Journal of Experimental Social Psychology 79 (2018) 200–216

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence

Jim A.C. Everett[a,*], Nadira S. Faber[a,b], Julian Savulescu[a], Molly J. Crockett[c,**]

[a] Uehiro Centre for Practical Ethics, University of Oxford, UK
[b] Department of Experimental Psychology, University of Oxford, UK
[c] Department of Psychology, Yale University, United States of America

ARTICLE INFO

Handling editor: Robbie Sutton

Keywords:
Morality
Deontology
Consequentialism
Utilitarianism
Impartiality
Person perception
Partner choice
Trust
Prosociality
Helping

ABSTRACT

Previous work has demonstrated that people are more likely to trust "deontological" agents who reject harming one person to save many others than "consequentialist" agents who endorse such instrumental harms, which could explain the higher prevalence of non-consequentialist moral intuitions. Yet consequentialism involves endorsing not just instrumental harm, but also impartial beneficence, treating the well-being of every individual as equally important. In four studies (total $N = 2086$), we investigated preferences for consequentialist vs. non-consequentialist social partners endorsing instrumental harm or impartial beneficence and examined how such preferences varied across different types of social relationships. Our results demonstrate robust preferences for non-consequentialist over consequentialist agents in the domain of instrumental harm, and weaker – but still evident – preferences in the domain of impartial beneficence. In the domain of instrumental harm, non-consequentialist agents were consistently viewed as more moral and trustworthy, preferred for a range of social roles, and entrusted with more money in economic exchanges. In the domain of impartial beneficence, preferences for non-consequentialist agents were observed for close interpersonal relationships requiring direct interaction (friend, spouse) but not for more distant roles with little-to-no personal interaction (political leader). Collectively our findings demonstrate that preferences for non-consequentialist agents are sensitive to the different dimensions of consequentialist thinking and the relational context.

1. Introduction

What unites psychologists, philosophers, and fiction writers? One thing stands out: a fascination with how people do, or should, respond when faced with a decision to sacrifice one innocent person to save a greater number of people. What should one do, for example, if the only way to prevent a major terrorist attack is to torture the child of the suspected terrorist until she releases the information of where her father is? In the academic literature, when someone endorses this harm in such "sacrificial dilemmas" they are typically said to be making a "consequentialist" (or "utilitarian") judgment in line with consequentialist ethical theories (Bentham, 1789/1983; Mill, 1863). These theories posit that consequences are the *only* thing that matters when making a moral decision - an action is good if it produces good consequences, and bad if it produces bad consequences. In contrast, when someone rejects inflicting harm on an innocent they are said to be making a "non-consequentialist", or "deontological" judgment in line with deontological ethical theories (e.g. Fried, 1978; Kant, 1797/2002; Rawls, 1971; Scanlon, 1998; W.D. Ross, 1930) positing that even if sacrificing someone to save the lives of five others is an action that maximises overall welfare ("the Good"), this does mean it is morally correct ("the Right").

Such dilemmas capture our imagination not just because they force an internal moral conflict, but because we recognize the reputational consequences that these impossible decisions might have for those who make them. Recent research has shown that agents who make consequentialist judgments in sacrificial dilemmas are seen as less moral, trustworthy and warm, chosen less frequently as social partners, and trusted less in economic exchanges (e.g. Bostyn & Roets, 2017; Sacco, Brown, Lustgraaf, & Hugenberg, 2017; Everett, Pizarro, & Crockett, 2016; Rom, Weiss, & Conway, 2017; Uhlmann, Zhu, & Tannenbaum, 2013). Such preferences are socially rational, because standard formulations of consequentialism require maximising the greater good even if this involves using, harming, and even killing innocent people.

* Correspondence to: J.A.C. Everett, Uehiro Centre for Practical Ethics, University of Oxford, Oxford, UK.
** Correspondence to: M.J. Crockett, Department of Psychology, Yale University, New Haven, CT, United States of America.
E-mail addresses: Jim.ac.everett@gmail.com (J.A.C. Everett), molly.crockett@yale.edu (M.J. Crockett).

https://doi.org/10.1016/j.jesp.2018.07.004
Received 25 November 2017; Received in revised form 15 July 2018; Accepted 16 July 2018

This consequentialist rejection of any constraints on the maximisation of welfare means that there is no place for rights, duties, and respect for individual persons: if by stealing your new laptop and selling it on the black market I could make a lot of money that I could donate to charities in the developing world to save children's lives, this is what I should do – regardless of whether I have previously made (potentially implicit) commitments not to steal from you. But expected adherence to such implicit commitments is critical when selecting a social partner for the purposes of cooperative exchange (e.g. friend, spouse, colleague). Indeed, we have argued that this tension between consequentialism and what we seek in social partners could, through mechanisms of partner choice (e.g. Baumard, André, & Sperber, 2013; Noë & Hammerstein, 1994), explain the prevalence of non-consequentialist moral intuitions (Everett et al., 2016). To the extent that people who make non-consequentialist moral judgments in sacrificial dilemmas are favored in a cooperation market - seemingly because these judgments signal a commitment to cooperation - deontological moral intuitions could therefore represent an evolutionarily prescribed prior that was selected for through partner choice mechanisms (Everett et al., 2016).

Although sacrificial moral dilemmas make good drama, these are not necessarily the most common conflict between consequentialist and deontological principles. As outlined in the two-dimensional model of utilitarian psychology (Kahane et al., 2018), consequentialist theories like utilitarianism involve more than just decisions about whether to sacrifice one to save a greater number ("*instrumental harm*"). At the core of utilitarianism is the idea of *impartial beneficence*, that we must impartially maximise the well-being of all sentient beings on the planet in such a way that "[e]ach is to count for one and none for more than one" (Bentham, 1789/1983), not privileging compatriots, family members, or ourselves over strangers – or even enemies. In general, people are attracted to impartiality, preferring fairness to unfairness (e.g. Fehr & Schmidt, 1999; Shaw, 2013; Tyler, 2000), and will choose equity over efficiency when these are in conflict, seemingly out of a desire to appear impartial (e.g. Choshen-Hillel, Shaw, & Caruso, 2015; Shaw, 2013). But there are also limits to preferences for impartiality, for example when balancing concerns of fairness and loyalty - as in the "whistleblowers dilemma" (Dungan, Waytz, & Young, 2014; Waytz, Dungan, & Young, 2013).

Critically, the strict consequentialist impartial standpoint departs substantially from common-sense morality, which posits *special obligations* to those with whom we have some kind of special relationship. Parents, for example, have certain duties and obligations to their children that are not shared by other people. These special obligations make it morally permissible (or even required) to save one's own child over, e.g., two strangers' children, and are incorporated into many forms of deontological ethics (e.g. Annis, 1987; Held, 2006; Kamm, 2007; Scanlon, 1998). Indeed – as we return to later - persistent philosophical criticisms of consequentialist theories centre on the fact that they fail to account for special obligations such as those found in friendship (e.g. Cocking & Oakley, 1995; Woodcock, 2009). Even if people prefer impartiality when deciding allocations between two family members, work colleagues, or strangers, we think it unlikely that people will prefer impartiality when deciding allocations between a single family member and a greater number of strangers. We know that when we read *The Iliad* we harshly judge Agamemnon for his consequentialist decision to sacrifice his daughter for the greater good. When reading Dicken's *Bleak House*, might we also think badly of the 'telescopic philanthropist' Mrs. Jellyby who spends most of her time setting up a charity for a far-off tribal community while ignoring the needs of her own family? We think the answer is yes.

Just as with instrumental harm, in the domain of impartial beneficence there is a deep conflict between what we seek in a social partner and the requirements of consequentialism. In the simplest terms, non-consequentialists should be preferred in the domain of instrumental harm because we do not want social partners who will harm us in order to maximise the greater good; non-consequentialists should be preferred in the domain of impartial beneficence because we want social partners who will help us even if it does not maximise the greater good.

Although we predict non-consequentialists would be preferred over consequentialists across both dimensions, it is also reasonable to assume that these preferences would be weaker when consequentialist preferences are expressed through endorsement of impartial beneficence than instrumental harm. Research on the omission bias shows that directly harming someone is judged as more morally wrong than failing to help or allowing harm to occur (Baron & Ritov, 1994; Ritov & Baron, 1990; Siegel, Crockett, & Dolan, 2017; Spranca, Minsk, & Baron, 1991), and psychologists have identified a general positive-negative psychological asymmetry whereby "bad is stronger than good" (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001), and negative information is seen as more diagnostic in impression formation and person perception (Skowronski & Carlston, 1989).

In this paper we report four studies in which we investigated social perceptions of non-consequentialist and consequentialist agents in both sacrificial dilemmas tapping endorsement of instrumental harm, and impartiality dilemmas tapping endorsement of impartial beneficence. As well as theoretically extending the conceptual space in which non-consequentialists might be preferred, we also investigate this across a much greater range of dependent measures than has previously been used. Specifically, we study partner preference by looking at two different economic games (the Trust Game and the Prisoner's Dilemma); several distinct dimensions along which the agent's character could be perceived (warmth; competence; morality); the different social roles in which the agent would be preferred (as a friend, a spouse, a boss, and as a political leader); and the different processes or motivations perceived to influence the agent's moral decision (reason vs. emotion; strategic considerations; altruistic motivations).

2. Study 1

2.1. Method

2.1.1. Open science

We report all measures,[1] manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/yuv2m/.

2.1.2. Ethics statement

For all studies, relevant ethical guidelines were followed and the research was approved through University of Oxford's Central University Research Ethics Committee, with the reference number MS-IDREC-C1-2015-098.

2.1.3. Participants

We recruited 201 participants via MTurk, and paid them $1.00 for their time. Participants were excluded from completing the survey if they had participated in related studies by us in the past, and were excluded from analysis if took the survey more than once ($N = 4$) or failed a simple comprehension check asking them to indicate the judgment their partner made in the dilemma ($N = 5$). This left us a final sample of 192 participants (98 female; $M_{age} = 33$, $SD = 9.84$). Our sample size was determined through an a priori power analysis (see supplementary methods for details) and a sensitivity power analysis for our main ANCOVA analysis, assuming an α of 0.05 and power of 0.80, indicated that the minimum effect size we had power to detect was a small-to-medium effect of $f = 0.20$.

[1] We report all measures in Study 1, with one exception: as an exploratory question for different purposes, we asked participants how much they intended to vote for Donald Trump or Hilary Clinton in the then-ongoing US elections. These were not relevant to our purposes here, and are not analyzed or reported here (or elsewhere).