



# Using Ethical Dilemmas To Predict Antisocial Choices With Real Payoff Consequences: An Experimental Study

By: **David L. Dickinson** and David Masclat

## Abstract

In this paper we investigate the relationship between ethical choices and antisocial behaviors. To address this issue we ran a within-subjects laboratory experiment that included both a classic (hypothetical) moral dilemma (using the well-known Trolley problem) and a real-payoff money-burning experiment. A main contribution is that our Trolley dilemmas separate purely utilitarian from more clearly immoral choice options. Our results show that choices in both environments respond to incentives (i.e., the relative price of the ethical decision), and Trolley problem decisions are consistent with previously known results — individuals prefer no action over action, as well as indirect over direct responsibility, when negative consequences would be similar in either instance. In analyzing the determinants of anti-social money burning, our data identify money burning due to inequality aversion, but we also find some evidence of pure nastiness. Importantly, we find that utilitarian behavior in the Trolley dilemma is not linked to antisocial money burning, which contrasts with previous conclusions in the literature. Nevertheless, we observe that the willingness to commit more clearly ethically dubious acts in the Trolley problem significantly predicts money burning and, more specifically, nastiness. We conclude that choices in hypothetical environments may be useful for predicting antisocial behaviors that have real payoff consequences and efficiency implications.

**Dickinson, D.** & Masclat, D. (2019). Using ethical dilemmas to predict antisocial choices with real payoff consequences: An experimental study, *Journal of Economic Behavior & Organization*, v. 166, 2019. Pages 195-215. <https://doi.org/10.1016/j.jebo.2019.08.023>. Publisher version of record available at: <http://www.sciencedirect.com/science/article/pii/S0167268119302719>



# Using ethical dilemmas to predict antisocial choices with real payoff consequences: An experimental study

David L. Dickinson<sup>a,\*</sup>, David Masclet<sup>b</sup>

<sup>a</sup>Appalachian State University, IZA, ESI

<sup>b</sup>Université de Rennes 1, CREM, CNRS

## ARTICLE INFO

### Article history:

Received 7 February 2019

Revised 23 August 2019

Accepted 24 August 2019

Available online 18 September 2019

### JEL classification:

C90

C91

Z10

### Keywords:

Experiments

Money burning

Ethical dilemmas

Anti-social behavior

Trolley problem

## ABSTRACT

In this paper we investigate the relationship between ethical choices and anti-social behaviors. To address this issue we ran a within-subjects laboratory experiment that included both a classic (hypothetical) moral dilemma (using the well-known Trolley problem) and a real-payoff money-burning experiment. A main contribution is that our Trolley dilemmas separate purely utilitarian from more clearly immoral choice options. Our results show that choices in both environments respond to incentives (i.e., the relative price of the ethical decision), and Trolley problem decisions are consistent with previously known results—individuals prefer no action over action, as well as indirect over direct responsibility, when negative consequences would be similar in either instance. In analyzing the determinants of anti-social money burning, our data identify money burning due to inequality aversion, but we also find some evidence of pure nastiness. Importantly, we find that utilitarian behavior in the Trolley dilemma is *not* linked to antisocial money burning, which contrasts with previous conclusions in the literature. Nevertheless, we observe that the willingness to commit more clearly ethically dubious acts in the Trolley problem significantly predicts money burning and, more specifically, nastiness. We conclude that choices in hypothetical environments may be useful for predicting antisocial behaviors that have real payoff consequences and efficiency implications.

© 2019 Elsevier B.V. All rights reserved.

## 1. Introduction

Unethical behavior within organizations is not rare and often results in high costs for the entire society. Antisocial behaviours can result in relational, workplace, or other costs to society that are nontrivial. Cyber-sabotage is now a growing concern, for example (see [Line et al., 2014](#)), and survey data from the U.S. and Europe document antisocial workplace behaviours that include mistreatment, verbal abuse, and sabotage, with estimates indicating these may impact 10%–35% of people in the workplace (see [Charness et al., 2013](#)).<sup>1</sup> Field data examples often pose difficulties in our attempt to under-

\* Corresponding author.

E-mail addresses: [dickinsondl@appstate.edu](mailto:dickinsondl@appstate.edu) (D.L. Dickinson), [david.masplet@univ-rennes1.fr](mailto:david.masplet@univ-rennes1.fr) (D. Masclet).

<sup>1</sup> Data from the U.S. includes research from the Workplace Bullying Institute (<http://www.workplacebullying.org/>) and the Bureau of Labor Statistics ([www.bls.gov](http://www.bls.gov), considering that at least some of the workplace stoppage data represents an exercise of incurring some cost in order to impose even larger costs on a counterpart), and data from the French Ministry of Employment are from the SUMER medical monitoring survey of workplace risks (surveying over 50,000 workers in the 2010 wave, see <https://www.eurofound.europa.eu/observatories/eurwork/articles/working-conditions-france-working-conditions-and-occupational-risks-sumer-2010>).

stand the core determinants of antisocial tendencies given that they may be confounded with self-interest, hidden from view, or contaminated by reputational concerns.<sup>2</sup> While the estimated prevalence of clinical-level antisocial personalities disorders in the general population ranges from 1–4% (Werner et al., 2015), subclinical levels of anti-social personality disorders are more common and on the rise in young adults (Twenge and Foster, 2010). Behavioral metrics that help identify the likelihood that someone engages in antisocial behaviours can therefore be a useful way to prevent antisocial behavioural costs and improve overall welfare (e.g., improved job matching, delegation of authority, mate selection).

At first glance antisocial behaviors appear morally inappropriate. However, some choices that may be considered antisocial may be deemed morally acceptable using an alternative moral metric. For example, when U.S. President Harry Truman decided to drop atomic bombs on Hiroshima and Nagasaki in 1945 to end WWII, he was faced with a great ethical dilemma. Although the bombs would result in many civilian deaths, Truman estimated that it would ultimately cost fewer lives compared to the alternative.<sup>3</sup> This reasoning is based on utilitarian moral principles, according to which the goodness or evil of an action is determined solely by its consequences (Mill, 1861; Bentham, 1789). In other words, if somehow you can save 10 lives by sacrificing one person, then it is justified to sacrifice that person. This view of morality, however, is at odds with the Kantian deontological view, according to which some actions can never be justified by their consequences; they are absolutely forbidden (Kant, 1787). In other words, it is always wrong to sacrifice an innocent person even if additional lives are saved as a result.

In this current study we address a question raised in the literature: is there a connection between utilitarian and anti-social or immoral choice? Additionally, do moral choices obey the law of demand? To address these issues we ran a within-subjects laboratory experiment to study choices in a classic moral dilemma, the well-known (hypothetical) Trolley problem, as well as choices in a consequential (i.e., real payoffs) money-burning experiment.

The Trolley dilemma has captivated moral philosophers for decades (Foot, 1967; Thomson, 1985; Spranca et al., 1991; Petrinovich et al., 1993). The dilemma describes a runaway trolley that, unless an action is taken, will run over several individuals on a track who are unable to escape. Action typically results in the death of a different individual but research shows upwards of 90% of individuals are willing to endorse the sacrifice of one to save (typically) five others (see Navarrete et al., 2012 and references therein). Various versions of the problem exist (see Shallow et al., 2011), but we focus on perhaps two of the most classic scenarios. The first assumes a runaway trolley will kill several individuals on a main set of tracks unless one pulls a lever to divert the trolley onto a side track where it will kill anyone who may be on the side track. Such a decision scenario is considered an “indirect” (or impersonal) moral choice in the sense that pulling the lever to save lives indirectly but intentionally results in the death of those on the side track. A second version is a “direct” (or personal) moral choice scenario where one must instead push an individual onto the main track (and that person will die) in order to save those others on the main track (Thomson, 1985).

The Trolley dilemma has come under fire for its lack of realism, low external validity, sensitivity to varied contextual details, inability to truly instruct us about utilitarianism, and failure to evoke psychological processes similar to other moral dilemmas (Rai and Holyoak, 2010; Bauman et al., 2014; Kahane, 2015). Nevertheless, others have found it useful for studying various components of moral reasoning (e.g., Cushman et al., 2006; Greene et al., 2001; Greene et al., 2009), such as the identification of behavioural norms or highlighting that certain moral dilemmas preferentially engage emotional centers in a way that may be important in predicting choice (e.g., Greene et al., 2001; Navarrete et al., 2012). Still others have noted how the Trolley dilemma can highlight the difference between acts of omission versus commission (Spranca et al., 1991; Cox et al., 2017), which is a relevant distinction in courts of law. And, while past criticism of the Trolley dilemma may have seemed justified due to the unrealistic nature of the decision it presents, the relevance of the Trolley dilemma is at a higher level than perhaps ever before with the recent rise in ethical concerns surrounding self-driving vehicles. Bonnefon et al. (2016) highlight how the moral dilemma relates to the social dilemma of Autonomous Vehicle (AV) adoption, whereby most survey respondents agreed that an AV should be programmed to sacrifice its passenger(s) if more pedestrians are saved as a result, but these same individuals thought it much less appropriate to program the AV as such if one's own life were at stake. Another recent study (Awad et al., 2018) documented how global views on Trolley dilemmas vary by culture, but summary statistics from large sample studies are focused on mean tendencies as opposed to examining “outlier” response patterns that may be informative regarding antisocial tendencies.

In economics, studies focusing on the antisocial dimension of behavior include the seminal studies by Zizzo and Oswald (2001) and Zizzo (2004), whose results show that many subjects are willing to incur a real cost in order to reduce other's payoffs—“money burning”. Money burning may be explained by inequality aversion (Zizzo and Oswald, 2001), but it may also result from a more antisocial pleasure of being nasty (Abbink and Sadrieh, 2009; Abbink and Herrmann, 2011).<sup>4</sup> Of course, Becker's (1968) seminal work on the economics of crime was highly influential and focused attention on the

<sup>2</sup> For instance, while there exists strong evidence that workers do not hesitate to engage in unethical activities in contests, it remains difficult to clearly disentangle whether sabotage activities are driven by pure anti-social tendencies or by monetary benefits associated with an increase in chance of winning the context by reducing the output of the opponent (e.g. Lazear, 1989).

<sup>3</sup> The atomic bombs dropped, resulted in the deaths of about 250,000 Japanese (Holt, 1995). The alternative was to launch an invasion. Truman claimed in his memoirs that this would have cost another half a million American lives.

<sup>4</sup> Individual characteristics may be yet another factor that explains money burning decisions. For instance, some previous studies have shown that high basal testosterone is associated with an increased threshold for conflict (see Carney and Mason, 2010, and references therein).

cost-benefit calculus of many decisions in the moral domain. Other relevant work relates to [Fehr and Gächter's \(2000\)](#) seminal paper examining peer punishment in group contribution environments, which could represent an environment where antisocial punishment is exhibited. These same authors noted the potential for peer-punishment to be prosocial as opposed to antisocial given certain conditions are met ([Fehr and Gächter, 2002](#)).

Our goal is to contribute to the literature in the following ways. First, we exploit a within-subjects strategy method design to examine the ability of moral identifiers, uniquely derived from Trolley dilemma decisions, to predict consequential choices in the money burning game. The predictive validity of ethical dilemma responses has been of interest in the recent literature, though not without debate. A recent study argues that hypothetical ethical dilemmas are *not* useful for predicting behavior in real dilemmas ([Bostyn et al., 2018](#)), though their study differs from ours on critical dimensions.<sup>5</sup> Other studies have already suggested a connection between antisocial personality types and a willingness to make morally difficult utilitarian choices ([Koenigs et al., 2007](#); [Bartels and Pizarro, 2011](#); [Gao and Tang, 2013](#); [Bracht and Zylbersztejn, 2018](#)),<sup>6</sup> though not all studies have supported this conclusion (e.g., see [Cima et al., 2010](#)). Those studies that do suggest the antisocial-utilitarian link, however, suffer from a key confound. Specifically, in their studies it is *always* utilitarian to sacrifice the life because more would be saved. Thus, existing studies cannot separate utilitarian behavior from less savory preferences. In contrast, our Trolley dilemmas help solve this identification problem and allow us to construct relatively unambiguous moral identifiers.<sup>7</sup>

Secondly, our paper will also contribute to the literature by investigating the extent to which costly money burning decisions and Trolley choices obey the law of demand. Responses to ethical dilemmas surrounding the adoption of autonomous vehicle technologies, which bear resemblance to the Trolley dilemma, were recently shown to be sensitive to the relative number of lives saved in the scenario ([Bonnefon et al., 2016](#)). Within the context of demand for costly punishment, [Nikiforakis and Normann \(2008\)](#) showed that voluntary contributions to provide a public good increase monotonically in punishment effectiveness, and [Anderson and Putterman \(2006\)](#) found that the price of punishment is a significant determinant of punishment demand. These previous studies suggest that even the moral domain of choice should obey the law of demand. To our knowledge, no previous study has attempted to investigate the role played by relative cost in the context of money burning decisions. Our set of Trolley dilemmas allows us to explore efficiency (utilitarian outcomes), which implies that the dominant concern should be minimizing the number of lives lost, even when we vary the price of inefficiency.

Finally, our analysis will distinguish between the different types of money burning as well as the different types of immorality we can infer from Trolley choices. We appreciate that not all money burning should be considered immoral or antisocial (e.g., inequality aversion would not be considered antisocial), and immoral choices may involve acts of omission or commission. Together these distinctions yield a more rich set of outcome and morality variables to explore in our data.

To preview our main findings, we find that outcomes in the Trolley dilemma are both consistent with previously known results but also make new contributions to the literature in an important way. Specifically, our data indicate that the relative cost of the ethical decision matters, as should be expected. Regarding the determinants of money burning, we find evidence that inequality version is present, but nastiness is also observed because some individuals burn a counterpart's money even when already at a payoff advantage. We also report that utilitarian behavior in the Trolley dilemma is *not* linked to antisocial money burning, which contrasts with previous conclusions in the literature suggesting that antisocial types are more utilitarian ([Koenigs et al., 2007](#); [Bartels and Pizarro, 2011](#); [Gao and Tang, 2013](#); [Bracht and Zylbersztejn, 2018](#)). However, those making ethically dubious choices in the Trolley problem are also more antisocial in their money burning choices. That is, we find that "Trolley immorality" can statistically predict inefficient antisocial (even nasty) money burning choices.

<sup>5</sup> [Bostyn et al \(2018\)](#) examine whether Trolley dilemma responses predict one's propensity to deliver electric shocks to mice in dilemmas with similar but nonfatal scenarios. In addition to the fact that their hypothetical judgment scenarios involved humans and not mice (which were the focus of their « real-life » ethical dilemmas), their study also involved deception. Specifically, the mice were not actually sacrificed as per the participants' decisions and so a de-briefing was also used in their study.

<sup>6</sup> Using self-report measures of antisocial personality tendencies ([Bartels and Pizarro, 2011](#); [Gao and Tang, 2013](#)) or patients with brain damage in regions important to emotion generation ([Koenigs et al, 2007](#)), these studies find tendencies towards increased utilitarianism in individuals with antisocial personality traits. Another recent study ([Bracht and Zylbersztejn, 2018](#)) is quite related to ours in that it also examines ethical choice in hypothetical dilemmas as well as in a consequential money transfer game. The differences in our study are notable, however. First, we do not pool data across *indirect* versus *direct* moral choices as they do, which is important given we identify a highly significant ( $p < .01$ ) impact of this factor on one's willingness to take action (we also show that one of their key results is qualified in our findings by conditioning on the direct versus indirect nature of the dilemma). Secondly, both our hypothetical and consequential choice experiments vary the relative efficiency or cost of one's action, thus allowing a more thorough examination of ethical and antisocial choice. Finally, we use a morality measure derived from the Trolley dilemma to predict behavior in the consequential money burning game, while [Bracht and Zylbersztejn \(2018\)](#) examine the reverse causation. While of potential interest, we find the causation of hypothetical-to-consequential choice more valuable in terms of implications and use as a potential screening or identification mechanism (e.g., job application/interview screener). Additionally, it is still the case that the dilemmas used in their study suffer from the key confound whereby utilitarian preferences cannot be separated from certain types of immoral preferences.

<sup>7</sup> It is therefore important to note that many moral dilemmas confound the utilitarian choice from the choice one might make for non-utilitarian reasons. For example in the typical Trolley dilemma, it is utilitarian to pull the switch or push the individual, and yet one may be willing to act not because more lives are saved than lost, but rather because one prefers or perversely enjoys being responsible for someone's death.

**Table 1**  
Trolley Dilemmas.

Trolley Dilemma #	INDIRECT				DIRECT			
	Are you willing to pull a lever to divert the trolley to a different track to save X people, where Y on that side track will die.				Are you willing to kill Y people by pushing them onto the track to save X people?			
	Total people killed	Total people saved	Yes	No	Total people killed	Total people saved	Yes	No
1	Y=6	X=6	Yes	No	Y=6	X=6	Yes	No
2	Y=5	X=6	Yes	No	Y=5	X=6	Yes	No
3	Y=4	X=6	Yes	No	Y=4	X=6	Yes	No
4	Y=3	X=6	Yes	No	Y=3	X=6	Yes	No
5	Y=2	X=6	Yes	No	Y=2	X=6	Yes	No
6	Y=1	X=6	Yes	No	Y=1	X=6	Yes	No
7	Y=1	X=5	Yes	No	Y=1	X=5	Yes	No
8	Y=1	X=4	Yes	No	Y=1	X=4	Yes	No
9	Y=1	X=3	Yes	No	Y=1	X=3	Yes	No
10	Y=1	X=2	Yes	No	Y=1	X=2	Yes	No
11	Y=1	X=1	Yes	No	Y=1	X=1	Yes	No
12	Y=0	X=6	Yes	No	Y=0	X=6	Yes	No

Note: Trolley dilemmas numbered here for discussion in the text (dilemmas were not numbered for subjects).

## 2. Experimental design

### 2.1. Overview

Both the Trolley and the money burning experiments were administered in strategy method format, where decisions were elicited on multiple decisions prior to a randomized draw of one (in the incentivized money burning task) for real payoff. Table 1 describes the menu of dilemmas administered in our version of the Trolley dilemma. Importantly, we highlight that our choice menu allows us to examine how the likelihood of taking action responds to the number of people saved (X) relative to killed (Y). We are also able to examine preferences for inaction over action when the number of lives lost would be unaffected (i.e.,  $X=Y$  dilemmas). And finally, we can examine how one's likelihood to take action differs if action is indirect (i.e., pull a lever to divert the runaway trolley) versus more direct (i.e., push an individual(s) onto the track to stop the runaway trolley), which we call *INDIRECT* versus *DIRECT* decision scenarios.

In what follows, we scored immorality as derived from the Trolley dilemma choices as follows: *Immoral Omission* is an indicator variable equal to one if a subject chose to *not* take action in the *DIRECT* and *INDIRECT*  $(X,Y)=(6,0)$  scenarios, where action would save 6 individuals without any lives being lost as a result. Another dichotomous variable, *Immoral Commission*, equals one if the subject chose action in both the  $(X,Y)=(6,6)$  and  $(1,1)$  scenarios of both the *INDIRECT* and *DIRECT* choice dilemmas. In the case of *Immoral Commission*, the subject prefers to be responsible (via action) for a given number of deaths rather than passively allow that same number of deaths to occur. We created a final variable by taking a subject's average propensity to act in the remaining scenarios not used in the construction of the *Immoral Omission* or *Immoral Commission* variables. Such a variable, *Action Propensity*, represents one's willingness to take action, though it also describes utilitarian preferences in our dilemmas.

For the money burning game, a key treatment variable is whether only one in the pair (the “decider”) or both individuals could burn money.<sup>8</sup> Specifically, in the *Bilateral Burn* treatment, both players could mutually and simultaneously destroy a portion of each other's payoffs. That is, each of the two subjects in a randomly matched pair made money burning decisions and two random decisions were selected such that each subject was both a decider and passive recipient (i.e., potential money burn victim) in a consequential money burning choice. In *Unilateral Burn*, subjects were randomly assigned as decider or passive recipient *before* decision making, and only the deciders made decisions. We chose to run these two variants of the money burning to check whether individuals may be motivated by “pre-emptive retaliation (see Abbink and Sadrieh, 2009). Indeed, in the bilateral burn treatment, one may expect more destruction of wealth because subjects may expect their partner to destroy money. Consequently, one may “respond” to this expectation by also burning money.

After decisions were made in all 9 money burning scenarios, deciders and recipients were randomly matched and one scenario was selected at random to determine the payoff of both players in the money burning game.<sup>9</sup> This process was

<sup>8</sup> We also varied the ordering of the Money Burning  $(x,y)$  pairs in the menu received (presenting the decision maker's endowment in *Increasing*, *Decreasing*, or *Random* order. Each subject saw only one ordering). We did not have a formal hypothesis regarding the ordering of the money burning scenarios, and later analysis documents that the varied ordering does not significantly impact outcomes in the task. We considered variation in the order more exploratory. Of course, there is no theoretical reason to believe that the ordering should matter, but this possibility has been investigated on the more well-known Holt and Laury (2002) risky choice lottery menu (see Bruner, 2009).

<sup>9</sup> Our design made use of the strategy method, as opposed to direct elicitation method, in order to generate multiple observations from each subject in each decision experiment (other than the passive recipients in the money burning game, which were randomly selected *prior* to decision making in that game). Brandts and Charness (2011) survey experimental results comparing strategy method versus direct elicitation and conclude that the strategy method for response elicitation, in general, provides a conservative estimate of what choice would be using direct response elicitation—in our case, money burning choices may therefore be a conservative estimate of outcomes one would find using direct elicitation of just a single response in a single scenario.

**Table 2**

Money burning choice tasks (*Increasing* treatment).  
Subjects chose the start or end distribution for each of the 9 tasks.

Task #	Start Distribution	Damage	Burning costs	End Distribution
A1	(50, 250)	50	10	(40, 200)
A2	(50, 200)	50	10	(40, 150)
A3	(50, 150)	50	10	(40, 100)
A4	(50, 100)	50	10	(40, 50)
A5	(50, 50)	50	10	(40, 0)
A6	(100, 50)	50	10	(90, 0)
A7	(150, 50)	50	10	(140, 0)
A8	(200, 50)	50	10	(190, 0)
A9	(250, 50)	50	10	(240, 0)

**Table 3**

Summary of the money burning treatments.

Session	Participants	Treatment Description
1	18	<i>Increasing</i> —Unilateral burn
2	16	<i>Increasing</i> —Bilateral burn
3	18	<i>Increasing</i> —Bilateral burn
4	18	<i>Decreasing</i> —Unilateral burn
5	16	<i>Decreasing</i> —Bilateral burn
6	14	<i>Decreasing</i> —Bilateral burn
7	18	<i>Random</i> —Unilateral burn
8	16	<i>Random</i> —Bilateral burn
9	16	<i>Random</i> —Bilateral burn
Total	150	( $n = 96$ Bilateral Burn, $n = 54$ Unilateral burn)

common knowledge. Table 2 shows the money burning task decisions faced by the subjects (with one's own payoff listed first and increasing as one goes down the Table of scenarios).

## 2.2. Experimental procedures

The experiment was computerized and administered using the Z-tree platform (Fischbacher, 2007). We recruited 150 subjects at the University of Rennes 1 (France), each subject participated in only session, and none had participated in a similar economic experiment. A total of 9 sessions were conducted (with 14 to 18 subjects per session), where in each session subjects were administered the Money Burning game followed by the Trolley dilemma.<sup>10</sup> Table 3 contains summary information about number of participants in each treatment of the Money Burning game, which is identified by the order of presentation of the scenarios (see footnote 4) and whether the *Unilateral* or *Bilateral Burn* treatment. Importantly, subjects were given the choice to opt out of the Trolley Dilemma for whatever reason. A total of 12 subjects (8%) chose to opt out of the Trolley dilemma task, and we use this “opt-out” in the analysis of money burning choices below.

A session lasted approximately one hour (this includes the time spent to read the instructions). At the end of the experiment, one task was randomly selected for each pair of randomly matched subjects (and random role assignments, in the case of *Unilateral Burn* treatments).<sup>11</sup> Payments were made anonymously at the end of the session and the average earnings were 25.52 Euros per subject.

## 3. Behavioural predictions and theoretical foundations

Though we describe a set of predictions based on behavioural considerations, it is important to highlight that one can identify theoretical underpinnings of these behavioural predictions. Consider, for example a model based on Figueres et al. (2013), that considers intrinsic moral obligations within the utility function (see also, Nyborg 2000, Brekke et al., 2003, Dickinson et al., 2018). The idea is that a utility function may include a moral obligation grounded in a Kantian categorical imperative (Laffont, 1975; Harsanyi, 1980).<sup>12</sup>

<sup>10</sup> This ordering is important in order to limit the potential for morality priming prior to the Money Burning game (i.e., we considered it much less a concern that playing the Money burning game would prime participants prior to the Trolley Dilemma given its neutral context compared to the sensation and obviously morality context of the Trolley Dilemma).

<sup>11</sup> This common payment procedure helps eliminate the potential for portfolio effects in the strategy method choice set to influence choice across the set. Only one choice will count for payoff and, with each choice being equally likely, subjects should treat each choice as the one that may dictate their payoff.

<sup>12</sup> The inclusion of moral values into motivations is part of the early history of economic thinking and dates back at least as far as Smith (1759).

Assume, for example, that one's action,  $a$ , generates both benefits,  $b$ , and costs,  $c$ . Further assume a function  $v(a - \hat{a})$  where  $\hat{a}$  describes one's moral imperative or obligation, and a deviation from this moral standard of action,  $a$ , generates disutility. Then, one's utility function can be described by:

$$U = b(a) - c(a) - v(a - \hat{a}) \quad (1)$$

Here, we assume  $b' > 0$ ,  $c' > 0$ ,  $b'' < 0$ ,  $c'' > 0$ , such that utility benefits and costs are increasing in the action, and benefits increase at a decreasing rate while costs increase at an increasing rate. The disutility of deviations from one's moral ideal are captured by assuming  $v' > 0$  if  $a > \hat{a}$ ,  $v' < 0$  if  $a < \hat{a}$ , and  $v' = 0$  if  $a = \hat{a}$ . That is, moral disutility decreases by moving towards one's moral obligation from either direction. We also assume that  $v'' > 0$  such that marginal disutility increases at an increasing rate as one's action gets further from the moral obligation. Note that an "action" here is quite general (i.e., higher morality could imply a higher or lower level of the action). All that matters is that actions generate utility costs and benefits, and there is disutility in moving away from one's moral obligation, whatever that may be.

We show in [Appendix B](#) that the first order condition from the utility maximization problem can be used to derive the following (intuitive) comparative static result:

$$\frac{\partial a^*}{\partial \hat{a}} > 0$$

In other words, one's optimal action moves in accordance with one's moral obligation. This implies that differences in moral choice across individuals in hypothetical environments are either the result of cost and/or benefit differences due to the action, or they are the result of differences in moral obligations across individuals.<sup>13</sup>

### 3.1. Trolley problem

First consider the theoretical predictions in the trolley problem. In absence of moral considerations, purely selfish decision makers (the *homo economicus*) should be indifferent between action and inaction since their material payoff remains unaffected in both cases. In sharp contrast, a *utilitarian* should always take an action when the number of lives saved is higher than the number of lives sacrificed since it maximizes the aggregate welfare ([Bentham, 1789](#); [Mill, 1861](#)). Consequently assuming that agents are utilitarian, we posit that a decreased relative cost of action should increase action likelihood. In other words, we predict a downward sloping demand curve for lives saved in this moral dilemma. Let us now consider that agents have morality concerns. The introduction of morality concerns into the utility function may prevent agents from acting despite the existence of a net gain in terms of lives saved. If the moral cost of taking an action in the Trolley problem is sufficiently high, individuals should never act, irrespective of the material aggregate net benefit from doing so. This is summarized in hypothesis [H1](#):

**H1 (Trolley):** **a)** According to utilitarian principles, the likelihood of action will increase in the relative number of lives saved. **b)** Individuals with sufficiently high moral concerns will never take action in the Trolley problem.

**Proof of H1:** see [Appendix B](#).

Our second assumption concerns the specific dilemma in the Trolley problem where lives lost are unaffected ( $X=Y$  dilemmas) or when action is costless (any  $(X,0)$  dilemma). In our set of Trolley dilemma choices we can then focus on  $(X,Y)$  pairs (6,6) and (1,1), where an equal number of individuals would perish whether or not action is taken. In these cases, we hypothesize a lesser likelihood to act given a preference to not be responsible (via action) for the deaths. To take action in such cases could be considered an immoral act of *commission*. Also of interest would be the (0,6) Trolley dilemma, where action costs no lives. In such a dilemma, to *not* take action would be consider an immoral act of *omission*. Both individuals and courts of law consider an act of omission to be a lesser "sin" than an act of commission that results in similar consequences (see [Cox et al., 2017](#)). This principle with respect to the Trolley dilemma has been labelled the "action principle". This is stated in hypothesis [H2](#):

**H2 (Trolley):** When lives lost are unaffected ( $X=Y$  dilemmas), inaction is preferred over action (moral omission). Also, action is preferred over inaction when action is costless (any  $(X,0)$  dilemma) (moral commission).

Rejection of [H2](#) implies acts of *Immoral Omission* or *Commission*.

**Proof of H2:** see [Appendix B](#)

Our last assumption regarding the Trolley problem concerns the role of framing. The literature identifies two clear predictions we can make regarding outcomes in the Trolley dilemma. First, a widely reported result is that individuals are more willing to take action and save lives in the *INDIRECT* frame where a level is pulled, as compared to the *DIRECT* frame where

<sup>13</sup> Alternatively, differences in moral choices may also result from mistakes in maximization (i.e., error, or perhaps lack of motivation due to hypothetical nature of choice). However, if immorality as identified in our hypothetical Trolley environment can predict other unethical or antisocial choice, it implies the hypothetical choices are not mistakes but rather reflect fundamental differences in individuals' morality views that may help predict choices in other moral domains that are consequential.

an individual is pushed onto the track, holding constant the relative number of lives saved. This is related to the distinction between personal and impersonal moral dilemmas (Greene et al., 2001). Thus, our third hypothesis stems from this “contact principle” (Cushman et al., 2006). This hypothesis implies that for each pair (X,Y) of lives (saved, lost), we predict an individual is more likely to take action in the *INDIRECT* frame.

**H3** (Trolley): For each (X,Y) dilemma, action is more likely in the *INDIRECT* Frame

**Proof of H3:** see [Appendix B](#)

### 3.2. Money burning

Consider now the theoretical predictions of the money burning game. Purely selfish individuals should never burn money since there are no material benefit associated with money burning decisions. The same predictions apply for *utilitarian* agents who would never choose to reduce total welfare. Thus, under the assumption of either pure selfishness or *utilitarianism*, there should be no money burning. The same predictions should apply for the *homo moralis*, i.e. agents with sufficiently high moral concerns. Only individuals with *nasty* preferences may be incited to burn money. This is stated in assumption

**H4**

**H4** (Money Burning): under the assumption of either pure selfishness or *utilitarianism*, there should be no money burning. Only individuals with *nasty* preferences will burn money.

**Proof of H4:** see [Appendix B](#)

In addition to pure nastiness, individuals' decisions in the money burning game may be also motivated by inequality aversion concern (Zizzo and Oswald, 2001; Abbink and Sadrieh, 2009; Abbink and Herrmann, 2011). According to inequality models utility depends not only on one's own payoff but also on the equality of the income distribution (see Fehr and Schmidt, 1999).<sup>14</sup> In our framework, if disadvantageous inequality aversion matters, one should therefore observe money burning when  $x < y$ , while money should not be burnt in cases of advantageous inequality (i.e.,  $x \geq y$ ).

**H5** (Money Burning): Strongly inequality averse people should burn money in case of disadvantageous inequality only (i.e.,  $x < y$ ).

**Proof of H5:** see [Appendix B](#).

Let's now focus our attention on the behavioural changes induced by the differences between the *Unilateral burn* and *Bilateral burn* treatments. Because decisions in the *Bilateral Burn* treatment are impacted by any (unmeasured) expectations of others' money burning choices, the pure effect unconfounded by expectations is measured from the comparison with the *Unilateral Burn* treatment. If individuals burn others' money and it is common knowledge that money burning is *Bilateral*, then another motivation for money burning is anticipatory negative reciprocity. This type of “pre-emptive retaliation” relies on the fact that in the simultaneous choice *Bilateral* treatments one may burn the counterpart's money on the expectation that the counterpart may burn some of one's payoff (see Abbink and Sadrieh, 2009). Because of pre-emptive money burning, we should therefore expect more money burning in the *Bilateral* treatment than in the *Unilateral Burn* treatment, *ceteris paribus*. Thus, we have the following money burning hypothesis **H6**:

**H6** (Money Burning): Money burning will be higher in *Bilateral Burn* treatment compared to the *Unilateral Burn* treatment.

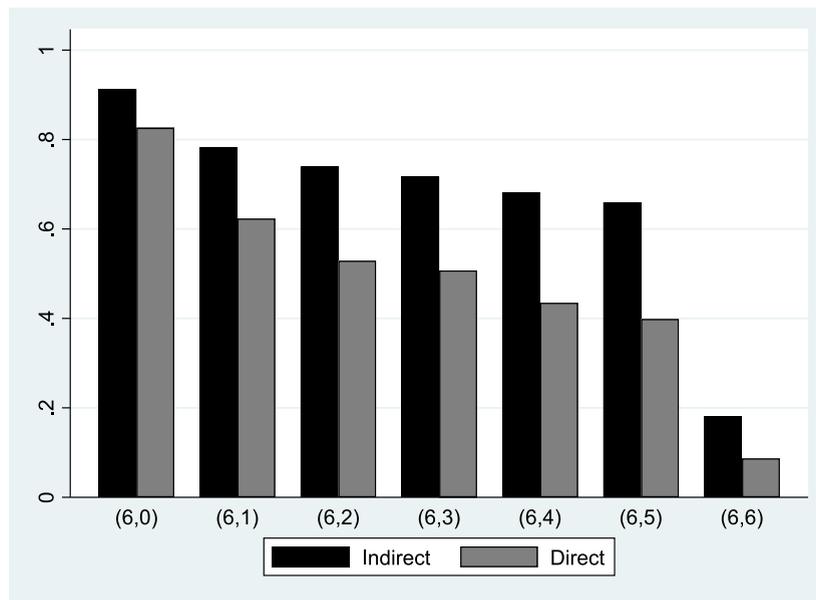
**Proof of H6:** see [Appendix B](#)

Just as in the case of the Trolley dilemma, we expect that antisocial tendencies to burn resources of others will nevertheless respond to the price of doing so. Previous studies have shown that punishment decisions in a VCM context obey the law of demand. (Nikiforakis and Normann, 2008; Anderson and Putterman, 2006). Based on these papers' findings one may reasonably conjecture that money burning decisions also obey the law of demand and, though the cost of burning is fixed in our design, the amount burnt varies. This implies that the cost of burning money (namely, the cost *relative* to one's payoff) in the chosen payoff distribution varies and we can expect an increase in money burning when the relative cost of burning money is low. This leads to **H7**.

**H7** (Money Burning): Burning money will be negatively related to its relative cost.

Finally, an important contribution we offer in the paper is to consider moral descriptors of one's choices in the Trolley dilemma as an explanatory variable regarding one's choice to burn money. Immoral acts of commission and omission are defined in **H2** based on the subset of Trolley dilemmas that did not present confounded explanations of one's choice. Someone who takes action in the (X,Y) Trolley dilemmas not implicated in **H2** can be said to have a higher *Action Propensity* (or

<sup>14</sup> Indeed, a very appealing hypothesis about distributional preference is inequality aversion (see Loewenstein et al. 1989; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002). These approaches assume utility depends not only on one's own payoff but also on the equality of the income distribution.



**Fig. 1.** Frequencies of taking action in the Trolley dilemma by treatment (number saved unchanged).  
Notes: (X,Y) dilemmas represent number saved (X) and number sacrificed (Y).

is more utilitarian). The morality of those with higher *Action Propensity* is difficult to assess given that one may be willing to sacrifice one or more lives to save others for more than one reason. Such reasons may include both ethically dubious reasons (i.e., I prefer to push someone to save others) as well as utilitarian reasons (e.g., I will do whatever leads to the most lives saved (least lives lost)). However, our morality variable constructs are intended to separate utilitarian actors from the immoral actors. For this reason, clean moral descriptors of immorality for our final hypothesis focus on metrics derived from a subset of the Trolley dilemmas, and comparison with results in the existing literature linking utilitarianism to antisocial choice can be made by focusing on the *Action Propensity* impacts.

**H8** (Money Burning): Moral descriptors derived from the Trolley dilemmas—( $X=Y$ ) and ( $X,0$ ) dilemmas—will predict increased money burning.

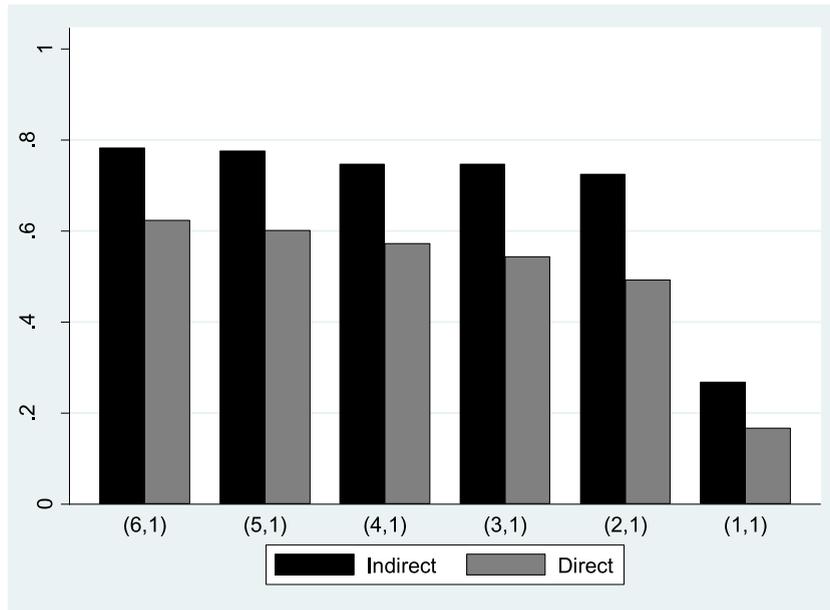
## 4. Results

### 4.1. Trolley results

We first share results from the Trolley Dilemma that we used to construct predictor variables used in the analysis of the Money Burning game data. We start by showing summary data from the subjects who made Trolley dilemma choices in Figs. 1 and 2 (9 Trolley dilemma choices per subject). Of the 150 participants in our experiment,  $n=12$  subjects opted out of the Trolley dilemma, leaving us with  $n=138$  Trolley subject decision makers (we code these “Trolley opt-out” subjects for later use as a regressor in the money burning estimations in the next section).<sup>15</sup> Fig. 1 shows the proportion of choices in each treatment (*Direct* and *Indirect* dilemmas) for the subset of dilemmas that hold constant the number of lives saved. Left to right on the horizontal axis shows dilemmas that increase the number of individuals sacrificed for a constant  $X=6$  individuals saved. Two things stand out in Fig. 1: the proportion of individuals who take action decreases as the relative cost,  $Y/X$ , increases; more surprisingly, greater than 20% of subjects did *not* choose to take action in the (6,0) dilemma where 6 individuals could be saved at zero cost, and some chose (indirect) action in the (6,6) dilemma where the same number of individuals would perish even if nothing were done. Both represent instances of what we call “Trolley immorality”.

Fig. 2 organizes the remaining subset of Trolley choices to hold constant the number of individuals who perish at  $Y=1$  and the number of lives saved decreases going from left to right in the figure. We again see that action in the Trolley dilemma is responsive to the relative cost (or effectiveness) of the action—subjects are less likely to take action when the relative cost,  $Y/X$ , increases (or, as the relative benefit  $X/Y$  decreases). In Fig. 2, we also see that a nonzero number of

<sup>15</sup> We conducted a probit estimation of the determinants of the decision to opt out of the Trolley dilemma. Though few subjects opted out, we found one variable, “happiness” (self-reported current level of happiness in life) was a marginally significant determinant of the opt-out choice ( $p < .10$ ). Specifically, those self-reporting higher levels of life happiness were marginally more likely to opt out of the Trolley dilemma.



**Fig. 2.** Frequencies of taking action in the Trolley dilemma by treatment (number killed unchanged).  
Notes: (X,Y) dilemmas represent number saved (X) and number sacrificed (Y).

subjects choose an immoral act of commission in Trolley dilemma (1,1) where action was chosen even though an individual would perish even with inaction.

As noted in Section 3 (Experimental Design), we elicit choices in the strategy method to maximize data generated per subject. Due to multiple decisions per subject, all models in Table 4 include standard errors clustering at the individual subject level. The model structure is a Probit estimation where the dependent variable is equal to one if that subject chooses to take “action” (i.e., pull the lever or push the individual(s) in that particular dilemma scenario). The different columns of Table 4 show estimations using different sets of independent variables. The first two columns use a dummy variable for each (X,Y) pair of lives saved (X) and sacrificed or killed (Y) compared to the omitted baseline scenario of (X,Y)=(6,0). Columns 3–5 replace the dummy variables with continuous variables measuring the number of lives sacrificed and saved.

The dummy variable identifying the *DIRECT* version of each Trolley dilemma has a consistently negative and significant coefficient estimate across all models, which supports Hypothesis H3. Individuals are significantly less likely to take action when it is a more personal moral dilemma (action would be direct) compared to impersonal (action would be indirect). Interestingly, this effect is somewhat muted for male subjects as seen by the significant and positive coefficient on *Male\*DIRECT* in model 5.<sup>16</sup> Because many of the dilemmas confound morality of choice with utilitarian actions, we next examine hypothesis H2 using only the subset of Trolley dilemmas (X,Y)=(6,6), (1,1), and (6,0). The comparison of coefficients in our Table 4 estimations are not a transparent way to assess whether a statistically significant number of subjects chose *action* in the (6,6) and (1,1) dilemmas, or *inaction* in the (6,0) dilemma. Rather, we can test the null hypothesis that the proportion of subjects choosing *action* in the (X=Y) dilemmas is equal to zero against the alternative that it is greater than zero. For the test of immoral action, we test the null hypothesis that the proportion of subjects choosing *action* in the (6,0) dilemma is equal to 100% against the alternative hypothesis that it is less than 100%. For the case of  $n = 138$  observations, the observed proportions in both the case of *DIRECT* and *INDIRECT* framing of the Trolley dilemmas lie outside of the 95% confidence interval. This evidence implies rejection of H2 in favor of the existence of immoral acts of omission and commission being greater than zero.<sup>17</sup>

Finally, we show support for Hypothesis H1 by using estimates in models (4) and (5) of Table 4. Here the marginal effect on # Lives Sacrificed (Y) holds constant the # Lives Saved (X), and vice versa. Thus, the negative and positive, respectively, effects of these variables on the likelihood of taking action confirm that Trolley choices respond to the relative number of

<sup>16</sup> This result is somewhat related to the gender result found in Bracht and Zylbersztejn (2018), who find males more likely to take action in a set of moral dilemmas. The study includes a variety of dilemmas in addition to a limited number of Trolley dilemmas, but they do not distinguish dilemmas in their set that involve a *direct* versus *indirect* action in the moral choice. As such, our result is an important qualification of what they report given our evidence suggests the gender effect may not be as general as they suggest.

<sup>17</sup> For the sample proportions tests, the Z statistic cannot be calculated for the boundary hypothesized proportions of 0% and 100%, and so we rather calculate our tests using null hypothesis proportions of 1% and 99%, respectively. Our conclusions remain intact even if allowing for a 5% “error” in decision making (at the  $p < .10$  level for the (6,6) *DIRECT* and (6,0) *INDIRECT* dilemmas, but at the  $p < .01$  level in all other cases). That is, if assuming that a small percentage of subject may make mistaken choices in our sample, our conclusions regarding the H2 result are largely unchanged.

**Table 4**  
Probability of action (pull level or push person) in trolley dilemma.

Marginal Effects Reported (robust st errors in parenthesis)						
Independent variable	(1)	(2)	(3)	(4)	(5)	(6)
	All	All	All	All	All	No-Switch <sup>^</sup>
<i>DIRECT</i> Action	−0.2061*** (0.0314)	−0.2083*** (0.0317)	−0.1917*** (0.0302)	−0.1997*** (0.0305)	−0.2974*** (0.0502)	−3054*** (0.0507)
<i>Male</i> * <i>DIRECT</i>	–	–	–	–	0.1863*** (0.0570)	0.1932*** (0.0576)
(X,Y)=(6,0)	Reference	Reference	–	–	–	–
(X,Y)=(6,6)	−0.6207*** (0.0263)	−0.6243*** (0.0268)	–	–	–	–
(X,Y)=(6,5)	−0.3996*** (0.0422)	−0.4026*** (0.0427)	–	–	–	–
(X,Y)=(6,4)	−0.3765** (0.0448)	−0.3794*** (0.0454)	–	–	–	–
(X,Y)=(6,3)	−0.3298*** (0.0488)	−0.3324*** (0.0495)	–	–	–	–
(X,Y)=(6,2)	−0.3095*** (0.0495)	−0.3120*** (0.0503)	–	–	–	–
(X,Y)=(6,1)	−0.2391*** (0.0529)	−0.2403*** (0.0542)	–	–	–	–
(X,Y)=(5,1)	−0.2547*** (0.0508)	−0.2570*** (0.0517)	–	–	–	–
(X,Y)=(4,1)	−0.2851*** (0.0468)	−0.2877*** (0.0476)	–	–	–	–
(X,Y)=(3,1)	−0.2991*** (0.0472)	−0.3019*** (0.0478)	–	–	–	–
(X,Y)=(2,1)	−0.3329*** (0.0443)	−0.3359*** (0.0449)	–	–	–	–
(X,Y)=(1,1)	−0.5839*** (0.0266)	−0.5881*** (0.0268)	–	–	–	–
# Lives Sacrificed (Y)	–	–	−0.1102*** (0.0073)	−0.1117*** (0.0072)	−0.1126*** (0.0073)	−0.1161*** (0.0070)
# Lives Saved (X)	–	–	0.0873*** (0.0062)	0.0885*** (0.0061)	0.0892*** (0.0062)	0.0909*** (0.0060)
Religion ∈ [1 ,10] (10=very important)	–	−0.0026 (0.0118)	–	−0.0022 (0.0113)	−0.0023 (0.0114)	−0.0027 (0.011)
Happiness ∈ [1 ,10] (10=highest current life happiness)	–	0.0205 (0.0204)	–	0.0196 (0.0196)	0.0198 (0.0197)	0.0200 (0.0198)
Age	–	0.0289 (0.0187)	–	0.0271 (0.0178)	0.0270 (0.0178)	0.0278 (0.0181)
Male (=1)	–	0.1020* (0.0599)	–	0.0979* (0.0576)	−0.0013 (0.0657)	−0.0025 (0.0668)
Observations	3312	3312	3312	3312	3312	3264
#Clusters	138	138	138	138	138	136 <sup>^</sup>
Log likelihood	−1921.8550	−1889.6102	−1998.7432	−1968.3391	−1954.1628	−1911.9069

Notes: \*.10, \*\*.05, \*\*\*.001 for the 2-tailed test. Standard errors clustered at the individual subject level.

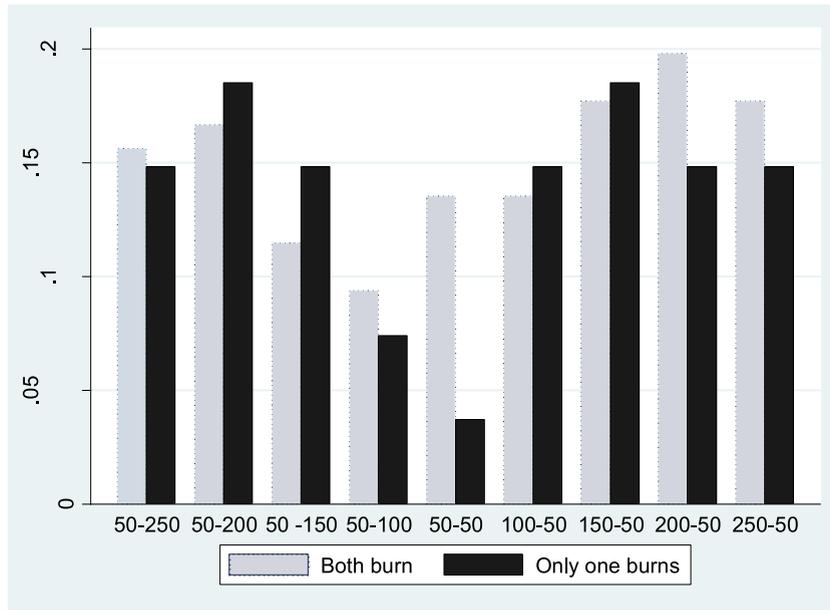
Total observations reflect  $n = 138$  subjects who opted to complete the Trolley dilemma task. Each of the 138 made 12 *Direct* and 12 *Indirect* Trolley dilemma choices.

<sup>^</sup> Reduced by subject who inconsistently switched choices in the Trolley Dilemma.

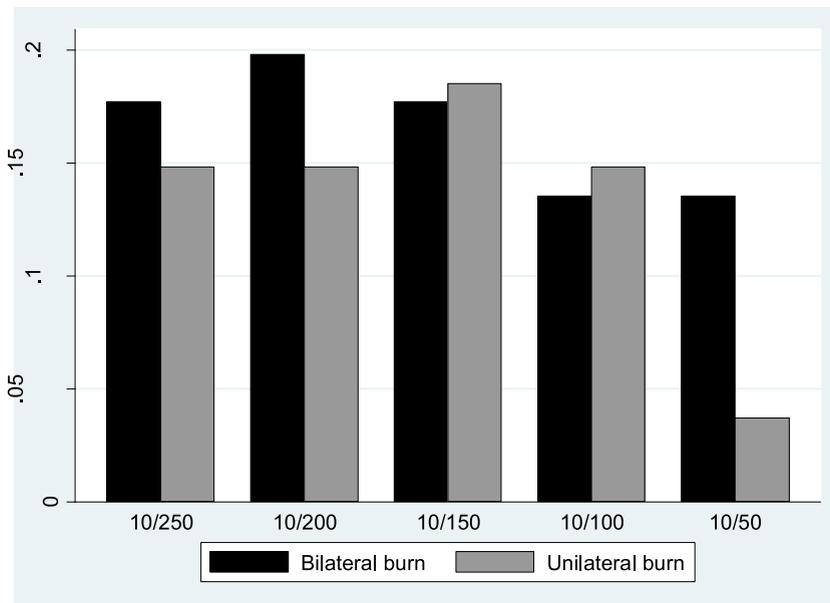
lives saved to lost, which supports Hypothesis 3. In short, action in the Trolley dilemma responds to incentives and displays a downward sloping demand curve for lives saved. Model (6) re-estimates model (5) for the subset of subject who do not display multiple switches in choice across the choice list to highlight that our results are not an artefact of irrational switching behavior in our task design. Nevertheless, a nontrivial number of subjects make choices that can be classified as immoral acts of commission or omission in our unique set of Trolley dilemmas. Having established the results from our Trolley Dilemma, shown them consistent with the extant literature, and also documented our morality metrics as revealing, we next turn to the results from the Money Burning game.

#### 4.2. Money burning results

Summary results from the Money Burning game are shown in Figs. 3 and 4, and in Table 5. Figs. 3 and 4 summarize the frequency of money burning choices for the different (x,y) allocation pairs. Fig. 3 shows money burning choices in each possible scenario, for both instances of *Unilateral Burn* and *Bilateral Burn*. Fig. 4 highlights the apparent downward trend in



**Fig. 3.** Frequencies of money burning decision by treatment.  
Notes: Allocation x-y describes own payoff-recipient payoff.



**Fig. 4.** Money burning per relative cost.  
Notes: ratio along horizontal axis reflect the size of the burning cost, 10, relative to the decider's payoff level prior to burning the recipient's payoff. Left to right along the axis reflects an increasing cost of burning the recipient's payoff, relative to one's own payoff in the allocation.

money burning as the cost of burning money is larger relative to the recipient's budget—money burning also obeys the law of demand (H7). Table 5 shows the total number of instances (out of 9 scenarios) in which the subject burned money, on average (top row), along with summary information on the proportion of money burning choices for the different possible types of money burners. Depending on the relationship between the decider's payoff,  $x$ , and the passive recipient's payoff,  $y$ , one can consider decisions to burn money reflect disadvantageous inequality aversion (burning money when  $x < y$ ) or nasty preferences (burning money when  $x > y$ ). Others may never burn money (*Homo Economicus* or Utilitarian preferences), and some burn money in all 9 scenarios and reveal an unconditional desire to behave antisocially (i.e., destroy resources and reduce total welfare of the pair). Though we have only limited data from deciders in the *Unilateral Burn* treatments, the *Bilateral Burn* data in Table 5 reflect similar proportions of burn choices in both *Bilateral* and *Unilateral Burn* treatments. This

**Table 5**  
Descriptive statistics of money burning decisions.

	All	Unilateral Burn	Bilateral Burn
# Money Burning choices (out of 9)			
<i>Mean</i>	1.33	1.22	1.35
[standard deviation]	[2.11]	[1.82]	[2.19]
<b>Never burn</b>	<b>82</b>	<b>18</b>	<b>64</b>
<i>Homo Economicus or Utilitarian</i>	(66.67%)	(66.67%)	(66.67%)
<b>Burn only when income &lt; other's</b>	<b>6</b>	<b>2</b>	<b>4</b>
<i>(Disadvantageous inequality aversion)</i>	(4.87%)	(7.40%)	(4.17%)
<b>Burn only when income ≥ other's</b>	<b>19</b>	<b>4</b>	<b>15</b>
<i>(Pure nastiness)</i>	(15.45%)	(14.81%)	(15.62%)
<b>Always burn</b>	<b>2</b>	<b>0</b>	<b>2</b>
<i>(Unconditionally anti-social)</i>	(1.63%)	(0%)	(2.08%)
<b>Other</b>	<b>14</b>	<b>3</b>	<b>11</b>
	(11.38%)	(11.12%)	(11.46%)
<b>Total # Subjects</b>	<b>123</b>	<b>27</b>	<b>96</b>

Notes: # subjects in **bold**, % subjects in parenthesis ( ).

most likely indicates that *Bilateral Burn* choices are *not* driven primarily by expectations that others will burn money. We next examine more formal econometric tests of our money burning hypotheses.

Tables 6 and 7 show results from Probit estimations of the probability that someone makes the dichotomous choice to burn money and select the End Distribution over the Start Distribution in the Table 2 scenarios. Errors in both tables are clustered at the level of the individual subject, and we report marginal effects in the tables. The set of independent variables in Table 6 includes: controls for the presentation order of the  $(x,y)$  distributions in the Money Burning menu set (*Random, Increasing, or Decreasing*); an indicator variable for the scenarios where burning was *Unilateral* (*Bilateral* is the reference group)<sup>18</sup>; indicator variables capturing payoff equality/inequality in the different  $(x,y)$  payoff distributions; a variable measuring the relative cost of money burning compared to one's own payoff; a variable measuring the simple utilitarian preference to take action (*Action Propensity*); a set of subject-specific controls. Importantly, model (3) in Table 6 and the models in Table 7 include indicator variables to identify whether the subject committed *Immoral Commission* or *Immoral Omission* in the Trolley problem (6,6), (1,1), and (6,0) dilemmas. So, these two indicator variables capture a sense of the moral preferences of the subject as derived from the Trolley choices, and the test of significance on their coefficients is a test of whether such measures from hypothetical decision scenarios may yet hold power to predict decision in consequential decision tasks that contain at least some type of moral element. Table 7 focuses on estimates separating the subsamples of the data for the  $(x,y)$  distributions where  $x < y$  (disadvantageous payoff inequality) versus  $x \geq y$  (advantageous inequality).

We first focus on the results in Table 6.<sup>19</sup> The statistically insignificant coefficient on the *Unilateral Burn* indicator variable leads us to reject hypothesis H6—money burning is not greater in *Bilateral* compared to *Unilateral Burn*. This suggests that beliefs that others will burn money do not impact money burning decisions in our data. Statistically significant positive coefficients on *Income < other* in all three models support rejection of our selfish or utilitarian hypothesis H4 in favor of hypothesis H5 where disadvantageous inequality aversion (Fehr and Schmidt, 1999) motivates money burning. The marginally significant ( $p < .10$ ) coefficient on the *Relative Cost* of burning money indicates that money burning is responsive to how much of one's payoff the burning choice will cost—a lower relative budget impact of burning marginally increases the likelihood that one burns money, which supports hypothesis H7. Model (2) includes an indicator variable for those who opted out of the Trolley dilemma, and we find a marginally significant impact of *Opt-Out* on the probability that one will burn money. This variable is absent in model (3) where we include the Trolley immorality measures as regressors, which necessarily implies we focus on the money burning data from those who also completed the Trolley dilemma choices. Importantly, in model (3) of Table 6, we find evidence that making a morally dubious choice(s) in the Trolley dilemma predicts a significantly increased likelihood of money burning.<sup>20</sup> This is support for hypothesis H8. Thus, we offer first evidence in the literature, to our knowledge, that moral indicators from a hypothetical dilemma can predict significant increases in anti-social money burning choices with real payoff consequences.<sup>21</sup> And, importantly, the lack of significance on the coefficient estimate for *Action Propensity* both Tables 6 and 7 models is a more clean test of whether utilitarianism is linked to antisocial

<sup>18</sup> Appendix C contains a separate robustness estimation of our Table 6 results in Table C1 Table 6, model ((3) is included as model (1) of Table C1). Here, we interact *Unilateral Burn* with the key immorality variables to establish that, not only is there no main effect difference between money burning tendency among those in *Unilateral* versus *Bilateral Burn*, but also that if anything *Unilateral Burn* participants appear even more likely to burn money if identified as a person of immoral commission.

<sup>19</sup> The coefficient estimates on the ordering dummy variables (allocation pairs were presented to different subjects in increasing, decreasing, or random order of one's own payoff) indicate that there the ordering of the  $(x,y)$  options as presented to subjects does not matter.

<sup>20</sup> The difference between the impact of *Immoral Commission* versus *Immoral Omission* is not statistically significant ( $p > .10$  for the Wald test of coefficient equality).

<sup>21</sup> Model (3) also indicates a marginally significant impact of higher self-reported life happiness predicting a lower probability that one burns money.

**Table 6**  
Probability of burning money.

Independent Variable	(1)	(2)	(3)	(4)
	Marg. Effect (st. error)			
	All	All	All	No-Switch <sup>^^</sup>
Increasing (x,y) order (=1)	−0.0077 (0.0502)	−0.0047 (0.0498)	−0.0196 (0.0510)	−0.0065 (0.0514)
Decreasing (x,y) order (=1)	−0.0186 (0.0467)	−0.0173 (0.0460)	−0.0147 (0.0493)	−0.0201 (0.0524)
Unilateral Burn (=1)	−0.0146 (0.0456)	−0.0110 (0.0454)	−0.0156 (0.0440)	−0.0348 (0.0417)
Income < other (x < y)	0.0005*** (0.0001)	0.0005*** (0.0003)	0.0005*** (0.0002)	0.0005*** (0.0002)
Income > other (x > y)	−0.0002 (0.0003)	−0.0002 (0.0003)	−0.0002 (0.0003)	−0.0001 (0.0003)
Income = other (= 1)	0.0525 (0.0524)	0.0513 (0.0519)	0.0597 (0.0583)	0.0866 (0.0617)
Relative cost	−0.9551* (0.5447)	−0.9398* (0.5370)	−1.0741* (0.5787)	−1.0033 <sup>#</sup> (0.6175)
Trolley Opt-Out (=1)	—	−0.1171* (0.0409)	—	—
Action Propensity (Trolley dilemmas 2–10)	—	—	0.0675 (0.0546)	0.0723 (0.0526)
Immoral Commission (=1) (action in Trolley 1&11)	—	—	0.2655*** (0.1154)	0.2188** (0.1360)
Immoral Omission (=1) (inaction in Trolley 12)	—	—	0.3428*** (0.1282)	0.3178** (0.1802)
Male (=1)	—	—	−0.0157 (0.0433)	−0.0238 (0.0430)
Happiness ∈ [1 ,10] (10=highest current life happiness)	—	—	−0.0278* (0.0161)	−0.0292* (0.0154)
Religion ∈ [1 ,10] (10=very important)	—	—	0.0053 (0.0078)	0.0023 (0.0075)
Age	—	—	−0.0135 (0.0133)	−0.0100 (0.0127)
Observations	1107	1107	1026	972
# Participants <sup>^</sup>	123	123	114 <sup>^</sup>	108 <sup>^^</sup>
Log likelihood	−458.2919	−452.741	−406.183	−362.746

Notes: \*.10, \*\*.05, \*\*\*.001 for the 2-tailed test. Standard errors clustered at the individual subject level.

Increasing, Decreasing, Random (reference group) control for the order of the money burning allocation scenarios. Relative Cost = the 10 experimental monetary units (EMU) cost divided by the payoff in EMU if choosing not to burn money. Trolley Opt-Out = 1 if subject chose not to complete the Trolley dilemma task.

<sup>^</sup> reduced as a result of those opting out of the Trolley dilemma choice, which is used to score morality variables.

<sup>^^</sup> reduced by number of subjects who inconsistently switched choices in the money burning task.

cial choices. We find that it is not, which contrasts with existing results in the literature (Koenigs et al., 2007; Bartels and Pizarro, 2011; Gao and Tang, 2013; Bracht and Zylbersztejn, 2018). In other words, only for those Trolley dilemmas that can identify immorality in a more unambiguous way do we find the connection between Trolley immorality and money burning. Model (5) in Table 6 re-estimates the previous model (4) from the subset of participants who do not display inconsistent switching behavior in their choices. While the estimation precision on the key independent variables is somewhat reduced with this subset of data, our results remain unchanged and are still statistically significant ( $p < .05$ ).

Table 7 shows results of related estimations where the subsample of  $x < y$  versus  $x \geq y$  are used as a way to identify general money burning from “nastiness”, which would be defined as a willingness to burn money for  $x \geq y$  payoff distributions (i.e., a willingness to pay to burn money even when my payoff is at least as higher my counterpart’s). The results from Table 7 show that the *Diff Income* (=  $|y - x|$ ) represented in the Start Distribution only predicts a significantly higher probability of money burning when a subject’s payoff is lower than the counterpart’s, which again implies rejection of H4 in favor of disadvantageous inequality aversion that is sensitive to the size of inequality. Looking at the *advantageous inequality* subset of data in model (2), we see that the relative cost of burning money marginally matters in terms of anti-social “nasty” choices ( $p < .10$ ). The higher the advantageous payoff inequality is in our design, the lower the relative cost to make the money burning choice. For this reason, we see the predicted marginally higher nastiness in those scenarios where the decider is at the largest payoff advantage (see also right-half of Fig. 3). This offers some evidence of nasty preferences as the alternative hypothesis upon rejection of the utilitarian or *Homo Economicus* hypothesis H4.

Interestingly, the immorality measures from the Trolley dilemma are significant predictors of the probability one burns money H8. Model (1) shows that both immoral acts of commission and omission in the Trolley dilemma predict a 36%–38%

**Table 7**  
Probability of burning money.

Marginal Effect (st. error) displayed	(1)	(2)
Independent variable	Income $\leq$ other's	Income $\geq$ other's
Increasing (x,y) order (=1)	0.0546 (0.0705)	-0.0665 (0.0574)
Decreasing (x,y) order (=1)	0.0312 (0.0649)	-0.0510 (0.0597)
Unilateral Burn (=1)	-0.0326 (0.0527)	-0.0123 (0.0569)
Diff Income	0.0004*** (0.0001)	0.0007 (0.0005)
Equal Income (x=y)	0.0513 (0.0509)	0.2075 (0.1820)
Relative cost	—	-2.3937* (1.3978)
<i>Action Propensity</i>	0.0379	0.1121
(Trolley dilemmas 2–10)	(0.0572)	(0.0717)
<i>Immoral Commission</i> (=1)	0.3669*** (0.1797)	0.1326 (0.1146)
(action in Trolley 1&11)		
<i>Immoral Omission</i> (=1)	0.3823*** (0.1774)	0.3043*** (0.1387)
(inaction in Trolley 12)		
Male (=1)	0.0084 (0.0527)	-0.0441 (0.0541)
Happiness $\in$ [1,10]	-0.0203 (0.0172)	-0.0357* (0.0195)
(10=highest current life happiness)		
Religion $\in$ [1,10]	0.0085 (0.0092)	-0.0071 (0.0102)
(10=very important)		
Age	-0.0211 (0.0151)	-0.0105 (0.0176)
Observations	570	570
# Participants <sup>^</sup>	114	114
Log likelihood	-196.646	-239.281

Notes: \*.10, \*\*.05, \*\*\*.001 for the 2-tailed test. Standard errors clustered at the individual subject level.

<sup>^</sup> reduced as a result of those opting out of the Trolley dilemma choice, which is used to score morality variables.

increase in the likelihood one burns money ( $p < .01$  in both cases. The difference between these two effects is statistically insignificant,  $p > .10$ ). We identify predictors of nastiness in model (2) of Table 7 and a key result is that we find that the immoral act of omission in Trolley dilemma #12 (i.e., *not* acting when 6 lives could be saved at the expense of zero lost lives) predicts a 30% increased likelihood of making a “nasty” money burning choice ( $p < .01$ ). In a sense, our strongest way to judge morality from the Trolley dilemma is whether someone chose the immoral act of omission. In sum, we find strong support for hypothesis H8 and conclude that Trolley morality, though hypothetical, can be a significant predictor of consequential antisocial decisions.

## 5. Discussion

Economists have long challenged the assumption of *homo economicus* and recognize that people are not always own-payoff maximizing. Rather, they may be altruistic, fairness-minded, cooperative, or perhaps even anti-social. Using laboratory methods with real payoff, experimental economics has shown that participants in dictator games often share their endowment (Forsythe et al., 1994; Hoffman et al., 1994), they reciprocate in gift exchange or trust environments (Berg et al., 1995; Fehr et al., 1998), and they contribute positive amounts in public goods games (see surveys in Ledyard, 1995; Chaudhuri, 2011). Studies focusing on the darker side of human behavior are remarkably more limited in economics. This is all the more surprising given that unethical behavior within organizations is not rare and often results in high costs for the entire society. Anti-social behaviors, in general, can result in relational, workplace, or other costs to society that are nontrivial.

In this paper we attempted to identify some key determinants of costly antisocial behaviors using measures derived from both a money burning game and a moral thought experiment. While ethical dilemmas and thought experiments have been of significant interest to moral philosophers for decades, we believe our study to be unique. Our particular innovation has been to use responses in the iconic Trolley dilemma to generate immorality indicators that have predictive power regarding one's decisions in consequential environments. The consequential environment we explore allows for costly antisocial choice and may be considered a type of behavioural marker for the likelihood of costly actions in field settings.

Our results highlight the importance of the relative cost of the ethical behavior across the domains of both the hypothetical Trolley dilemma and the consequential Money Burning game. Subjects are more likely to make an ethically dubious

choice if the costs of doing so are lower. Aside from identifying typical response patterns in the Trolley dilemma, we identified choices made from our set of Trolley dilemmas that would constitute morally questionable acts of omission or commission. We then estimated a significant increase in the likelihood of burning money for those subjects identified as willing to commit an immoral act of omission or commission in the Trolley dilemma. Upon further investigation, we found that the immoral Trolley respondents' increased willingness to burn money was linked more strongly to disadvantageous inequality aversion than to nastiness. Nevertheless, we identified that choice in one Trolley scenario (not typically considered in the existing literature) is a highly significant predictor of the probability of nasty money burning. These results call into question some recent conclusions in the literature regarding increased Utilitarianism among those with anti-social personality traits (Koenigs et al., 2007; Bartels and Pizarro, 2011; Gao and Tang, 2013; Bracht and Zylbersztejn, 2018). Specifically, our research connects immoral, rather than utilitarian, choices to anti-social behavior in a stylized game.

As always, there are limitations to our study. First, it is likely the case that reputational concerns may be important if one is aware that selection in some field setting (e.g., hiring choice) based on Trolley dilemma responses may be at stake. And of course, the validity of a hypothetical ethical dilemma may always be a point of concern. For this reason, one of our main purposes is to highlight that response patterns in such hypothetical dilemmas may be instructive towards an understanding of consequential behavioural tendencies. At some level, the criticism of selection bias would apply to any number of hypothetical or self-response instruments used to screen individuals or assess situational risk. We believe the key is that we first understand the link between hypothetical responses and consequential behaviors, because researchers often have no alternative approach to study high stakes choices in the moral domain.

Our hope is that this research will stimulate further investigations into the value of hypothetical choices towards predicting outcomes in other non-hypothetical but related decision domains. These findings may have interesting implications for how hypothetical scenario instruments could be used to screen individuals for antisocial tendencies that could be costly to an organization. Because the type of anti-social decision making we studied involves resource destruction when outcome inequality is present, it is intriguing to consider that the markers for such behavioural tendencies may already exist in well-known hypothetical thought scenarios. Imagine that an employer could use responses to the Trolley dilemma as a way to identify workers who may be more willing to engage in antisocial resource destruction. While this may seem like the type of worker to avoid (i.e., do not hire such individuals in designing self-driving auto accident avoidance algorithms), those willing to destroy resources in a way that is *not* anti-social may have value to the employer in certain specialized roles (e.g., lead negotiator who must credibly be willing to walk away from a contractual arrangement or wage negotiations).

Our results may therefore be useful in identifying the benefits of improved screening in matching markets, in general. For example, previous studies would have suggested employers hire a "utilitarian", as identified by a moral dilemma battery, a suitable candidate to positions requiring difficult but necessary decisions. However, our results show reason for caution as these utilitarians might be masking underlying antisocial tendencies that would be destructive in the organization but cannot be separately identified using traditional a traditional behavioral questionnaire. As an alternative example, consider how improved screening in online dating markets may use creative approaches to identify desirable traits that are not unintentionally confounded with antisocial traits. For example, generic leadership suitability questions may be inadequate as they capture both desirable leadership qualities as well as antisocial tendencies that may appear disproportionately in certain leadership context (see Landay et al., 2019, regarding psychopathy and leadership).

Of course, such implications of our findings are themselves only a thought experiment, but we hope them to be useful at motivating why this may be a fruitful area for research extensions. If choices in hypothetical dilemmas can serve as behavioural markers that predict real world ethical choice, then we feel this is a useful step forward in an important area of behavioural research. Additionally, current technological developments (e.g. the utilization of drones and self-driving vehicles) render hypothetical moral dilemmas like the Trolley dilemma increasingly relevant to policy-makers as society attempts to understand barriers to technology adoption and implementation (e.g., Crockett, 2016). Our results have implications for how choices in hypothetical moral dilemmas may be used to understand or even possibly forecast certain types of behavior in real world environments.

## Acknowledgments

We would like to thank Elven Priour for programming for the experiment. Financial support from the [Agence Nationale de la Recherche](#) Felis Project ANR-14-CE28-0010 is gratefully acknowledged. We also thank the participants at the Beta workshop at Nancy University and at the Fall North American Economic Science Association meetings in Guatemala (2018).

## Appendix A. Experiment Instructions

### Money burning game instructions

*The following instructions are translated from French. These are for the Unilateral burn treatment (table presents allocation pairs (x,y) in an increasing order). The instructions for the other treatments are available from the authors upon request.*

You are participating in an economics experiment during which you can earn money. It is therefore important to read these instructions carefully. All earnings in this experiment will be expressed in terms of ECU (*Experimental Currency Units*). At the end of the sessions, these earnings will be converted to Euros as follows:

- 8 points = 1 Euro
- You will also receive a show-up fee of 8 Euros

At the beginning of the experiment, you will be randomly assigned a role: A or B. Therefore, you will be either a player of type A or of type B. You will keep the same role during the entire experiment.

#### Description of the Game

Suppose you are player A. You will be randomly matched with a player B in the room. A table as shown below will appear on your screen. For each row of the table you will have to answer the following question (yes or no):

« You receive an endowment of  $x$  points. Player B receive an endowment of  $y$  points. You have the opportunity to reduce player B's endowment by **50** points, which will cost you 10 points. In this case, you will get  $x-10$  points and the other player will get  $y-50$  points. » Do you want to reduce player B's payoff?

Each row of the table corresponds to a particular value of  $x$  and  $y$ .

Ligne	Votre gain	Gain de l'autre joueur	Décision de réduire le gain de l'autre joueur
1	50	250	Oui <input type="radio"/> <input type="radio"/> Non
2	50	200	Oui <input type="radio"/> <input type="radio"/> Non
3	50	150	Oui <input type="radio"/> <input type="radio"/> Non
4	50	100	Oui <input type="radio"/> <input type="radio"/> Non
5	50	50	Oui <input type="radio"/> <input type="radio"/> Non
6	100	50	Oui <input type="radio"/> <input type="radio"/> Non
7	150	50	Oui <input type="radio"/> <input type="radio"/> Non
8	200	50	Oui <input type="radio"/> <input type="radio"/> Non
9	250	50	Oui <input type="radio"/> <input type="radio"/> Non

Once you have filled out the entire table, the computer will randomly choose a row of the table that will determine your payoff as well as player B's payoff. At the end, you will observe your payoff and player B's payoff.

Your payoff is calculated as follows:  $x$ -cost of reduced points for player B.

In addition you receive a show up fee of 8 euros.

*Example 1:* suppose that row 5 is randomly chosen and that for this row you had decided to reduce player B's payoff, then your payoff is  $50-10=40$ .

*Example 2:* suppose that row 5 is randomly chosen and that for this row you had decided not to reduce player B's payoff, then your payoff is 50.

Suppose you are player B. You will be randomly matched with a player A in the room. In this game you have no decision to take and it is player A's decision that will determine your payoff.

Once player A has filled out the entire table, the computer will randomly choose a row of the table that will determine player A's payoff as well as your payoff. At the end, you will observe your payoff and player A's payoff.

Your payoff is calculated as follows:  $x$ - reduced points by player A

In addition you receive a show up fee of 8 euros.

*Example 1:* suppose that row 5 is randomly chosen and that for this row player A had decided to reduce player B's payoff, then your payoff is 0.

*Example 2:* suppose that row 5 is randomly chosen and that for this row player A had decided not to reduce player B's payoff, then your payoff is 50.

One row of the table will be randomly chosen to determine payoffs in this experiment

If you have any question regarding the instructions, please raise your hand. We will answer your questions in private.

Trolley dilemma instructions

The following instructions are translated from French (Note: participants could opt out of this task)

A runaway trolley will kill X persons on the track. You have the option to act to save these X people.

The table below describes different scenarios where each row corresponds to particular values of X and Y.

In the left part of the table, you have the possibility to **pull a lever** to divert the trolley to a different track. In this case, the train will run onto a different track on which there are Y people who will die. You have to answer yes or no to the following question: Are you willing to pull a lever to divert the trolley to a different track where Y on that side track will die to save X people on the main track?

In the right part of the table, you have the possibility to **push onto the main track** Y people who will be killed by the train but yet stop the train. You have to answer yes or no to the following question: Are you willing to kill Y people by pushing them onto the side track to save X people on the main track?

Note that you also have the opportunity to not answer these questions by clicking on the button below the table.

Un train hors de control va tuer X personnes qui se trouvent sur les rails. Vous avez la possibilité de réagir pour sauver toutes ces X personnes.

Le tableau ci-dessous décrit différents scenario où chaque ligne du tableau correspond à un nombre particulier de X et de Y.

Dans la colonne de gauche du tableau, vous avez la possibilité d'utiliser un  **aiguillage**  pour faire changer le train de voie. Dans ce cas, le train va prendre une autre voie sur laquelle sont Y personnes qui vont alors mourir. Vous devez répondre par oui ou non à la question suivante : souhaitez-vous utiliser l'aiguillage qui va faire que le train tue Y personnes pour sauver les X personnes sur le rail principal ?

Dans la colonne de droite du tableau , vous avez la **possibilité de pousser sur la voie principale** Y personnes qui vont mourir écrasées par le train pour faire arrêter le train. Vous devez répondre par oui ou non à la question suivante : Souhaitez-vous tuer Y personnes en les poussant pour sauver les X personnes sur le rail principal ?

Souhaitez-vous utiliser l'aiguillage qui va faire que le train tue Y personnes pour sauver les X personnes sur le rail principal ?			Souhaitez-vous tuer Y personnes en les poussant pour sauver les X personnes sur le rail principal ?		
Totalité des personnes tuées	Totalité des personnes sauvées		Totalité des personnes tuées	Totalité des personnes sauvées	
Y=6	X=6	Oui <input type="radio"/> Non <input type="radio"/>	Y=6	X=6	Oui <input type="radio"/> Non <input type="radio"/>
Y=5	X=6	Oui <input type="radio"/> Non <input type="radio"/>	Y=5	X=6	Oui <input type="radio"/> Non <input type="radio"/>
Y=4	X=6	Oui <input type="radio"/> Non <input type="radio"/>	Y=4	X=6	Oui <input type="radio"/> Non <input type="radio"/>
Y=3	X=6	Oui <input type="radio"/> Non <input type="radio"/>	Y=3	X=6	Oui <input type="radio"/> Non <input type="radio"/>
Y=2	X=6	Oui <input type="radio"/> Non <input type="radio"/>	Y=2	X=6	Oui <input type="radio"/> Non <input type="radio"/>
Y=1	X=6	Oui <input type="radio"/> Non <input type="radio"/>	Y=1	X=6	Oui <input type="radio"/> Non <input type="radio"/>
Y=1	X=5	Oui <input type="radio"/> Non <input type="radio"/>	Y=1	X=5	Oui <input type="radio"/> Non <input type="radio"/>
Y=1	X=4	Oui <input type="radio"/> Non <input type="radio"/>	Y=1	X=4	Oui <input type="radio"/> Non <input type="radio"/>
Y=1	X=3	Oui <input type="radio"/> Non <input type="radio"/>	Y=1	X=3	Oui <input type="radio"/> Non <input type="radio"/>
Y=1	X=2	Oui <input type="radio"/> Non <input type="radio"/>	Y=1	X=2	Oui <input type="radio"/> Non <input type="radio"/>
Y=1	X=1	Oui <input type="radio"/> Non <input type="radio"/>	Y=1	X=1	Oui <input type="radio"/> Non <input type="radio"/>
Y=0	X=6	Oui <input type="radio"/> Non <input type="radio"/>	Y=0	X=6	Oui <input type="radio"/> Non <input type="radio"/>

Ne souhaite pas répondre

OK

Appendix B. Comparative Static Predictions (theoretical framework)

We present a theoretical model that introduces considerations for intrinsic moral obligations in the utility function (e.g., Nyborg, 2000; Brekke et al., 2003; Figueires et al., 2013; Dickinson et al., 2018). Precisely we enrich the agent's utility function by introducing a function of moral motivation. Recall that utility in our behavioural model with moral obligation was defined as:

$$U = b(a) - c(a) - v(a - \hat{a}) \tag{1}$$

where  $a$ , is an action that generates both benefits,  $b$ , and costs,  $c$ .  $v(a - \hat{a})$  is a non-monetary moral function where  $\hat{a}$  describes one's moral imperative such that any deviation from this moral standard of action,  $a$ , generates disutility. We assume  $b' > 0$ ,  $c' > 0$ ,  $b'' < 0$ ,  $c'' > 0$ , such that utility benefits and costs are increasing in the action, and benefits increase at a decreasing rate while costs increase at an increasing rate. The disutility of deviations from one's moral ideal are captured by assuming  $v' > 0$  if  $a > \hat{a}$ ,  $v' < 0$  if  $a < \hat{a}$ , and  $v' = 0$  if  $a = \hat{a}$ . We also assume that  $v'' > 0$  such that marginal disutility increases at an increasing rate as one's action gets further from the moral obligation. Note that an "action" here is quite general.

Following Figueires et al. (2013), we assume that moral motivation is weak in the sense that it can be influenced by others' activities or expectation of others' activities. Precisely, we conceptualize the weak moral motivation (or obligation) of each agent as a combination of three arguments:i) an autonomous obligation denoted  $\tilde{a}_i \in [0, 1]$ , ii) a social influence

argument  $\bar{a}_{-i}$ , and iii) fairness considerations captured by a composite variable  $z$ . The autonomous logic is captured by an ideal, or “ethical,” level noted  $\hat{a}_i \in [0, 1]$ . Such an autonomous morality can be grounded on a Kantian categorical imperative, or on an unconditional commitment to a contribution (Laffont, 1975; Harsanyi, 1980). The second argument captures social influences through either the observation of others’ unethical activities and/or beliefs about others’ actions  $\bar{a}_{-i}$ . Finally, the third argument, noted  $z \in [-\hat{a}_i, \hat{a}_i]$ , captures fairness considerations in a broad definition (Rabin, 1993; Fehr and Schmidt, 1999) that can affect moral motivation; it includes feeling of being treated badly (well) by others but also feeling of being badly (well) treated by the Nature in term of bad/good luck. Depending on the nature of the action  $a$ , the value of parameter may be either positive or negative. Precisely, if a person feels he is treated badly (kindly) by others or by the nature, he may revise downward (upward) her moral ideal obligation.<sup>22</sup> Accordingly, following Figuières et al. (2013), we then define *strong moral motivation* as an unconditional commitment to stick to one’s ideal moral target. In contrast, *weak moral motivation* refers to one’s sensitivity to the observation/expectation of others’ actions, which can lead to a revision of one’s moral ideal target.<sup>23</sup> Overall, the qualified moral obligation,  $\hat{a}_i$  can be defined as a function of the aforementioned variables:  $\hat{a}_i = \hat{a}_i(\bar{a}_{-i}, z)$ . We assume that  $\frac{\partial \hat{a}_i}{\partial \bar{a}_{-i}} \geq 0$ ,  $\frac{\partial \hat{a}_i}{\partial z} \geq 0$  and  $\frac{\partial \hat{a}_i}{\partial z} \geq 0$

Individuals choose action,  $a$ , to maximize utility, yielding the following first order condition:

$$\frac{\partial U}{\partial a} = b'(a) - c'(a) - v'(a - \hat{a}) = 0 \tag{2}$$

This can be solved for the optimal action level  $a^* = a(\hat{a})$  such that the following identity holds:

$$b'(a^*(\hat{a})) - c'(a^*(\hat{a})) - v'(a^*(\hat{a}) - \hat{a}) \equiv 0 \tag{3}$$

From this we can derive the comparative static result of interest by differentiating with respect to one’s moral obligation:

$$b'' \frac{\partial a^*}{\partial \hat{a}} - c'' \frac{\partial a^*}{\partial \hat{a}} - v'' \left( \frac{\partial a^*}{\partial \hat{a}} - 1 \right) = 0 \tag{4}$$

This can be solved for:

$$\frac{\partial a^*}{\partial \hat{a}} = \frac{-v''}{b'' - c'' - v''} > 0 \tag{5}$$

Thus, the optimal level of action is positively linked to one’s moral obligation in the decision scenario.

### 3.1. Trolley problem

**Proof of H1:** Let’s consider the following maximization problem *without* the morality argument,  $v(a - \hat{a})$  in the utility function, and assuming multiple and separable benefits and costs of one’s action (this may facilitate consideration of each live saved or lost in the Trolley dilemma):

$$\max_{a_i} U_i = \sum_{j=1}^n b_j(a_i) - \sum_{k=1}^m c_k(a_i) \tag{6}$$

where  $\sum_{j=1}^n b_j(a_i)$  corresponds to the aggregate benefits in term of lives saved when taking action,  $a_i$ , and  $\sum_{k=1}^m c_k(a_i)$  is the aggregate cost in terms of lives sacrificed. From (6) we have the following first order condition:

$$\frac{\partial U}{\partial a_i} = 0 \rightarrow nb' = mc' \tag{7}$$

Assuming that  $b'=c'$ , such that the marginal value of a saved life equals the marginal cost of a sacrificed life, then a utilitarian should always choose action as long as  $n > m$  and should abstain from acting otherwise.

Let’s now relax some assumption and consider the case of agents with moral concerns represented by the following utility function (indexing  $v$  by the lives lost allows the moral imperative to potentially differ across lives sacrificed).

$$\max_{a_i} U_i = \sum_{j=1}^n b_j(a_i) - \sum_{k=1}^m c_k(a_i) - \sum_{k=1}^m v_{ik}(a_i - \hat{a}_i) \tag{8}$$

<sup>22</sup> For instance, imagine the case of a dictator game played twice. Suppose that player  $i$  is the dictator and player  $j$  is the receiver in period 1. In period 2, the role are reversed. Suppose also that player  $i$  keeps all his endowment for himself in period 1. In absence of information regarding the issue of the game player during the first period, player  $j$  will choose his ideal amount sent to player  $i$  based on his ideal moral obligation,  $\hat{a}_i$ . Suppose now that player  $j$  is informed of player  $i$ ’s decision in period 1 before taking his decision. Then he may revise downward his decision because he feels he is badly treated by player  $i$  ( $z < 0$ ). But player  $j$  may also feel badly treated by nature if for instance, player  $i$ ’s decision in period 1 in term of allocation of wealth is replaced by a random allocation. In this case, it is also possible that player  $j$  may revise downward her decision by the simple fact of being badly treated by nature.

<sup>23</sup> The extent of such a revision of moral motivation typically varies across individuals: strongly morally motivated agents will closely stick to their ideal target, whereas weakly motivated agents are prone to revise their morally ideal target whenever they observe or anticipate a gap between their own and others’ money burning decisions. Our idea is that most people are of the “mixed” type, i.e., their actual moral target is the outcome of a deliberative process through which their preferred moral target is balanced against others’ anticipate level of money burning.

From (8) we derive the first-order condition (F.O.C.):

$$\frac{\partial U}{\partial a_i} = 0 \rightarrow nb' = mc' + mv' \tag{9}$$

Eq. (9) indicates that there is now an additional marginal cost  $mv'$  to sacrifice  $m$  lives and that cost may counterbalance the *utilitarian* calculus described above in Eq. (6). Depending on the individual weight of moral concern in the utility function, it is now unclear the best action to maximize utility. Indeed, if the marginal moral cost of taking action that sacrifices  $m$  lives in order to save  $n$  lives outweighs what would otherwise be a net gain in utility,  $nb' - mc' < mv'$  then the individual will abstain from acting. This condition may also be written as  $nb' < m(c' + v')$ , which highlights that the relevant comparison is now the sum of all marginal benefits relative to the sum of all costs (traditional plus moral costs). It is clear from (9) that the likelihood of acting will increase in the number of lives saved,  $n$ , holding  $m$  constant.

**Proof of H2:** the specific dilemma in the trolley problem where lives lost are unaffected ( $X=Y$  dilemmas) corresponds to the case in our model where  $m$  equals  $n$ . Assuming  $b'=c'$  (otherwise, some lives matter more than others, which is a clear extension of this model), the F.O.C. in (9) reduces to  $0 = mv'$ . This condition is only met when the individual makes a choice precisely at one's moral obligation,  $a = \hat{a}$ . Only agents endorsed with more immoral or nasty preferences would be inclined to take action here since the moral obligation is to be actively responsible for the lives lost, rather than passively allow a similar number of deaths. This could be interpreted in our model as having a relatively high  $\hat{a}$  parameter (since by nature,  $\hat{a}$  close to zero means a highly moral agent) such that  $a < \hat{a}$  and  $v' < 0$ . In such case, there is a gain to increase action,  $a$ , such that the moral cost of deviating from one's target will decrease.

Let's now consider the (6,0) Trolley dilemma, where action costs no lives. In such case,  $c'=0$  and  $m=0$ . Here, there are no longer moral costs of lives lost since no one dies in the (6,0) dilemma, so inaction is only justified if one assumes moral costs would be incurred by *saving* lives. In this case, inaction in any (X,0) Trolley dilemma maximizes utility. In such a dilemma, to not take action would be consider an immoral act of *omission*.

**Proof of H3:** Our last assumption concern the role of framing. Due to the distinction between personal and impersonal moral dilemmas (Greene et al., 2001) and based on the “contact principle” (Cushman et al., 2006), we predict an individual is more likely to take action in the *INDIRECT* frame. In our theoretical model, framing effect is captured by the fact that moral cost is higher of an action is higher in the *DIRECT* frame treatment,  $v'_{DIRECT} > v'_{INDIRECT}$ .

### 3.2. Money Burning

**Proof of H4:** Under the assumption of either pure selfishness or utilitarianism, individuals should never burn money. Consider now the case of *homo moralis* agents represented by the following utility function:

$$\max_{a_{ij}} U_i = b - \sum_{j=1}^n c_i(a_{ij}) - \sum_{j=1}^n v_{ij}(a_{ij} - \hat{a}_i) \tag{10}$$

Here, the first term  $b$  corresponds to initial material endowment that is independent of action  $a_{ij}$ ; the second term is the total monetary cost for agent  $i$  of burning other player's  $j$  payoff by choosing action  $a_{ij}$ . The third term is the moral cost of burning others' resources.

From (10) we derive the first order condition:

$$\frac{\partial U}{\partial a_{ij}} = 0 \rightarrow -nc' - nv' = 0 \tag{11}$$

From (11) it is straightforward that whether the optimal action effort of money burning will be zero or positive depends on the sign of  $v'$ , which varies based on one being above or below one's moral obligation action.

If  $\hat{a}$  is low (for instance if  $\hat{a} = 0$ ), which could be interpreted as the fact of having high moral obligation, then any increase of  $a$  for  $a > \hat{a}$  will increase the moral cost such that  $v' > 0$ . In this case the non-monetary moral cost adds to the material cost  $c'$  and reinforce the tendency not to burn.

Only if  $\hat{a}$  is sufficiently high such that  $a < \hat{a}$  and  $v' < 0$ , there is a gain to increase effort  $a_i$  and thus to engage in money burning. Thus a relatively high  $\hat{a}$  parameter may be interpreted as *nasty* preferences (Abbink and Sadrieh, 2009; Abbink and Herrmann, 2011). Specifically, individuals with nasty preferences (i.e. those having a sufficiently high moral target  $\hat{a}$  such that any increase of effort  $a$  will reduce the nonmonetary cost ( $v' < 0$ ), as long as  $a < \hat{a}$ ) will engage in burning money if  $v' > c'$ .

**Proof of H5:** If disadvantageous inequality aversion matters, one should therefore observe money burning only when  $x < y$ , while money should not be burnt in cases of advantageous inequality (i.e.,  $x \geq y$ ). In our model, inequality aversion may be captured by a negative parameter  $z$  in the moral function of individuals with  $x < y$  that may lead them to revise downward their ideal moral motivation. A negative  $z$  parameter reflects here the fact of being unfairly treated by Nature (i.e., given a low endowment). This negative  $z$  parameter then motivates money burning due to a revised moral obligation target.

**Proof of H6:** Our theoretical framework allows us to account for pre-emptive retaliation by assuming that moral motivation is *weak*. Here *weak moral motivation* refers to one's sensitivity to the expectation of others' actions noted  $a_j$ , which can lead to a revision of one's moral ideal target. Thus in the bilateral treatment, individuals might revise upward their targeted level of money burning  $\hat{a}_i$  if they expect that the counterpart may burn, which may lead the individuals to engage in pre-emptive money burning to meet their ethical obligation.

**Proof of H7:** This follows directly from the F.O.C. in (11), while holding  $v'$  constant.

**Proof of H8:** This is a testable hypothesis based on our assumption that moral targets  $\hat{a}$  in the Trolley dilemma reflect one's ethics in other consequential decision problems. This is not mathematically proven.

## Appendix C

Table C1.

**Table C1**

Probability of burning money (model (1) reproduces model (3) in Table 6).

Independent Variable	(1) Marg. Effect (st. error)	(2) Marg. Effect (st. error)
Increasing (x,y) order (=1)	-0.0196 (0.0510)	-0.0220 (0.0513)
Decreasing (x,y) order (=1)	-0.0147 (0.0493)	-0.0105 (0.0495)
Unilateral Burn (=1)	-0.0156 (0.0440)	-0.0459 (0.0459)
Income < other (x < y)	0.0005*** (0.0002)	0.0005*** (0.0002)
Income > other (x > y)	-0.0002 (0.0003)	-0.0001 (0.0003)
Income = other (= 1)	0.0597 (0.0583)	0.0593 (0.0587)
Relative cost	-1.0741* (0.5787)	-1.0903*# (0.5785)
Action Propensity (Trolley dilemmas 2–10)	0.0675 (0.0546)	0.0551 (0.0463)
Immoral Commission (=1) (action in Trolley 1&11)	0.2655*** (0.1154)	0.2064** (0.1152)
Immoral Omission (=1) (inaction in Trolley 12)	0.3428*** (0.1282)	0.3055** (0.1699)
Immoral Commission (=1) * Unilateral Burn		0.3864** (0.1868)
Immoral Omission (=1) * Unilateral Burn		0.0913 (0.1697)
Male (=1)	-0.0157 (0.0433)	-0.0007 (0.0435)
Happiness ∈ [1 ,10] (10=highest current life happiness)	-0.0278* (0.0161)	-0.0267 (0.0166)
Religion ∈ [1 ,10] (10=very important)	0.0053 (0.0078)	0.0062 (0.0080)
Age	-0.0135 (0.0133)	-0.0166 (0.0139)
Observations	1026	1026
# Participants <sup>^</sup>	114 <sup>^</sup>	114 <sup>^</sup>
Log likelihood	-406.183	-403.302

Notes: \*.10, \*\*.05, \*\*\*.001 for the 2-tailed test. Standard errors clustered at the individual subject level.

Increasing, Decreasing, Random (reference group) control for the order of the money burning allocation scenarios.

Relative Cost = the 10 experimental monetary units (EMU) cost divided by the payoff in EMU if choosing not to burn money.

<sup>^</sup> reduced as a result of those opting out of the Trolley dilemma choice, which is used to score morality variables.

## References

- Abbink, K., Herrmann, B., 2011. The moral costs of nastiness. *Econ. Inq.* 49 (2), 631–633.
- Abbink, K., Sadrieh, A., 2009. The pleasure of being nasty. *Econ. Lett.* 105, 306–308.
- Anderson, C.M., Putterman, L., 2006. Do non-strategic sanctions obey the law of demand? the demand for punishment in the voluntary contribution mechanism. *Games Econ. Behav.* 54 (1), 1–24.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I., 2018. The moral machine experiment. *Nature* 563 (7729), 59.

- Bauman, C.W., McGraw, A.P., Bartels, D.M., Warren, C., 2014. Revisiting external validity: concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Soc. Personal Psychol. Compass* 8 (9), 536–554.
- Bartels, D.M., Pizarro, D.A., 2011. The mismeasure of morals: antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition* 121, 154–161.
- Becker, G.S., 1968. Crime and punishment: an economic approach. In: *The Economic Dimensions of Crime*. Palgrave Macmillan, London, pp. 13–68.
- Bentham, J., 1789. *An Introduction to the Principles of Morals*. Athlone, London.
- Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. *Games Econ. Behav.* 10 (1), 122–142.
- Bonnefon, J.F., Shariff, A., Rahwan, I., 2016. The social dilemma of autonomous vehicles. *Science* 352 (6293), 1573–1576.
- Bostyn, D.H., Sevenhant, S., Roets, A., 2018. Of mice, men, and trolleys: hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychol. Sci.* 29 (7), 1084–1093.
- Bracht, J., Zylbersztejn, A., 2018. Moral judgments, gender, and antisocial preferences: an experimental study. *Theory Decis.* 85 (3–4), 389–406.
- Brandts, J., Charness, G., 2011. The strategy versus the direct-response method: a first survey of experimental comparisons. *Exp. Econ.* 14 (3), 375–398.
- Brekke, K., Kverndokk, S., Nyborg, K., 2003. An economic model of moral motivation. *J. Public Econ.* 87, 1967–1983.
- Bruner, D.M., 2009. Changing the probability versus changing the reward. *Exp. Econ.* 12 (4), 367–385.
- Carney, D.R., Mason, M.F., 2010. Decision making and testosterone: when the ends justify the means. *J. Exp. Soc. Psychol.* 46 (4), 668–671.
- Charness, G., Masclat, D., Villeval, M.C., 2013. The dark side of competition for status. *Manage Sci.* 60 (1), 38–55.
- Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Exp. Econ.* 14 (1), 47–83.
- Cima, M., Tonnaer, F., Hauser, M.D., 2010. Psychopaths know right from wrong but don't care. *Soc. Cogn. Affect Neurosci.* 5 (1), 59–67.
- Cox, J.C., Servátka, M., Vadovič, R., 2017. Status quo effects in fairness games: reciprocal responses to acts of commission versus acts of omission. *Exp. Econ.* 20 (1), 1–18.
- Crockett, M., 2016. The trolley problem: would you kill one person to save many others? *The Guardian*, Dec. 12, 2016.
- Cushman, F., Young, L., Hauser, M., 2006. The role of conscious reasoning and intuition in moral judgments: testing three principles of harm. *Psychol. Sci.* 17, 1082–1089.
- Dickinson, D.L., Masclat, D., Peterle, E., 2018. Discrimination as favoritism: the private benefits and social costs of in-group favoritism in an experiment labor market. *Eur. Econ. Rev.* 104 (May), 220–236.
- Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. *Nature* 415 (6868), 137.
- Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* 90 (4), 980–994.
- Fehr, E., Kirchsteiger, G., Riedl, A., 1998. Gift exchange and reciprocity in competitive experimental markets. *Eur. Econ. Rev.* 42 (1), 1–34.
- Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition and cooperation. *Q. J. Econ.* 114, 817–868.
- Figuières, C., Masclat, D., Willinger, M., 2013. Weak moral motivation leads to the decline of voluntary contributions. *J. Public Econ. Theory* 15 (5), 745–772.
- Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* 10 (2), 171–178.
- Foot, P., 1967. The problem of abortion and the doctrine of the double effect. *Oxford Rev.* 5, 5–15.
- Forsythe, R., Horowitz, J.L., Savin, N.E., Sefton, M., 1994. Fairness in simple bargaining experiments. *Games Econ. Behav.* 6 (3), 347–369.
- Gao, Y., Tang, S., 2013. Psychopathic personality and utilitarian moral judgment in college students. *J. Crim. Justice* 41 (5), 342–349.
- Greene, J.D., Cushman, F.A., Steward, L.E., Lowerberg, K., Nystrom, L.E., Cohen, J.D., 2009. Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition* 111, 364–371.
- Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D., 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293 (5537), 2105–2108.
- Harsanyi, J., 1980. Rule utilitarianism, rights, obligations and the theory of rational behavior. *Theory Decis.* 12, 115–133.
- Hoffman, E., McCabe, K., Shachat, K., Smith, V., 1994. Preferences, property rights, and anonymity in bargaining games. *Games Econ. Behav.* 7 (3), 346–380.
- Holt, C.A., Laury, S.K., 2002. Risk aversion and incentive effects. *Am. Econ. Rev.* 92 (5), 1644–1657.
- Kahane, G., 2015. Sidetracked by trolleys: why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Soc. Neurosci.* 10 (5), 551–560.
- Kant, I., 1787. "On a supposed right to lie from benevolent motives," 1787. In: Beck, Lewis W. (Ed.), *The Critique of Practical Reason and Other Writings in Moral Philosophy*. University of Chicago Press, Chicago, pp. 346–350 1949.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., Damasio, A., 2007. Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature* 446 (7138), 908–911.
- Laffont, J.J., 1975. Macroeconomic constraints, economic efficiency and ethics: an introduction to kantian economics. *Economica* 42, 430–437.
- Landay, K., Harms, P., Credé, M., 2019. Shall we serve the dark lords? a meta-analytic review of psychopathy and leadership. *J. Appl. Psychol.* 104 (1), 183–196.
- Lazear, E., 1989. Pay equality and industrial politics. *J. Political Econ.* 97 (3), 561–580.
- Ledyard, J., 1995. Public goods: a survey of experimental research. In: Kagel, J., Roth, A. (Eds.), *The Handbook of Experimental Economics*. Princeton Univ. Press, Princeton, pp. 111–194.
- Line, M.B., Zand, A., Stringhini, G., Kemmerer, R., 2014. Targeted attacks against industrial control systems: is the power industry prepared? In: *Proceedings of the 2nd Workshop on Smart Energy Grid Security*, November (2014). ACM, pp. 13–22.
- Mill, J.S., 1861 [1998]. *L'Utilitarisme*. Paris: PUF.
- Navarrete, C.D., McDonald, M.M., Mott, M.L., Asher, B., 2012. Virtual morality: emotion and action in a simulated three-dimensional "trolley problem". *Emotion* 12 (2), 364–370.
- Nikiforakis, N., Normann, H.T., 2008. A comparative statics analysis of punishment in public-good experiments. *Exp. Econ.* 11 (4), 358–369.
- Nyborg, K., 2000. Homo economicus and homo politicus: interpretation and aggregation of environmental values. *J. Econ. Behav. Org.* 42, 305–322.
- Holt, J., 1995. *Morality, Reduced to Arithmetic*. The New York Times, p. 19 August 5, 1995, Section 1, page.
- Petrinovich, L., O'Neill, P., Jorgensen, M., 1993. An empirical study of moral intuitions: toward an evolutionary ethics. *J. Pers. Soc. Psychol.* 64, 467–478.
- Rabin, M., 1993. Incorporating fairness into game theory and economics. *Amer. Econ. Rev.* 83, 1281–1302.
- Rai, T.S., Holyoak, K.J., 2010. Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cogn. Sci.* 34, 311–321.
- Shallow, C., Iliev, R., Medin, D., 2011. Trolley problems in context. *Judgm. Decis. Mak.* 6 (7), 593–601.
- Smith, A., 1759 [1981]. *The theory of moral sentiments*. D.D. Raphael and A.L. Macfie, eds. Liberty Fund: Indianapolis.
- Spranca, M., Minsk, E., Baron, J., 1991. Omission and commission in judgment and choice. *J. Exp. Soc. Psychol.* 27, 76–105.
- Thomson, J., 1985. The trolley problem. *Yale Law J.* 94, 1395–1415.
- Twenge, J.M., Foster, J.D., 2010. Birth cohort increases in narcissistic personality traits among American college students, 1982–2009. *Soc. Psychol. Personal. Sci.* 1 (1), 99–106.
- Werner, K.B., Few, L.R., Bucholz, K.K., 2015. Epidemiology, comorbidity, and behavioral genetics of antisocial personality disorder and psychopathy. *Psychiatr. Ann.* 45 (4), 195–199.
- Zizzo, D., 2004. Inequality and procedural fairness in a money burning and stealing experiment. In: *Inequality, Welfare and Income distribution: Experimental approaches*. Emerald Group Publishing Limited, pp. 215–247.
- Zizzo, D., Oswald, A.J., 2001. Are people willing to pay to reduce others' incomes? *Annales d'Economie et de Statistique* 63–64, 39–62.