles[5]. This inherent conflict is well illustrated by the so-called trolley problem, which has long interested both philosophers and psychologists. One version of the dilemma is presented as follows[6]:

You are a railway controller. There is a runaway trolley barrelling down the railway tracks. Ahead, on the tracks, there are 5 workmen. The trolley is headed straight for them and they will be killed if nothing is done. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a side track and you can save the 5 workmen on the main track. You notice that there are 2 workmen on the side track. So there will be 2 workmen who will be killed if you pull the lever and change the tracks but the 5 workmen on the main track will be saved. Is it morally acceptable for you to pull the lever?

A deontological decision-maker would argue that pulling the lever is morally unacceptable, as it would be murder[1]. On the other hand, utilitarianism would suggest that it is morally acceptable to pull the lever, as it would maximize the number of saved lives.

In an alternative version of the dilemma, one has to push a man off a footbridge in front of the trolley ("footbridge" scenario). This man will die but will stop the trolley, and the five people in the way of the trolley will be saved. Interestingly, people are less likely to make a decision consistent with utilitarian perspectives[2] in the footbridge scenario compared to the switch scenario. The difference between the utilitarian response rate of the two problems became the basis of investigations of many influential cognitive theories in the field of moral judgement[3,7–13]. The fact that people respond differently to the two trolley dilemmas was proposed to be explained by people's adherence to the so-called doctrine of double effect (DDE[6,9]). A simple version of the doctrine is that harm is permissible as an unintentional side-effect of a good result. This doctrine is the basis of many policies in several countries all

[1] In fact, deontological rules are often more complicated than this. Some of the deontological rules would allow for killing in this situation. The terms "deontological" and "utilitarian/consequentialist" are labels we use to refer to certain responses.

[2] We will call these "utilitarian" responses but the fact that these decisions are consistent with utilitarianism does not indicate that people gave them out of utilitarian principles; the same is true for "deontological" responses[7,8].

over parts of the world, concerning issues such as abortion[6], euthanasia[14], international armed conflict regulations[15,16], and even international business ethics[17]. According to this doctrine, it is morally impermissible to bomb civilians to win a war, even if ending the war would eventually save more lives. However, if civilians die in a bombing of a nearby weapons factory as a side-effect, the bombing is morally acceptable. The way people perceive or act on these moral rules can influence the policies that are accepted or even followed - as we can already see in the case of driverless cars, which sometimes have to decide between sacrificing their own passengers and saving one or more pedestrians[5].

Greene et al.[20] and Cushman et al.[9], however, argued that the difference in utilitarian response rates cannot simply be explained by the DDE. Greene et al. presented evidence for the interaction of the intention of harm (i.e., harm as means or side effect; referring to the doctrine of double effect) and personal force (i.e., whether or not the agent had to use personal effort to kill the victim and save more people) on moral acceptability ratings. More concretely, people were less likely to judge sacrificing one person to save more people morally acceptable when they had to use their personal force to kill the person *and* the death of this person was required to save more people (this is what is meant by *intending* the harm). Hence, they concluded that people are more sensitive to the doctrine of double effect when they have to use their own physical force. Despite some exceptions[26,27], most of the evidence for this conclusion comes from samples of WEIRD (Western, Educated, Industrialized, Rich, Democratic[23,24]) societies, leaving the question open whether these effects are psychologically universal[25] and not culture-specific.

This study tests three cross-cultural hypotheses:

(1) The effects of personal force on moral judgements are culturally universal.
(2) The interactional effect of personal force and intention on moral judgements is culturally universal.
(3) Collectivism-individualism has a moderating effect on the degree to which personal force and intention affect moral judgements in a way that their effect is stronger in more collectivistic cultures.

The first and second hypotheses, that the effects of personal force and intention on moral judgements are culturally universal, come from their relatedness to interpersonal violence. People seem to exhibit a general tendency to avoid causing violent harms (e.g., murder)[19,20], and they are more likely to perceive actions as violent or harmful when they supposed to use personal force or intention[3]. As a result, people are more likely to behave in a deontological way when personal force or intention is present in the dilemma. As all cultures regulate interpersonal violence[21], the H, we expect to find that both intention and personal force, as well as their interaction, have an effect on moral judgements in different cultures. The literature seems to be in accordance with this thesis; Chinese[25–27] and Russian[28] participants seemed to produce similar effects on moral dilemmas to Americans and Western Europeans, and even small scale societies tended to be susceptible for the effect of intention[22,23].

Even though we anticipate that the effect of personal force and intention will emerge universally across cultures, we nonetheless expect cultural differences to moderate these effects.

The effect of personal force on moral judgement has been attributed to emotional processes[9,24–26], specifically social emotions (such as guilt, shame or regret)[25,27]; the potential use of personal force makes people feel guilt or shame before making a decision, and, therefore, they will rate actions that use personal force as morally less acceptable. There is a convincing argument that these social emotions are universal[28–30], with some cultural variation in their intensity and the social contexts in which they are experienced[28–30] . It has been argued that shame and guilt are more important in interdependent, collectivistic cultures (as their function is argued to be social control); people living in East Asian countries have reported experiencing them more frequently and more intensively[28–30]. Other findings suggest that it is anxiety that mediates the effect of intention and personal force[26], but anxiety (social anxiety in particular) has also been positively associated with collectivism[31], pointing to the same direction. Hence, we hypothesize that people living in collectivistic cultures will judge actions that involve personal force and intention as morally less acceptable than people in individualistic cultures. Utilitarian responding in moral dilemma judgements has also been associated with low levels of empathic concern[32]. People living in collectivistic cultures have also been suggested to exhibit higher levels of empathy[33,34]. Hence, we predict that individualism/collectivism factor will also have a main effect on utilitarian responding: collectivists will be less utilitarian in general, due to their higher levels of empathic concern.

Besides our hypothesis testing, we also collect a number of additional country-level as well as individual measures for exploratory purposes. These measures have been previously shown to be related to moral judgement such as economic status[35], individualism-collectivism[35], and religiosity[36] and an alternative measure of utilitarian responding[37–40].

The investigation of the question is crucial for advancing the field for the following reasons:

1) The original article has been very influential[3], but to the best of the authors' knowledge, a direct replication of the original study has never been conducted and/or published.
2) Our knowledge is scarce on the cultural universality of the effect of personal force and intention in moral judgements.
3) The resulting database (with many types of trolley problems and additional measures) could assist and guide future research and applications of moral thinking.

Overview

[3] Google Scholar listed 515 citations at the time of the writing of this manuscript.

In the first part of our study, we test the universality of the role of personal force in moral judgements with a direct replication of Study 1a conducted by Greene et al. that aimed to explore the effect of personal force on moral dilemma judgements (whether the decision maker uses personal force). In their study, Greene et al. found evidence that the application of personal force decreases moral acceptability of the utilitarian action (Hypothesis 1a, 1b). In the second part, we test the universality of the interactional effect of personal force and intention on moral dilemma judgements, by replicating Study 2 of Greene et al. (Hypothesis 2a, 2b) with partially different moral dilemmas. Furthermore, we test our hypothesis according to which collectivism moderates the effect of intention and personal force (Hypothesis 3). In addition, we collect various additional measures for exploratory purposes.

Methods

Participants

To recruit a large culturally and demographically diverse sample of participants, we have recruited collaborating laboratories through the Psychological Science Accelerator [41]. So far, the data collection team includes 146 labs from 52 countries. All of these participating laboratories obtained IRB approval (to be verified before the last round of Stage 1 submission). Combined, these labs committed to collect a minimum number of 18,637 participants. More labs are expected to be recruited before data collection commences. The participating labs and the country in which they will collect the data can be seen in Table S1 (the complete list of contributing labs by country can be found in our Supplementary Material A; available on our OSF page:https://osf.io/kzmhy/). Each lab will recruit participants for the study by sending out the survey link along with the consent form to their participant pool, online platforms (such as Mturk), or testing them in the research lab. Eligibility for participation will be based on age (≥ 18 years) and being a native speaker of the language of the test (more details on this criterion in the *Controlling for possible confounds* section).

We do not collect any identifiable private data during the project. Each lab ascertained that the local institutional ethical review board agrees with the proposed data collection. This study was conducted in accordance with the Declaration of Helsinki. The IRB approvals are available on our OSF project page: https://osf.io/j6kte/.

Materials and Procedure

Moral dilemmas. We will use a total of 6 trolley dilemmas, namely: *footbridge switch, standard footbridge, footbridge pole, loop, obstacle collide* (taken from Greene et al. 2009),

and *standard switch*[4]. Each of these scenarios represents a different condition; for example, in the *standard footbridge* scenario both intention and personal force are required to push the man off the bridge. All the materials are presented in the Supplemental material B at https://osf.io/z3v6f/. As in the original experiments, every participant will only be assigned to one of these dilemmas. The problems will be accompanied by a drawn sketch to aid understanding. Following the original procedure, after presenting each problem, participants will be asked whether the described action (e.g., pushing the man to save five people) is morally acceptable or not (Yes/No response). After this judgement, participants will have to indicate on a numbered Likert-type scale ranging from 1 (*completely unacceptable*) to 9 (*completely acceptable*) the extent to which they think that the given action is morally acceptable. Next, participants will be asked to write down the justification of their decision on the same page. After participants are presented with the first trolley dilemma, they will be presented with a second dilemma from the same condition, without drawn sketches. As second dilemmas, we will use the so-called "Speedboat dilemmas". These dilemmas will be taken from Study 1b and 2b of Greene et al., and can be found in Supplementary Materials B, with exception of the dilemmas in the obstacle collide and standard footbridge conditions, which were provided by Joshua Greene during the review of the paper.
The order will be fixed for dilemma presentation, so that the trolley version will always be presented first.

Additional measures. Although the exploration of individual-level factors behind moral thinking is not the aim of the present research, to enrich our database for future studies and secondary analyses, we will add additional questions to our surveys about individual-level: 1) total yearly household income; 2) place of living (urban or rural area); 3) position on the four-dimensional Individualism-Collectivism scale[34] (16 items) for disentangling cultural differences in participants' responses[42]; and 4) Religion: Specific religion of the participant will be asked, plus one question to measure their level of religiosity: "On a scale from 1 to 10, how religious are you?". Furthermore, we will include the Oxford Utilitarianism Scale[28] (9 items). Following these questions, participants' highest level of education, age, and sex will also be recorded. We will also record the participants' country of origin, and whether the participant is coming from an immigrant background. At the end of the survey, we will ask whether the participant experienced any technical problems or whether they have any comments on the experiment. All questions are available at: https://osf.io/z3v6f/.

Procedure. The experiment will be administered by using a centralised online survey, which participants can answer remotely via the internet or in the lab. We will use the original instructions, as presented in the Supplemental Materials. After responding to the dilemmas, participants will be asked to answer three questions: (1) a measure of careless responding (question about the specifics of the trolley scenario) (2) whether they found the material confusing; and (3) whether they found the description of the problem realistic. After these

[4] This item was provided by Joshua Greene during the review of this paper.

questions, participants will be directed to our series of questionnaires: the Oxford Utilitarianism Scale, followed by the Individualism-Collectivism Scale and the measures of religion. Next, we will administer the demographic questions (income, place of living, country of origin, immigrant background, level of education, age, and sex). Afterwards, we will add 3 further questions to measure careless responses, and, finally, we will ask for further comments/any experienced technical problems.

Controlling for possible confounds. To avoid second language effects on moral judgement [43], only participants who are native speakers of the language of the experiment can take part in it. To assure that only native speakers participate, we will ask participants to indicate their native language(s). Bilinguals can choose their preferred language. Anyone whose native language is different from the language of the survey will be excluded from our data analyses.

Following Greene et al.'s procedure, also data from those participants who report that they found the material confusing (response *a*. to the first question on p. 33 of the original Supplemental Material) will be excluded from the analyses.

Data of participants who report having experienced any technical problems during the experiment will be excluded from all analyses. To avoid careless responses, we are going to add three bogus items at the end of the survey. We are going to ask participants very basic questions (e.g., "I was born on February 30th.") to which incorrect answering indicates careless responding[44]. We will exclude data from all participants who gave an incorrect response to any of these questions. Moreover, we will introduce two additional questions (posed right after the moral dilemmas), asking participants about the specifics of the trolley and speedboat scenarios that they had been presented with, in order to test whether they were paying attention when reading the scenario. Specifically, they will be asked to select the option which most accurately describes the situation that they had been presented with. Each option will describe the nature of the physical action; the key manipulation in the experiment. As the attention on the trolley and speedboat dilemmas will be measured by different questions, when analysing the responses on the trolley dilemma, we will exclude the data of those giving an incorrect response to the trolley attention check question, and same for the speedboat dilemma. For example, people who gave a correct response on the trolley attention check question, but not on the speedboat version, will be included when analysing the trolley dilemma, and excluded when analysing the speedboat version. These questions can be found in Supplementary Material B.

Moral dilemmas are becoming more and more popular, and it is possible that people with previous knowledge on these dilemmas respond differently. To address this potential problem at the end of the experiment, participants will be asked the following question: "Before this experiment, were you familiar with moral dilemmas of this kind, in which you can save more people by causing the death of one person? Please rate it on a scale from 1 (absolutely not

familiar) to 5 (absolutely familiar)". Data of people who are familiar with the trolley problem or such moral dilemmas (respond with 4 or 5 on this scale) will be excluded from our analyses. Additionally, participating labs will be asked to avoid recruiting philosophers or philosophy students: philosophy students are likely to have heard about trolley problems, and we would like to minimise the number of participants to be excluded. Note that these filters will only be applied to our analyses but not on the data set which will be openly available.

Cultural classification of countries. To test our first hypothesis on the universality of the effect of personal force and intention on moral judgements, we used the cultural classification of Awad et al.[35]. To test the cultural universality hypothesis, a comprehensive cultural classification is needed that encompasses multiple sources of cultural variability. Based on surveyed moral preferences, they identified three distinct clusters of countries: Eastern, Southern, and Western. They argued that this cluster structure is broadly consistent with the alternative but more complex Inglehart-Welzel cultural map[34].

Notable deviations between this study and the design of Greene et al.

Besides the multinational data collection that forms the crux of our project, the first important methodological difference between this study and the original study is that the original study was conducted by paper and pencil, whereas we will administer the experiment online. Of note, recent research found no evidence for a difference between the behaviour of participants who took part in the experiment online versus those who took part in the experiment in the lab [45]. We also added one change in the introduction of the experiment (see Supplementary Material B); participants will not be given the opportunity to ask the researcher any questions before the experiment (as the experiment can be administered online, they will not have the opportunity to do so).

The second important change in this experiment is that participants will be presented with two moral dilemmas in one condition, instead of one. These additional dilemmas will be analysed separately, as they were in the original experiment. The third difference is that for Study 2, we will use different moral dilemmas than those that were used by Greene et al.; the standard switch and footbridge dilemmas will be used instead of the loop weight and obstacle push dilemmas, respectively. These dilemmas are not different from the ones used by Greene et al. in their structural characteristics, only on surface characteristics. That is, in the standard switch the harm is unintended and no personal force is required, while in the standard footbridge dilemma, the harm is intended and requires personal force. By including the standard switch and standard footbridge scenarios instead of the original ones, we gain further insight into the data. Imagine for example, that the personal force effect does not replicate in one of the cultural clusters. One explanation for this is that people are simply not sensitive to the effect of personal force in that cluster. However, it might also be the case that utilitarian response rates to similar dilemmas increase over time[46]. If so, we should see that the replicated difference between the standard footbridge and switch dilemmas is shrinking or

disappeared. Furthermore, by comparing the standard footbridge to the footbridge pole dilemmas, we can test the effect of physical contact, and by comparing the standard switch case to the footbridge switch case to confirm the effect of intention.
Finally, in the original experiment, Greene et al. excluded participants who did not manage to suspend disbelief. Nevertheless, as they noted, this had no effect on their results. Thus, we decided that we will not use this exclusion criterion.

Language adaptation. The participating labs will translate the survey items into the language of the participant pool following the translation process of the PSA[5]. The survey will be administered in the language of the local subject pool.

Table 1.
The Cultural Classification of Countries of Participating Labs Following Awad et al.

Western	Southern	Eastern
Australia, Austria, Brazil, Bulgaria, Canada, Croatia, Denmark, Germany, Greece, Italy, Kazakhstan, Kenya, Netherlands, New Zealand, Poland, Portugal, Romania, Russia, Serbia, Singapore, South Africa, Spain, Switzerland, United Kingdom, United States of America	Argentina, Chile, Colombia, Czechia, Ecuador, El Salvador, France, Hungary, Mexico, Peru, Philippines, Slovakia, Turkey	Mainland China and Hong Kong, India, Indonesia, Iran, Japan, Lebanon, Malaysia, North Macedonia, Pakistan, Thailand, United Arab Emirates

Confirmatory Replication Analyses

We focus our analyses on the question of universality of Greene et al.'s two most important claims. We will conduct independent analyses in each country cluster and report them separately. Note that these analyses will only be conducted on trolley problems. Separate analyses will be conducted on the remaining moral dilemmas (see Additional analysis).

Hypothesis 1a: We hypothesise to find an effect of personal force on moral judgement on the Western cluster (replication of the original effect).

Hypothesis 1b: If the effect of personal force is culturally universal, we should find an effect of personal force on the moral acceptability ratings (Greene et al., Study 1) in the Southern and Eastern cultural clusters as well.

[5] https://psysciacc.org/translation-process/

Hypothesis 2a: There should be an interaction between personal force and intention (Greene et al., Study 2) in the Western cluster (replication of original effects). More specifically, the intention factor is expected to be larger when personal force is present, as compared to when personal force is absent.

Hypothesis 2b: If this effect is culturally universal, we should find this effect in the Southern and Eastern cultural clusters as well.

Unlike in the original study, we will employ Bayesian analyses to gain information from our data concerning the strength of evidence for the null and alternative hypotheses. The Bayes factor indicates the relative evidence provided by the data comparing two hypotheses[47]. Regarding the threshold of good enough Bayesian evidence, we will follow the recommendations of [48] and set the decision threshold of BF_{10} to > 10 for H_1 and < 1/10 for H_0. We will use informed priors for the alternative model: a one-tailed Cauchy distribution with a mode of zero and a scale $r = 0.26$ (Hypothesis 1a and 1b) and $r = 0.19$ (Hypothesis 2a and 2b) on the standardized effect size using the BayesFactor package in R[49] for the analysis. These priors are based on the effect sizes that we expect to find as explained below in the sample size estimation section. We will implement all of our analyses with the R statistical software[50].

To test Hypothesis 1a and 1b, we will compare the moral acceptability ratings given on the footbridge switch problem and footbridge pole dilemma, with the moral acceptability rating of the footbridge switch dilemma expected to be higher. More concretely, we will perform three one-sided Bayesian *t*-tests with the same comparison in each cultural group. For each cultural cluster, we will conclude that we replicated the original effect if Bayes factor (BF_{10}) > 10, we will conclude that we found a null effect if $BF_{10} < 1/10$, and we will conclude that the results are inconclusive if we find a BF_{10} in between these numbers (see below for justification of these thresholds).

To test Hypothesis 2a and 2b, we will test the interaction of personal force and intention in each cultural cluster, separately. We will conduct Bayesian linear regression analysis in each cultural cluster. The Bayes factor of interest is defined as the quotient of the model including the interaction and two main effect (numerator) and the model including only the two main effects (denominator). For each cultural group, we will conclude that we replicated the original effect if the Bayes factor of the interaction (BF_{10}) > 10, we will conclude that we found a null effect if $BF_{10} < 1/10$, and we will conclude that the results are inconclusive if we find a BF_{10} in between these values (see below for justification of these thresholds). To further understand the direction of the interaction, we will plot out the results in each cultural cluster. To conclude the replication of the original effect, we should find that the intention effect is higher in the personal force condition than in the no personal force condition.

Note that we will conduct and report the frequentist version of the proposed analysis (e.g., *t* tests for each hypothesis, for each cultural class) for the sake of comparability of the original and our results. Nevertheless, we will regard the results of our Bayesian analyses the basis of our statistical inference. The frequentist statistics will only be added as the supplementary material, and no inference will be drawn from it.

Robustness analyses

To probe the robustness of our conclusions to the scaling factor of the Cauchy distribution used as the prior of H1, we will report Robustness Regions for each Bayes factor. Robustness Regions will be notated as RR[min, max], where min indicates the smallest and max indicates the largest scaling factor that would lead us to the same conclusion as the originally chosen scaling factor[51].

Sampling plan and stopping rule

As the data are planned to be collected globally, our knowledge is insufficient concerning the noise of the measurement and the rate of exclusion in the various samples, which are needed for an accurate sample size estimation. For this reason, we propose a sequential data acquisition. That is, first, we will launch Study 1 (Hypotheses 1a and 1b), and collect data in sequences from 500 participants per cluster per condition; from 3,000 participants altogether (after all exclusions). We will stop data collection after each sequence. At these stops, we will conduct our planned Bayesian analyses. Should the BF reach the preset thresholds in a given cluster, we will stop data collection for that cluster. If, in a cluster, the BF thresholds were not reached, we will continue data collection with 200 additional participants per cluster per condition, and then re-analyse the data, repeating this procedure until one of the BF thresholds is reached, or the participant pool is exhausted.

Should we not reach this limit with our present capacity of ~19,000 participants, we would extend the data collection to a new semester. In the unlikely case that we do not reach our evidence threshold within 12 months, we would report our final results, acknowledging the limited strength of the findings.

We will launch Study 2 data collection in a given cluster only when the analysis of Study 1 is conclusive. In Study 2, we will conduct the analysis only when we have exhausted our resources.

Sample Size estimation

To calculate our needs for data collection, we conducted a rough sample size estimation. Assuming that the original effect size is found in Study 1 ($d = 0.4$), our sample size estimation indicated that we would require 500 participants per condition per cluster (3,000 altogether), while if the original effect size is to be found in Study 2 ($d = 0.28$), our estimation indicated that we would need 1,800 participants per condition per cluster (21,600 altogether for Study 2) to obtain 95% of power in detecting the effect. A detailed description

of the Sample Size estimation can be found in Supplementary Materials C at: https://osf.io/vq67t/.

Testing the effect of country-level individualism/collectivism on the effects of personal force and intention.

Our third hypothesis proposed that collectivism increases the effects of personal force and intention. As a measure of country-level individualism and collectivism, we will add the Collectivism measure from the Cultural Distance WEIRD scale (countries difference in terms of individualism from the United States)[52] as a continuous variable in our model. We test whether collectivism interacts with personal force and intention (Hypothesis 3), as explained in the introduction. Hypothesis 3 expects to find a three-way interaction between collectivism, intention, and personal force, for which we will use the dilemmas we used to test Hypotheses 2a and 2b. In this analysis, we will use a Cauchy distribution with a scale of r = 0.37 (same we used to test Hypothesis 2a and 2b, i.e., the test of the interaction) as prior. Should we find evidence for null effect (BF < 1/10) of the interaction of individualism/collectivism, personal force, and intention, we will conclude that individualism/collectivism does not moderate the effect of personal force and intention.

Additional data analysis

Analysis of the additional moral dilemmas

Study 1.

As we explained above, each participant will have to give response on two moral dilemmas. For Study 1 (effect of personal force), we will conduct the same analysis on the rest of the moral dilemmas, without the trolley versions, as in the original study (Study 1b; Greene et al.).

Study 2.

We will conduct the same analysis (interaction of personal force and intention) on the rest of Speedboat dilemmas, without the trolley versions.

Further tests

Effect of physical contact and intention. With this set of items, we will be able to assess the effect of physical contact, by comparing the standard footbridge and footbridge pole dilemmas. We will also assess the effect of intention by comparing the standard switch case with the footbridge switch case. These analyses will be done in every cluster, and we will use Bayesian t-tests for these comparisons. We will use the same prior we use for the assessment of the effect of physical force ($r = 0.26$). This analysis will be done separately on the trolley and speedboat dilemmas.

Comparing the standard switch and standard footbridge dilemmas. For the reasons explained earlier, we are planning to compare the standard footbridge and standard switch dilemmas, in each cultural cluster. For this, we will conduct a Bayesian t-test, with the same prior previously used for the assessment of the effect of physical force ($d = 0.26$). This analysis will be done separately for the trolley and speedboat dilemmas.

Oxford Utilitarianism Scale. We will compute the mean scores and 95% confidence intervals on the Oxford Utilitarianism Scale for each cultural cluster to explore potential cultural differences.

Individual-level horizontal and vertical individualism-collectivism. Triandis and Gelfand[45] defined individualistic and collectivistic cultural tendencies among 4 dimensions: *vertical individualism, vertical collectivism, horizontal individualism,* and *horizontal collectivism*. We will add these continuous measures to our Bayesian linear regression analysis. The predictive power of all four measures will be assessed separately.

Including familiar participants. A potentially large number of participants will be excluded due to familiarity with the trolley dilemma, and there is a possibility that this exclusion criterion will affect the data from some countries or cultural clusters more than others. To avoid this potential sampling bias, we will compute all above-listed analyses on moral dilemmas (confirmatory and exploratory) on the full sample in which we do not exclude the participants who were familiar with the trolley problem. Second, we will compute all analyses specifically on data coming from people who were familiar with the trolley problem in order to compare the results of "familiar" and "unfamiliar" participants.

Additional information

Data pre-processing. Data will be aggregated from the online database of our survey. All exclusion criteria will be applied before all statistical analyses. The data for the excluded participants will still be saved with the full data but will not be analysed. The data management plan for the research can be found on our OSF page: https://osf.io/m9nuq/.

Pilot testing. To ascertain that the survey software operates without any technical problems, we plan to conduct a pilot test in which each participating lab will be expected to collect data from 10 participants. We will only assess the expected functioning of the survey software without analysing the collected data.

Timeline. We plan to finish collection within six months from Stage 1 in principle acceptance and we plan to submit our report within one month from then.

Discussion

[Here, we will discuss the implications of the results, and put them in a greater framework, including the potential criticisms of the paradigm]

Open Science Statement

The study will be publicly preregistered on the OSF site after Stage 1 in principle acceptance. Collected anonymised raw and processed data will be publicly shared on the OSF page of the project. Code for data management and statistical analyses will be written in R and will be made open access too. All materials (including the translated items to all languages) will also be made available through OSF.

Data availability statement

All anonymised data collected will be openly available on the project's OSF page.

Code availability statement

All code used for data management or analysis will be openly available on the project's OSF page.

Competing interest statement

The authors declare no competing interests.

Acknowledgements

We would like to thank Joshua Greene and seven anonymous reviewers for their hard work on commenting on this proposal. We would like to thank the PSA network for continuous support in all aspects of the project. We gratefully acknowledge support from ANR grant ANR-17-EURE-0010 (Investissements d'Avenir program, BB), ANR Labex IAST (BB) and the Scientific Research Fund Flanders (FWO-Vlaanderen, BB).

Author Contribution

BB and BA developed the study concept, proposed the initial design, and wrote the initial proposal draft. ZK and BB conducted the power analysis, TN and BP helped with creating the R code of the power analysis. JP and CC were responsible for project management. MK created the data management plan. RK and BP shared some suggestions on the data analysis plan. The rest of the authors provided comments on the manuscript, submitted ethical approvals to the responsible ethical boards, and will contribute with recruiting participants and conducting the experiment. BA and CC managed the OSF repository of the project. The rest of the co-authors took part in the early development of the project and contributed to the writing of the Stage 1 registered report.

References

1. Mill, J. S. & Bentham, J. *Utilitarianism and other essays*. (Penguin, 1987).
2. Kant, I. *Groundwork for the metaphysics of morals.* (Yale University Press, 1785).
3. Greene, J. D. *Moral tribes: emotion, reason and the gap between us and them*. (Penguin Press, 2013).
4. London, J. A. How should we model rare disease allocation decisions? *Hastings Cent. Rep.* **42**, 3 (2014).
5. Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
6. Foot, P. The problem of abortion and the doctrine of double effect. *Oxf. Rev.* **5**, 5–15 (1967).
7. Baron, J. Utilitarian vs. deontological reasoning: method, results, and theory. in *Moral inferences* (eds. Bonnefon, J.-F. & Trémolière, B.) 137–151 (Psychology Press, 2017).
8. Baron, J. & Gürçay, B. A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Mem. Cognit.* **45**, 566–575 (2017).
9. Cushman, F., Young, L. & Hauser, M. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychol. Sci.* **17**, 1082–1089 (2006).
10. Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M. & Cohen, J. D. The neural bases of cognitive conflict and control in moral judgment. *Neuron* **44**, 389–400 (2004).
11. Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. An fMRI investigation of emotional engagement in moral judgment. *Science* **293**, 2105–2108 (2001).
12. Gürçay, B. & Baron, J. Challenges for the sequential two-system model of moral judgement. *Think. Reason.* **23**, 49–80 (2017).
13. Mikhail, J. Universal moral grammar: Theory, evidence and the future. *Trends Cogn. Sci.* **11**, 143–152 (2007).

14. Boyle, J. Medical ethics and double effect: the case of terminal sedation. *Theor. Med. Bioeth.* **25**, 51–60 (2004).
15. Gross, M. L. Bioethics and armed conflict: Mapping the moral dimensions of medicine and war. *Hastings Cent. Rep.* **34**, 22–30 (2004).
16. Gross, M. L. Killing civilians intentionally: double effect, reprisal, and necessity in the Middle East. *Polit. Sci. Q.* **120**, 555–579 (2005).
17. Tully, P. A. The doctrine of double effect and the question of constraints on business decisions. *J. Bus. Ethics* **58**, 51–63 (2005).
18. Greene, J. D. *et al.* Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition* **111**, 364–371 (2009).
19. Cushman, F. Action, outcome, and value: A dual-system framework for morality. *Personal. Soc. Psychol. Rev.* **17**, 273–292 (2013).
20. Cushman, F., Gray, K., Gaffey, A. & Mendes, W. B. Simulating murder: The aversion to harmful action. *Emotion* **12**, 2 (2012).
21. Ellsworth, R. M. & Walker, R. S. Sociobiology of lethal violence in small-scale societies. in *The Routledge International Handbook of Biosocial Criminology* 85–102 (Routledge, 2014).
22. Abarbanell, L. & Hauser, M. D. Mayan morality: An exploration of permissible harms. *Cognition* **115**, 207–224 (2010).
23. Barrett, H. C. *et al.* Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proc. Natl. Acad. Sci.* **113**, 4688–4693 (2016).
24. Cushman, F., Young, L. & Greene, J. D. Our multi-system moral psychology: Towards a consensus view. in *The Oxford handbook of moral psychology* 47–71 (2010).
25. Koenigs, M. *et al.* Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* **446**, 908–911 (2007).
26. Perkins, A. M. *et al.* A dose of ruthlessness: Interpersonal moral judgment is hardened

by the anti-anxiety drug lorazepam. *J. Exp. Psychol. Gen.* **142**, 612–620 (2013).

27. Szekely, R. D. & Miu, A. C. Incidental emotions in moral dilemmas: The influence of emotion regulation. *Cogn. Emot.* **29**, 64–75 (2015).
28. Johnson, R. C. *et al.* Guilt, shame, and adjustment in three cultures. *Personal. Individ. Differ.* **8**, 357–364 (1987).
29. Tracy, J. L. & Matsumoto, D. The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proc. Natl. Acad. Sci.* **105**, 11655–11660 (2008).
30. Scollon, C. N., Diener, E., Oishi, S. & Biswas-Diener, R. Emotions across cultures and methods. *J. Cross-Cult. Psychol.* **35**, 304–326 (2004).
31. Heinrichs, N. *et al.* Cultural differences in perceived social norms and social anxiety. *Behav. Res. Ther.* **44**, 1187–1197 (2006).
32. Gleichgerrcht, E. & Young, L. Low levels of empathic concern predict utilitarian moral judgment. *PloS One* **8**, e60418 (2013).
33. Luo, S. *et al.* Interaction between oxytocin receptor polymorphism and interdependent culture values on human empathy. *Soc. Cogn. Affect. Neurosci.* **10**, 1273–1281 (2015).
34. Cheon, B. K. *et al.* Cultural influences on neural basis of intergroup empathy. *NeuroImage* **57**, 642–650 (2011).
35. Awad, E. *et al.* The Moral Machine experiment. *Nature* **563**, 59–64 (2018).
36. Koenig, L. B., McGue, M., Krueger, R. F. & Bouchard Jr, T. J. Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings. *J. Pers.* **73**, 471–488 (2005).
37. Kahane, G. On the wrong track: Process and content in moral psychology. *Mind Lang.* **27**, 519–545 (2012).
38. Kahane, G., Everett, J. A., Earp, B. D., Farias, M. & Savulescu, J. 'Utilitarian'judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good.

Cognition **134**, 193–209 (2015).

39. Kahane, G. *et al.* Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychol. Rev.* **125**, 131–164 (2017).
40. Conway, P., Goldstein-Greenwood, J., Polacek, D. & Greene, J. D. Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition* **179**, 241–265 (2018).
41. Moshontz, H. *et al.* The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Adv. Methods Pract. Psychol. Sci.* **1**, 501–515 (2018).
42. Bond, M. H. & van de Vijver, F. J. R. Making scientific sense of cultural differences in psychological outcomes: Unpackaging the Magnum Mysterium. in *Culture and psychology. Cross-cultural research methods in psychology* (eds. Matsumoto, D. & van de Vijver, F. J. R.) 75–100 (Cambridge University Press, 2011).
43. Costa, A. *et al.* Your morals depend on language. *PloS One* **9**, e94842 (2014).
44. Meade, A. W. & Craig, S. B. Identifying careless responses in survey data. *Psychol. Methods* **17**, 437–455 (2012).
45. Klein, R. A. *et al.* Many Labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).
46. Hannikainen, I. R., Machery, E. & Cushman, F. A. Is utilitarian sacrifice becoming more morally permissible? *Cognition* **170**, 95–101 (2018).
47. Dienes, Z. Using Bayes to get the most out of non-significant results. *Front. Psychol.* **5**, 781 (2014).
48. Schönbrodt, F. D. & Wagenmakers, E.-J. Bayes factor design analysis: Planning for compelling evidence. *Psychon. Bull. Rev.* **25**, 128–142 (2018).
49. Morey, R. D., Rouder, J. N. & Jamil, T. *BayesFactor: Computation of Bayes factors for common designs*. (2015).

50. R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* (2018).
51. Dienes, Z. How do I know what my theory predicts? Accessed via: https://psyarxiv.com/yqaj4. *Adv. Methods Pract. Psychol. Sci.* **2**, 364–377 (2019).
52. Muthukrishna, M. *et al.* Beyond WEIRD psychology: measuring and mapping scales of cultural and psychological distance. *Available SSRN 3259613* (2018).