No harm done?
An experimental approach to the non-identity problem

Forthcoming in *Journal of the American Philosophical Association*

Matthew Kopec Northeastern University, m.kopec@northeastern.edu
Justin Bruner University of Arizona, justinpbruner@email.arizona.edu

Abstract: Discussions of the non-identity problem presuppose a widely shared intuition that actions or policies that change who comes into existence don't, thereby, become morally unproblematic. We hypothesize that this intuition isn't generally shared by the public, which could have widespread implications concerning how to generate support for large-scale, identity-affecting policies relating to matters like climate change. To test this, we ran a version of the well-known dictator game designed to mimic the public's behavior over identity-affecting choices. We found the public does seem to behave more selfishly when making identity-affecting choices, which should be concerning. We further hypothesized that one possible mechanism is the notion of harm the public uses in their decision-making and find that substantial portions of the population seem to each employ distinct notions of harm in their normative thinking. These findings raise puzzling features about the public's normative thinking that call out for further empirical examination.

Keywords: the non-identity problem, experimental philosophy, climate change, harm, altruism

1. Introduction

If human life is to continue on this planet, we need to motivate the public to care more deeply about the welfare of people who don't yet exist. A substantial proportion of the population seems largely indifferent about what kind of world we leave to those future people – unfortunately, a large enough proportion to perpetually erase any fleeting progress made on climate change mitigation. Although some recent authors have explored conservationist arguments to motivate even Homo economicus (e.g., Broome 2018), most economists agree that solving the climate crisis comes at a substantial cost to the current generation. So, it seems our only viable democratic option is *moral* motivation. The public must come to recognize and act upon our moral obligation to leave a healthy planet to those people who don't yet exist.

Unfortunately, our obligations to people who don't yet exist are more slippery than most. Since our choices between policies will, over a long enough time-scale, change which people come into existence, it is difficult to make the argument that we are leaving an unhealthy planet to *these* people under examination. Afterall, they wouldn't have existed if we had chosen another policy. This moral quirk – the non-identity problem (Parfit 1984) – is widely recognized as a serious philosophical problem, one we must solve before we can fully understand our duties to future generations. And environmental ethicists (e.g., Gardiner 2012) tend to see this problem as a key roadblock in generating action on

climate change (a component of his "intergenerational storm"). Even the recent IPCC 2014 report mentions the non-identity problem as a major theoretical roadblock to action on climate change (Kolstad et al. 2014: 216).

But does the non-identity problem really make a difference when it comes to motivating the public to treat future generations responsibly? Some might think the obvious answer is "No!", since the public doesn't generally know about the problem, nor are they able to really grasp the issues involved. But if we take philosophers' intuitions as a guide to moral intuitions and judgments more generally, there is a deeper reason for skepticism about the practical relevance of the non-identity problem. A central driving force in the literature is the strong intuition that, just because a certain choice changes the identity of those affected, the choice doesn't thereby lose its morally problematic features. The identity-affecting choices discussed in the literature *seem* just as wrong as the parallel choices that don't change who comes into existence. This intuition is so widespread among philosophers working on this problem that even those who think the intuition is ultimately mistaken admit to sharing it anyway (e.g., Boonin 2014). So, philosophers don't generally think we are morally off the hook when our choices affect who comes into existence. And if the intuitions of the public match the intuitions of these philosophers, then the non-identity problem at least raises no *special* problem for moral motivation after all (even if it remains theoretically puzzling).

We set out to examine whether this is really the case. Our experience teaching the non-identity problem clashed with what is generally accepted: non-philosophers often *do* see a substantial moral difference between identity-affecting cases and the parallel cases that don't change who comes into existence. Since many students saw a substantial moral difference between the cases meant to motivate the problem, they didn't see why this was a puzzle worth theorizing about in the first place. We thus worried that the optimistic story from the previous paragraph is simply mistaken, in which case the non-identity problem may indeed pose a special problem for moral motivation. To test this, we developed a behavioral economic experiment, a version of the well-known dictator game, designed to probe what the public's behavior in identity-affecting choice problems might be. We admit that the study we developed is only suggestive, since it doesn't really change which people come into existence, and the control group isn't obviously and straightforwardly "harmed". (Getting such a study past a university's ethics board would be tricky, to say the least!) That said, what we found is surprising nonetheless.

We found evidence both that most in the population are fully able to follow the details of identity-affecting choice problems, and they also tend to be more selfish when confronted with such choices. One possible explanation for this change is that many people might employ a version of what has been called the 'counterfactual comparative account' of harm in their normative thinking. On this notion of harm, an agent cannot be harmed if she hasn't been made worse off, and so the agents who are causally downstream in identity-affecting actions wouldn't count as being harmed (since they wouldn't have existed otherwise). Since we designed our study to mimic this particular feature of identity-affecting choice problems, we were also able to probe the relationship between giving behavior and judgments of harm. What we found suggests that something akin to the counterfactual comparative notion of harm does indeed seem to play a role in the normative thinking of a substantial portion of the population – at least in the context of the game – and the role it plays does seem to have some effect on giving behavior. That said, there actually seems to be a substantial split within the public

over which notion of harm to employ when making these kinds of normative evaluations, which calls out for further empirical examination.

We should pause to say, from the outset, that we only take our exploration of the non-identity problem to be suggestive, and we recognize that there are features of properly identity-affecting decisions that are not faithfully captured by our experiment. We nonetheless hope that the results we sketch below are striking enough to motivate further exploration. We think the end result would be a more complete understanding of the public's normative reasoning in their decision making, which could then be used to better motivate the public to support policies that protect future generations.

In Section 2, we offer a sketch of the non-identity problem, which we understand as a clash between moral intuitions and other common assumptions or judgments made by normative ethicists. In Section 3, we lay out the details of our experiment and list those data we collected that we see as most relevant to the non-identity problem and notions of harm in general. In Section 4, we argue from the data to some provisional conclusions concerning the non-identity problem and its practical implications. In Section 5, we explain the possible relevance of the data to debates over how the public understands the notion of harm. We conclude in Section 6.

2. Background on the Non-Identity Problem

The non-identity problem arises because there are some acts that strike us as intuitively immoral, and yet the acts effectively change which people come into existence. For example, consider the following case, inspired by Parfit (1984):

> The US Government in the year 2025 is split between two very different large-scale social policies. The first policy, CONSERVATION, involves a range of changes in environmental policy (something akin to the so-called "Green New Deal"). The second, DEPLETION involves a business as usual strategy, where resources will be used, carbon will be emitted, and other pollutants will be dispersed into the environment at roughly their current levels. The 2025 US Government ends up choosing DEPLETION, even though CONSERVATION would have required only relatively modest sacrifices to the current generation. In the year 2200, the US population lives in a heavily degraded environment. If the 2025 US Government had instead chosen CONSERVATION, then by 2200 a completely different population would have inhabited an environment roughly similar to the one we enjoy today.

It's intuitively plausible that the US Government's choice was immoral. The initially plausible reason is that the future people under DEPLETION have been wronged by that choice. And the natural reason behind this judgment is that the people within that future population have been harmed by that policy choice, because they were left with a dirty and depleted environment. But, on closer inspection, this is a difficult case to make. The people within the population under DEPLETION weren't made worse off, because, if CONSERVATION had been chosen, a completely different set of people would have existed by 2200. So, given the natural thought that you can't harm someone if you don't make that person worse off, no particular individual within the future population under DEPLETION was harmed by the US Government choosing that policy. So, it looks like no individuals in that population were wronged, and it seems that the choice wasn't immoral after all. Yet this clashes with our strong initial intuition about the case.

What, precisely, is the problem here? As we see it, it involves a clash between the strong initial intuition we tend to have about cases like these and a series of other seemingly reasonable judgments or assumptions. First, we have a strong intuition that the act or policy in question is immoral. Second, we tend to assume that if an act or policy is immoral, this must stem from the fact that the act or policy wrongs someone. Third, we also tend to assume that wronging another person requires doing harm to that person. And, finally, we tend to assume that doing harm to another person requires making that person worse off than they would have been otherwise. But in these non-identity cases, no one is made worse off, and, if our other assumptions are correct, the act couldn't have been immoral in the first place. So, either our initial intuition is wrong, or one of our other reasonable seeming assumptions or judgments must be wrong.

Philosophers have attempted different solutions to the problem by tackling each of these four collectively inconsistent pieces of the puzzle. Taking them in reverse order, some philosophers have argued that harming another doesn't require making that person worse off, e.g., Hanser (1990), Harman (2004, 2009), Rivera-López (2009), and Shiffrin (2009), or that we can otherwise account for how the person/people in the non-identity cases are indeed harmed, e.g., Gardner (2015). Some have attempted to deny that wronging a person requires that you have harmed that person, e.g., Kumar (2003, 2015) and somewhat related contractualist solutions by Finneron-Burns (2016) and Gibbs (2016), as well as Hurley and Weinberg (2015). Some have attempted to explain how the act or policy could be immoral without strictly speaking *wronging* anyone, which was Parfit's own attempted solution (1984, 2017) (see also, e.g., Buchanan et al. 2000 and Steinbock 2009). And some have argued that our strong intuition, i.e., that the act or policy is immoral in much the same way that it would be if the same person or people were affected, is simply mistaken, e.g., Boonin (2014) and Weinberg (2014).

But regardless of which strategy these philosophers prefer, each of them shares the strong *intuition* that non-identity cases are immoral in much the same way as their parallel same-person cases are immoral. Even those like Boonin (2014) who end up arguing that this intuition is mistaken, and that the actions in non-identity cases are not actually immoral, still agree that this intuition is both strong and widespread. So, in a sense, this assumption that the act or policy does not receive moral absolution simply because different people come into existence seems to be a central driving force in the literature on the non-identity problem. If we take these authors at their word, they all share the strong intuition, their colleagues tend to share the intuition (Parfit 1984: 359, 363), and their students tend to share the intuition (Boonin 2014: 25).

As alluded to earlier, we've had a different experience: it is difficult to motivate the non-identity problem as a genuine puzzle because a number of students see an intuitive moral difference once they realize different people will come into existence depending on which choice is made. In other words, it seemed to us that the central intuition that drives the whole literature on the non-identity problem might not be widely shared outside of our (relatively small) community of philosophers. If our hunch is correct, perhaps this could raise some debunking worries for that central intuition, since it could be that philosophers are relying upon intuitions not widely shared in the general public. (For some examples where the intuitions of philosophers seem to differ from the folks, see Tobia, Buckwalter and Stitch (2013) and Machery (2017: chapter 2).) Although we think there is probably something to that line of concern, our worry here is of a more practical nature.

Think back to the choice between CONSERVATION and DEPLETION. Regardless of what analysis any particular philosopher settles upon concerning that case, we assume that this philosopher's intuition tells her that a nation that chooses DEPLETION over CONSERVATION has done something seriously immoral, all else being equal. Now, many philosophers working on the non-identity problem see it as having genuine practical relevance. The idea is that if we can't "solve" the non-identity problem, then we will have a hard time justifying why people ought to support CONSERVATION over DEPLETION. But, from a *motivational* perspective, the solution they seek might be somewhat irrelevant. If the general public shares the intuitions and judgments of the philosophers working on the non-identity problem, then they should also think that the fact that different populations come into existence under each policy doesn't make a moral difference. So if the public's moral motivation in policy choices is predominantly driven by how wrong a policy intuitively seems to them, then we should assume that the public would be just as motivated to support CONSERVATION over DEPLETION after they grasp that the policies change who exists as they were before they had that realization. So, if philosophers' intuitions are fairly typical among the population at large, then the non-identity problem may prove to be a mere theoretical puzzle of little practical relevance. (Although, this is not to say that there won't be other issues regarding climate change distinct from the non-identity problem that are practically relevant; see Gardiner (2012: chapter 5) for various others.)

If, on the other hand, the intuitions of philosophers in non-identity cases *aren't* representative of the population, we are in a very different situation. If the public's moral intuitions and judgments quickly shift once they grasp that they are dealing with a non-identity case, then we should expect that they will be less likely to support policies that involve some sacrifice on their part. This, in turn, would have some implications for how we, philosophers, ought to talk about these kinds of problems with the public. For example, it might be a morally bad idea to broadcast this particular philosophical problem to the public, e.g., in public talks, podcasts, or popular periodicals. (Indeed, it might be a morally bad idea for us to have written this article.) As will become clearer, it also might speak against the seemingly promising strategy of focusing on considerations of harm when attempting to motivate the public to act more conscientiously toward the environment, as suggested by Rottman et al. (2015) among others. If the public's intuitions on this matter are very different than those of philosophers, we might be in real trouble if we both focus our moral motivational efforts on harm considerations and then inform the public about the non-identity problem.

So, do the intuitions and judgments of philosophers mirror those of the general public?

3. Two Experiments

Our experiments are designed with two main goals in mind: to determine whether individuals are less likely to make altruistic sacrifices in identity-affecting choice problems and to examine the role that considerations of harm might play in any changes of behavior when making identity-affecting choices. To satisfy these dual goals, we rely on a mix of traditional survey methods and experimental methods from economics. While philosophers are largely familiar with survey methods, experimental methods from economics have received significantly less attention. (Some exceptions are Bicchieri and Xiao (2009) Bicchieri and Chavez (2010), and Bruner et al. (2018).) And yet, the experimental methods of economics, since they

explicitly were devised to explore behavior, are particularly promising given our goal of identifying the behavioral consequences of identity-affecting considerations. Such experiments proceed by observing how subjects behave in a particular scenario, as opposed to merely reporting how subjects *believe* they would behave. To do this, the subjects' decisions must have real consequences. In typical economics experiments, subjects make decisions with full knowledge that those decisions have *financial* consequences for themselves and other participants.

To best understand whether identity-affecting choices will encourage more self-interested behavior, we ideally would want to place subjects in conditions that approximate the salient features of the actual scenario of interest. This is why methods from experimental economics are so helpful for this kind of examination. Theories with genuine practical implications only rarely involve how subjects *say* they would behave in some environment (i.e. only in those cases where the phenomenon of practical relevance is itself what people tend to utter in such and such circumstances). Instead, they make predictions about how subjects will actually behave. And there's substantial evidence that, in many cases, what agents *say* they will do and what they will *actually* do come apart e.g., FeldmanHall et al. (2012) and Vlaev (2012). Second, and more importantly, methods from experimental economics have an established track-record for effectively probing the other-regarding preferences of individuals. For obvious reasons, we cannot conduct an experiment that forces individuals to make choices that in turn directly cause different individuals to come into existence. Instead, we recreate strategic scenarios that approximate the relevant and salient features of the non-identity cases of interest. Behaviors in this proxy condition provide useful insights. We now turn to descriptions of the experiments themselves.

3.1 Experiment 1 (the Control Study)

Our first experiment, which we are using as a control, is a prerequisite for registering shifts in the behavior and normative attitudes of individuals when confronted by an identity-affecting decision. We then compare these to the results of our second experiment, which we are using as our treatment study, to determine how behavior and normative attitudes are altered when identity-affecting issues become relevant.

The main task of Experiment 1 is the so-called dictator game. The dictator game consists of two individuals: a proposer (Player A) and a recipient (Player B). Player A is given 1.00 USD and must determine how to allocate this amount between herself and Player B. Player A can choose to retain the whole amount for herself. Or she can choose to share with Player B, by transferring to her counterpart any amount up to $1.00, in $0.10 increments. Importantly, the allocation chosen by Player A cannot be contested by Player B (i.e., Player B is unable to veto the proposed allocation, as in the often used "Ultimatum Game").

If subjects only care about maximizing their payoff, we would expect every Player A to pocket the entire bonus. But this is not what economists observe. Often, Player A transfers some non-zero amount to Player B. According to a meta-analysis conducted by Christoph Engel (2011), proposers on average give roughly 28% of the bonus to their counterpart. This is compelling evidence that individuals have other-regarding preferences of some form or another, which makes the dictator game a promising tool for probing the nature and limits of other-regarding preferences.

3.1.1 Experiment 1: Experimental set-up

After receiving approval from the Human Ethics Office at Australian National University (Protocol: 2017/415), we recruited 354 subjects from the US using Amazon Mechanical Turk (MTurk) in the winter of 2017. (We limited the subjects to US participants, because using smaller stakes, like we did, has been shown to have little effect on giving behavior in the dictator game within the US population, but not, for example, within the Indian population; see Raihaini et al. (2013) for details.) Subjects were provided a fixed participation fee of 0.50 USD and told they might earn up to an additional 1.00 USD. Anonymity was guaranteed as no identifying information was collected. Finally, the experiment on average took participants just over three and a half minutes to complete, and subjects were asked some basic demographic questions before completion and payment.

The main task of the experiment was the dictator game. Participants were first provided with a description of the dictator game, and, to ensure comprehension, participants were asked to complete a quick two-question quiz. Failure to correctly answer a comprehension question meant the subject was forced to restart the experiment. Subjects were either assigned to the role of dictator ("Player A", n = 177; male = 107, female = 68, other/decline = 2) or recipient (Player B) and were told they would remain in these roles for the entirety of the experiment. Finally, Player As were all asked to determine how they would like to split the $1.00 bonus between themselves and their randomly chosen Player B counterpart. Player A was informed that this was the only task of the game and was reassured that their identity would not be revealed to their counterpart. The allocation was later revealed to Player B, and this was the only form of contact between the participants.

After the allocation was chosen, or after that allocation was revealed to Player B, each subject was given a compulsory exit survey respectively. In addition to demographic information, we also asked three questions of philosophical relevance. First, Player A participants were asked whether they thought the allocation they gave to Player B harmed Player B, and Player B participants were likewise asked whether they felt the allocation they received from Player A had harmed them. Call these the "specific harm questions." Second, both participants were asked whether they felt that any Player A that gives a $0 allocation to her respective Player B does harm by making that choice. Call this the "generic harm question." Finally, both participants were asked whether they thought the allocation given was 'fair'. Call this the "fairness question." (The exact wording for all questions is available in the Supplementary Materials available on Open Science Framework at https://osf.io/fcpxk/.)

3.1.2: Experiment 1: Results

Player A participants transferred an average of $0.238. As mentioned, a recent meta-analysis of dictator games uncovered an average transfer rate of 28%. This suggests our results are fairly consistent with what has previously been observed in the literature, though a bit more on the selfish side than expected. A total of 56 Player A participants (32%) chose to keep the entire allotment, while 38 (21%) opted for an equal division.

Regarding the generic harm question, 43% of Player A participants and 48% of Player B participants responded that a dictator who transfers zero does, in fact, harm the other player. So, although subjects are divided on whether exceptionally low offers in the dictator game result in a harm to the recipient, a sizable portion of subjects thought that such low transfers do harm. And the fact that there is such a small difference in the attitudes of dictators and recipients on

this matter, even though some of those dictators had just decided to make a zero transfer, is noteworthy. Furthermore, 19% of Player As thought their chosen allocation harmed their counterpart while 31% of Player B participants thought Player A's choice harmed them. We found an even greater disparity between Player As and Player Bs on whether they found Player A's transfer to be fair, with 71% of the dictators reporting that it was fair but only 42% of receivers saying it was fair.

3.2 Experiment 2 (the Treatment Study)

Our second experiment investigates a variant of the standard dictator game designed to add identity-affecting considerations into the choice problem. As in Experiment 1, a dictator (Player A) was allocated $1.00 to divide. However, unlike the previous experiment, there were a total of 11 Player B participants paired with each Player A participant. We refer to these 11 experimental subjects as the Player B Group, and assign each a label (B0, B1, …, B10). Each member of the Player B Group was matched to a particular outcome of the dictator game: Player B0 was matched to the outcome where Player A transfers no funds, Player B1 was matched to the outcome where Player A transfers $0.10 to their counterpart, and so on. Like before, Player A then selected an allocation, which was later revealed to the member of the Player B Group who was pre-determined to receive that amount. Importantly, this revelation was the only form of contact between the Player As and any members of their respective Player B Groups.

Those Player B Group participants that *did not* match the selected allocation were simply paid the $0.50 participation fee and asked to take an unrelated survey. These Player B participants were not provided with any information about the game they had just "participated" in or the behavior of the Player A they were paired with. In other words, all but one of Player B Group participants were kept completely ignorant of the strategic scenario. Importantly, each Player A was told that those members of the Player B Group not matched to the selected allocation would never find out about Player A's behavior. As a result, this variation of the game mimics some of the salient identity-affecting considerations that generate the non-identity problem. Player A's decision not only determines both how much money is transferred but also which individual will be drawn into the strategic scenario. And, importantly, any member of a Player B Group chosen to receive a bonus through this process had no chance of doing any better than she did. If Player A had chosen a more generous allocation, a completely different member of the Player B Group would have reaped that reward. Lastly, those not chosen to receive a bonus in the process will never find out about the decision, just as those who don't come into existence because of some choice will never have the opportunity to find out about that fact.

To make these details a bit more concrete, Figure 1 visually sketches the difference between the choice problem faced by Player A in Experiment 1 and Experiment 2. Hereafter, we will refer to the former as the "standard dictator game" and the latter as the "non-identity dictator game".

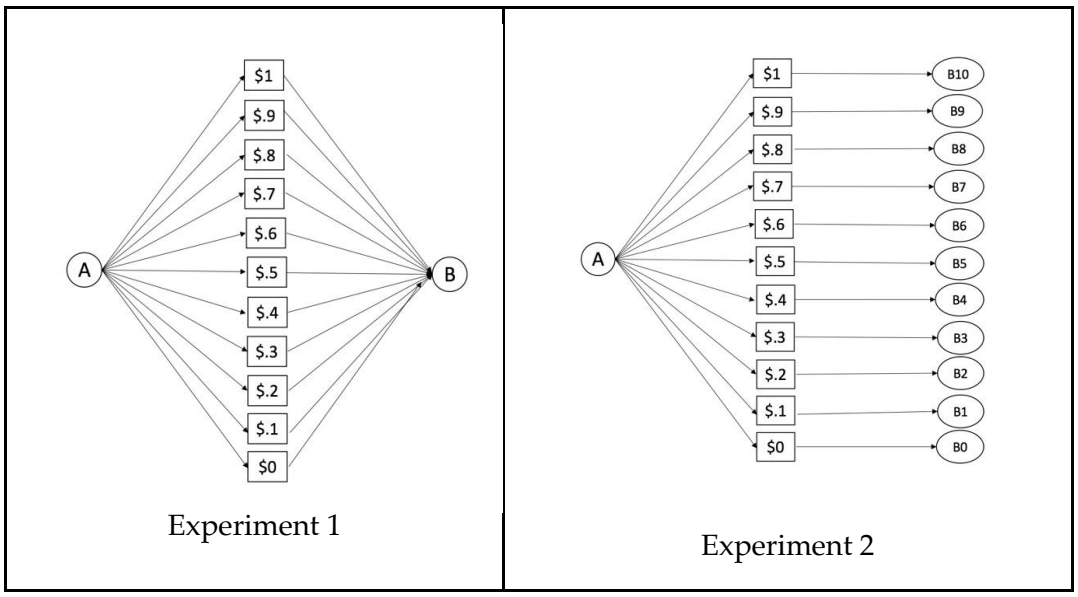

Figure 1: Representation of the choice problem faced by Player A in Experiment 1 and Experiment 2

As noted earlier, we found in Experiment 1 that subjects appeared to be driven at least partially by other-regarding behavior. Our question is whether the identity-affecting nature of our non-identity variant of the dictator game will have a noticeable effect on the behavior of our subjects. We now turn to the experimental set-up and results. We then compare behavior in the standard dictator game to that in the non-identity dictator game.

3.2.1 Experiment 2: Experimental set-up

We again used subjects from MTurk, recruiting a total of 176 Player A participants (male = 100, female = 72, other/decline = 4) and 1936 Player B participants. Once again, our interactions with subjects took place over the online interface, and subject anonymity was guaranteed. All subjects received a $0.50 participation fee upon successful completion of the experiment and were told that they might receive up to an additional $1.00. On average, the experiment took the active subjects an average of just over four minutes to complete.

Player A participants were presented with a detailed description of the non-identity dictator game. They were then administered a three-question quiz to ensure they not only understood how their choice affected their own compensation, but also determined which of the 11 Player B participants would receive the relevant allocation. Participants had to correctly answer all of these comprehension questions correctly before they were allowed to proceed and choose their preferred allocation. Finally, subjects answered a slightly revised set of debriefing questions, and a demographic survey.

As noted earlier, the 11 Player B Group participants were each randomly assigned to one of the 11 possible allocations. The Player B Group participant who corresponded to the allocation chosen by Player A (whom we shall refer to as the 'active Player B') was told about the experimental set-up and was also required to complete the set of comprehension questions. The active Player B was then informed of their predetermined position in the Player B group and that their

position matched Player A's allocation. All of the remaining participants of the Player B Group were simply given an unrelated survey to complete and paid the $0.50 participation fee.

The final task for Player As and active Player Bs was an exit survey, which was once again required for payment. The survey included slightly revised versions of the specific harm, generic harm, and fairness questions from Experiment 1, and the same demographic questions. (See Supplementary Materials for details: https://osf.io/fcpxk/.)

3.2.2 Experiment 2: Results

On average Player As transferred $0.155. A total of 81 Player As (45%) chose to keep the entire endowment for themselves, while only 21 (12%) opted for an equal division. Regarding the responses to the harm/fairness questions, 85% of Player As and 77% of active Player Bs responded negatively to the generic harm question. This indicates that most subjects do not think zero transfers in the non-identity dictator game result in any harm. On the specific harm question, we found that 92% of Player As thought they did not do harm, while 84% of active Player Bs thought the transfer didn't harm them. The latter is rather surprising, given that offers tended to be very low. On the fairness question, 68% of player As and 57% of active Player Bs took the transfer to be fair.

3.3 Comparison of Experiment 1 and Experiment 2

We compare the behavior and attitudes of participants in the standard dictator game and those in the non-identity dictator game. First, Player As were significantly more generous in the standard dictator game than Player As in the non-identity version (see Figure 2). In particular, many more Player As were willing to take an even split in the standard dictator game than in the non-identity version, and many more Player As made exceptionally low transfers in the non-identity dictator game. This clearly suggests that, at least in the aggregate, identity-affecting choice problems tend to generate more self-interested behavior on the part of the dictator. The differences between amounts given in the two studies scored as highly significant (with $p < 0.001$) on every statistical test we ran on the data. (For example, on the full data set, for the Student's t-test $p = 0.00043$, for the Komogorov-Smirnov test $p = 0.00027$, and for the Wilcoxon rank sum test, $p = 0.00027$.) See Figure 2 for the full histogram.

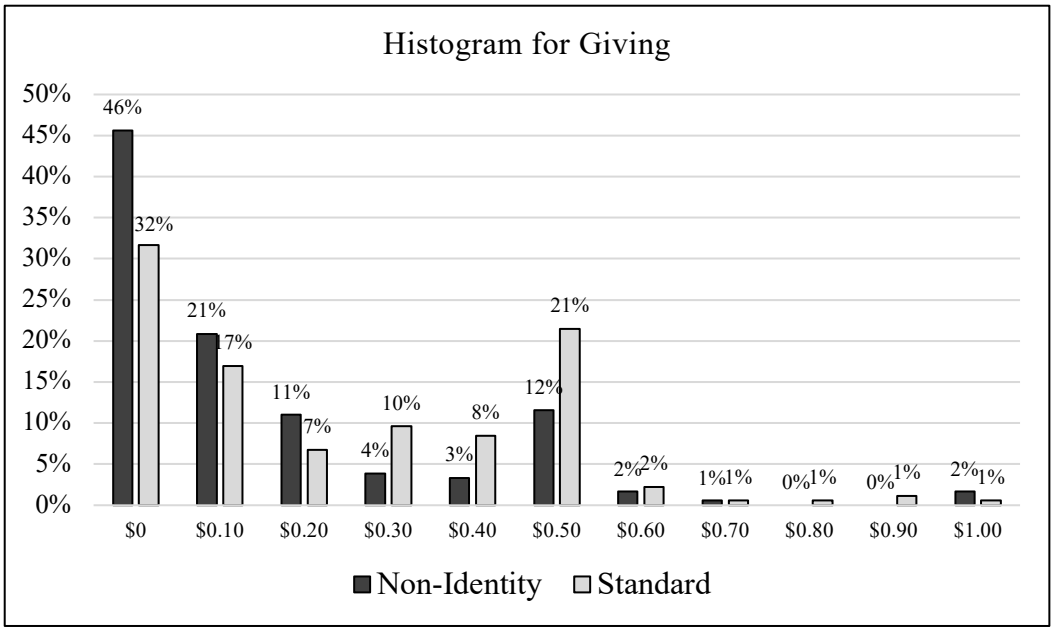

Figure 2: Proportion of Player A transfers in standard and non-identity dictator games

Attitudes about harm were also substantially different between the two experiments. Both Player As and Player Bs were much more likely to register a harm in the standard dictator game than in the non-identity dictator game despite the fact that offers in the non-identity game were much lower than those made in the standard dictator game. As noted, 31% of Player Bs in the standard dictator game felt they were harmed by the specific transfer, but only 16% of active Player Bs in the non-identity version felt harmed by their corresponding Player As' choice. Likewise, substantially more Player As in the standard dictator game felt their behavior resulted in harm than Player As in the non-identity version (19% compared to 8%, respectively).

Figures 3 and 4 provide a more fine-grained look at the harm attitudes of both Player A and Player B participants. In particular, for both experiments we list the proportion of Player A and Player B participants who believed that low transfers (i.e., $0.30 or less) resulted in harm. We find that both players in the non-identity dictator game were much less likely to think that harm was done than their counterparts in the standard version. This difference can be quite substantial.

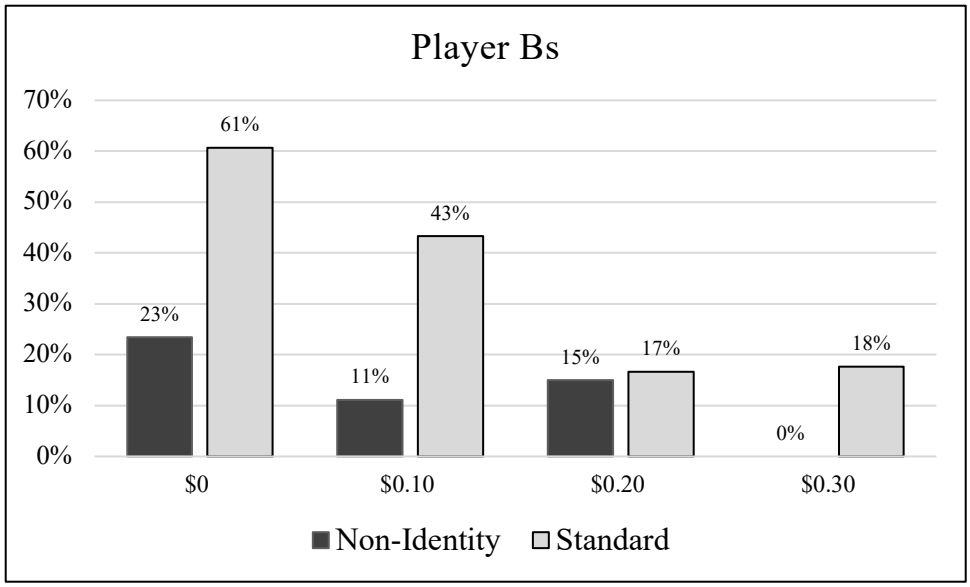

Figure 3: Proportion of Player Bs receiving a low transfer who believed the transfer did them harm

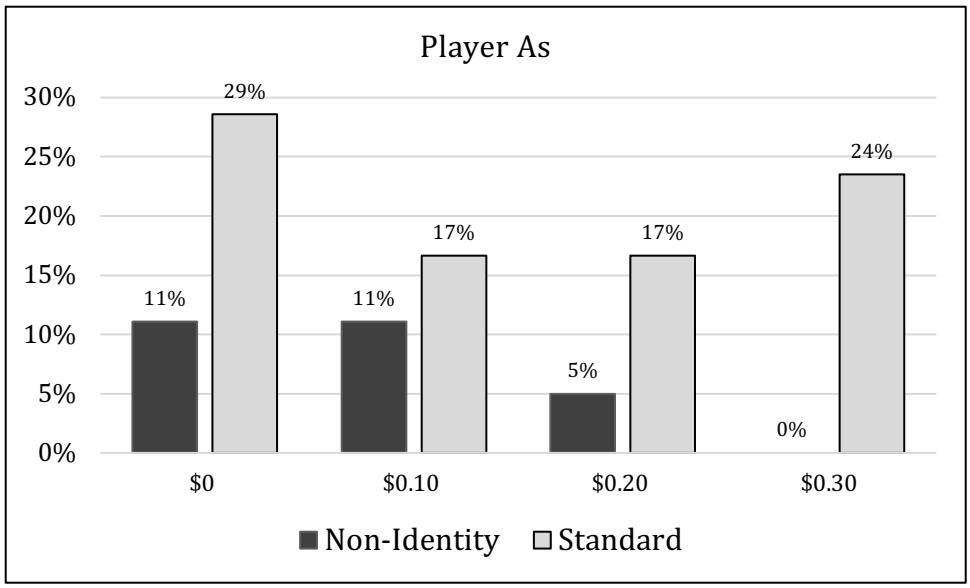

Figure 4: Proportion of Player As giving a low transfer who believed the transfer did the recipient harm

Finally, all subjects in the standard dictator game were more likely than subjects in the non-identity dictator game to agree with the statement that a Player A participant who transfers $0 to her counterpart harms her counterpart (45% vs. 19%; Student's t-test: 7.72, $p < 0.00001$).

4. Practical Implications for the Non-identity Problem

These studies were intended to examine the practical importance of the non-identity problem. Recall that while various authors claim that the non-identity

problem is of great practical importance, this claim is slightly puzzling, because all parties in the debate seem to agree that the paradigm identity-affecting actions are intuitively just as immoral as their analogous same-person actions. Even authors who ultimately argue that such intuitions are mistaken also admit to sharing the intuition. If the public generally shares these intuitions of the philosophers, we should expect that the realization that they are in an identity-affecting choice problem would make little difference to their respective moral evaluations and actions. In other words, if they saw a certain choice as being immoral in the first place, that choice should still seem immoral once they realize it is an identity-affecting choice. There admittedly may still be grave practical roadblocks to generating the motivation to help future generations on this version of the story, but the non-identity problem wouldn't raise any special concerns on that front. And even if the public's intuitions are slightly different from those of philosophers, the public might not generally be able to grasp that a choice is identity-affecting in the first place. Grasping the non-identity problem, after all, requires some rather sophisticated reasoning. And if the public generally cannot grasp that they are in an identity-affecting choice problem, it's unlikely that the identity-affecting nature of the problem will cause any changes in actual behavior.

So what do the data from our study suggest about the practical implications of the non-identity problem? Let's start with the latter point about the public's ability to track the identity-affecting nature of a choice problem. First, if most people's reasoning skills were inadequate, thus making them incapable of tracking whether they are in an identity-affecting choice problem, then we would expect it to be very difficult for participants to complete the comprehension questions in our non-identity dictator game. While we did notice that participants in the non-identity version failed more comprehension questions overall, we only required roughly 33% more trials (622 for Standard vs. 824 for Non-ID) to collect roughly the same number of data points (354 for Standard vs. 352 for Non-ID). And some of this increase in comprehension failures are likely due to the simple fact that we asked an additional comprehension question in the Non-ID version. This suggests that while there probably is indeed some portion of the population that finds such details difficult to follow, it is certainly not a large enough proportion to support the claim that the general public is so naive that we should expect identity-affecting choices to make very little impact on their behavior. And the substantial portion who make it through the comprehension question seem fully capable of seeing the normative difference in an identity-affecting choice problem. For example, the substantial drop in the proportion of those who think that a 0 offer harms the player given 0, which went down to 19% from 45% in the standard version, suggests that these participants are able to track the moral intricacies of the choice problem. (That said, the experiment was only designed to make some of the non-identity features salient to the participant, and we admit that those same individuals could fail to register these features when faced with real non-identity problems.)

Do the data suggest that the identity-affecting nature of a choice problem has an effect on the public's behavior? We believe they clearly do. First, there was a rather pronounced drop in giving behavior, from a mean of $0.238 to $0.155. Although an 8.3 cent drop in giving might not seem so drastic in absolute terms, it is worth noting that *this is a decrease in total giving of roughly one third*. This was paired with a 13 percentage point increase in Player As who kept the whole amount and a 7 percentage point increase in those who kept most but not all of it (i.e., 80 or 90% of the bonus). Even splits decreased by 9 percentage points (from 21% in the standard dictator to 12% in the Non-ID version). If we assume that

moral pressure is what drives altruistic choices in dictator games—a common assumption in the literature—these results suggest that a substantial portion of the population does feel less moral pressure to be altruistic in identity-affecting choice problems.

We must admit that the kinds of effects we saw in our identity-affecting choice problem are merely suggestive of what we might see when dealing with large scale identity-affecting problems like when a nation is deciding on climate change policy. Even beyond any worries one might have about how robust and reproducible our results will prove to be, there are a number of important features of climate policy choices that our scenario doesn't capture. First, there are *severity* differences, since climate policy addresses a very serious problem, and, as such, the relevant outcomes are of a very different sort than small differences in earnings. In particular, the different choices will endow each respective future population with drastically different levels of welfare, and our choice scenario doesn't capture this kind of serious inequality. Second, there are *existence* differences, since our identity-affecting scenario involves choices that differentially affect existing humans, as opposed to choices that change which individuals exist. Put another way, in the non-identity dictator game there is a group of living people who aren't benefited in the way they would have been if a different choice had been made; in a genuine non-identity case the possible future people who aren't benefited don't even exist. Third there are *temporal* differences, since in our game the outcomes of identity-affecting choices are relatively immediate, whereas the outcomes of climate policy choices take generations to eventually bear their identity-affecting results. And there may be shorter-term reasons we could appeal to in order to generate action on climate change, and this kind of alternative path to generating altruism is lacking in our study.

We believe that each of these differences might ground legitimate concerns, and we feel that all of them call out for further study. For example, regarding severity, one could perform a successor study that either raised the monetary stakes or one that introduced actual harms into the study (perhaps by using a paradigm like the one used by Crockett et al. (2014)), or both. For example, if the choice continuum ranged from a $1,000 gift to 10 mild electrical shocks, we may see a much higher level of altruistic behavior in that version of an identity-affecting case. And although we, on balance, prefer our experimental paradigm over the use of surveys that employ vignettes, the latter could yield important insights into whether the severity or inequality of outcomes affects participants' moral intuitions in identity-affecting cases. Regarding existence concerns, it is much more difficult to imagine how our paradigm could be modified to more properly capture the parallels, so vignettes would be needed to further probe those differences. Our hunch is that the mere fact that some of the "affected" individuals don't actually exist isn't really what's doing the work, since the simple fact that those who miss out on the bonus don't ever know they missed out should be enough to get the analogy going. But this is admittedly an open question. And one author has been working on a more recent collaborative project that attempts to tease out whether a phenomenon referred to in the psychological literature as "compassion fade" (see Butts et al. 2019) might be partly responsible for increases in selfish behavior in multiplayer dictator games.[1] Finally, there are various ways one could adjust our paradigm to capture the temporal aspects better, for example

[1] Details on some of those studies will appear in (Grodeck, Handfield and Kopec ms) and through pre-registration on Open Science Framework: https://osf.io/abpd4/ and https://osf.io/uwyqp/. We thank Matthew Lindauer for the lead on Compassion Fade.

by paying out the second players' bonuses at later dates and by varying those durations. We predict these temporal changes would further decrease the level of altruism, as would be expected if participants are temporal discounters, but this is again an open question worthy of further study.

These caveats notwithstanding, we do think what we've uncovered gives us some good initial reasons to be careful in how we choose to motivate actions aimed at future generations. The fact that our subjects in the identity-affecting choice problem were substantially more selfish suggests that the public may be less willing to make even small sacrifices, e.g., modest tax increases or slightly higher prices on certain products, if made aware of the identity-affecting nature of such policy choices. Furthermore, the fact that the increases in selfish behavior appear to correlate with a decreased perception of having harmed anyone gives us a reason to be cautious when we attempt to motivate the population to support conservationist policies by pointing to the harms that inaction could cause over very long timeframes. At least until future studies come along to debunk these possibilities, we feel caution is advised.

5. Implications for Normative Reasoning on Harm

When Parfit (1984) set the stage, the non-identity problem was deemed a problem because of the common-sense notion of harm that's often employed in our moral theorizing. According to this common-sense notion of harm, often referred to as the counterfactual comparative account, to harm someone is to make them worse off than they would have been otherwise. And in a non-identity case, it's simply not true that you've made the relevant person or people worse off than they would have been otherwise. So, if this counterfactual comparative account of harm is correct, then there's no legitimate sense in which you've harmed the person in a non-identity case. This gets the whole problem rolling. As noted in Section 2, one way to solve the non-identity problem would thus be to devise a generally acceptable account of harm that entails that the affected individuals in non-identity cases are actually being harmed after all. So various authors have seen the non-identity problem predominately as a problem for the counterfactual comparative account of harm, and thus as a reason to seek out alternative accounts of harm (e.g., Harman 2004, 2009, Shiffrin 2009, Gardner 2015). Given the parallels, the choice problem we've generated in our game is well situated to offer some insight into which notions of harm members of the public are employing in their normative reasoning, at least at the moment of making their choice. So here we will look at some of the competing accounts of harm that have been proposed in the literature and examine what the data we've collected could tell us about which notions of harm might be affecting the public's behavior.

First, it helps to have a precise statement of what those like Parfit take to be the common-sense account of harm:

> Counterfactual Comparative Account: A's act harms B if and only if A's act makes B worse off than B would have been otherwise.

The main competitor for the counterfactual comparative account in the literature is typically referred to as the non-comparative account of harm, which we could state as:

Non-comparative Account: A's act harms B if and only if A's act causes B to be in an intrinsically bad state.[2]

To give an example of how these two accounts of harm can come apart, think back to the environmental policy example from earlier. If the US adopts DEPLETION the US population in 2200 lives in relative squalor. On the counterfactual comparative account, the current US generation hasn't harmed the citizens of 2200, since it's not the case that these future individuals would have been better off if CONSERVATION were adopted instead. But the non-comparative account has a very different verdict. By selecting the policy DEPLETION, the current US generation has caused the citizens of 2200 to be in an intrinsically bad state. So, on the non-comparative account, the 2025 US Government did indeed harm the US population in 2200.

Although these are the two main accounts of harm in the literature, there is one other account that is worth mentioning. This account, which is often called the temporal comparative account, could be stated as:

Temporal Comparative Account: A's act harms B if and only if A's act makes B worse off than B was before A's act occurred.

Philosophers tend to quickly discredit this account as a legitimate notion of harm, due to the various supposed counterexamples that are fairly easy to devise, e.g., (Hanser 2008), (Thomson 2011) and (Shiffrin 1999). (Although see (Foddy 2014) for a defense.) But since we are interested in which notions of harm play a role in the normative reasoning of the public, philosophers' counterexamples don't give us a reason to toss out the concept by fiat. It's now well established that the public uses various deeply problematic concepts in their normative thinking. So here we'll treat the temporal comparative count as a live alternative.

So what can the data from our study tell us about which notions of harm the public employ? Let's start with our subjects' attitudes about harm in the standard dictator game. In the standard version, there is no sense in which Player B is put into an intrinsically bad state when Player A takes all of the bonus money for herself. After all, Player B will still get the show up fee. So if the public predominantly employed a non-comparative notion of harm in their normative thinking, we would expect a very low percentage of our subjects to think that a 0 transfer harms Player B. We found quite the contrary. Since 45% of our subjects in the standard dictator game believed that a 0 transfer harmed Player B, this suggests that a rather substantial portion of the population employs something like the counterfactual comparative notion of harm in their normative thinking, at least when faced with certain kinds of choice problems.

Notice further that one way to frame what is going on when Player A takes the whole bonus for herself is that she is failing to benefit the Player B she was paired with. Many philosophers have a strong intuition that simply failing to benefit someone is very different from harming that person, and this intuition has been used as a motivation to reject the counterfactual comparative account, since it tends to conflate the two. For example, Bradley (2012), who calls this the "problem of omission", rejects the counterfactual account on just these grounds. (Although see (Feit 2019) for critique of Bradley's argument.) It turns out that a

[2] For clarity, we should note that Harman (2004, 2009), who is typically associated with non-comparative accounts of harm, would likely not endorse the formulation we give here, since she doesn't require that placing someone into an intrinsically bad state is a necessary condition for harm.

substantial portion of the population doesn't seem to share this intuition, since they are happy to diagnose something that is a clear case of failing to benefit as an instance of harming. This, perhaps, suggests that we should proceed with some caution before rejecting the counterfactual comparative account of harm based solely on the problem of omission, although here we don't intend to take a stand on which notion of harm is the "correct" one.

The substantial drop in harm perceptions when we move to the non-identity dictator game further supports the claim that a substantial portion of the population employs a counterfactual comparative notion of harm. Recall that a much smaller proportion of Player As believe they have harmed the member of the Player B Group with their choice of transfer, even though the offers were on the whole much lower. And we also saw a sizable drop in the proportion of subjects that thought that a 0 transfer does harm. Both of these facts further support the claim that something like the counterfactual comparative notion of harm is playing a role in these subjects' normative thinking, at least in their assessments of this particular problem.

But we think the data also suggest that an even larger portion of the population employs some other notion of harm when assessing these kinds of choices. For example, 55% of subjects in our standard dictator game didn't think that Player A did harm by taking the whole bonus, which you wouldn't expect if these subjects employed a counterfactual comparative notion. So, one thing the data from our study show is that there are substantial limits on what portion of the population could be using the various notions of harm in their normative thinking. For example, the data suggest that it's very unlikely that in future studies researchers will find that the vast majority of the population uses a non-comparative notion of harm in their normative thinking, and similarly for the counterfactual comparative notion. It's much more likely that we'll find substantial heterogeneity in the population. It also would be fully consistent with our data if the notion of harm the public employs didn't even play the role of a singular criterion concept in the first place. Perhaps harm, in the public's thinking, is a much messier kind of concept, that gloms onto different features of cases depending on context.[3] Either way, it is simply unlikely that we'd find one of the singular criterion notions preferred by philosophers dominated the public's thinking.

Much like before, our study came with various limitations for probing the public's ideas of harm, and there are important unanswered questions that call out for further study. For example, even if the public did think of harm as a singular criterion concept, our data wouldn't distinguish which of the alternative accounts various portions of the population employ. While it's true that a Player B who is given no bonus is not caused to be in an intrinsically bad state, it's also true that she isn't made worse off than she was before she was given no bonus. After all, she started out with no bonus. Thus, this portion of our subjects could just as well be utilizing a temporal comparative notion of harm as a non-comparative one.[4] In future work, we plan to ask questions to help us to tease apart which of these

[3] We thank an anonymous referee for this journal for suggesting this possibility.

[4] David Boonin, in correspondence, suggests another competitor, his students seem to employ: a kind of morally laden counterfactual comparative account, where an act harms if and only if it *wrongfully* makes an individual worse off than she would have been otherwise. Our data likely also don't distinguish between this competitor and the temporal or non-comparative account. Subjects might admit that a 0 transfer makes Player B worse off than she would have been otherwise, and yet not think that it does so *wrongfully*.

competitors this substantial portion of the population are using. Additionally, a non-negligible portion of our subjects in the non-identity dictator game still thought that a Player A who keeps the whole bonus harms Player B0 by making that choice (19%). We wouldn't really expect this on any of the notions of harm covered here (including Boonin's contender mentioned in footnote 4). It could very well be that there is a notion of harm that is playing a role in the public's normative thinking that philosophers have thus far completely ignored. And, as noted above, intuitive instances of harm might get triggered by a multiplicity of different considerations depending on the specific context the individual encounters. Further study is needed to tease all of this apart.

6. Conclusion

In this paper, we have tried to show that the non-identity problem likely does pose a potentially serious obstacle to moral motivations on problems like climate change. In particular, agents tend to act more selfishly when they find themselves in identity-affecting choice problems, where they seem less willing to make small, altruistic sacrifices. If the kind of behavior we uncovered scales up to large scale policy choices, then this would give us a good reason to take steps to limit the damage done by the identity-affecting considerations. We admit that the issue of whether the behavior is likely to scale up is largely an empirical question—our studies are only intended to be suggestive of the problem that concerns us. But we hope that the behavior we have uncovered is striking enough to motivate a further examination of how identity-affecting considerations can influence how real people reason, decide, and behave when facing these kinds of moral problems. Similarly, we hope that the diverse, even counterintuitive, attitudes about harm that we've uncovered in the process will motivate future research projects that will in turn give us a clearer picture of the role harm considerations play in our everyday normative thinking.

Acknowledgements: We thank David Boonin, Ben Grodeck, Toby Handfield, Duncan Purves, and audiences at CU - Boulder, Northeastern University, National University of Singapore, the 6th Australasian Workshop in Moral Philosophy at Kioloa, and the Workshop on Experimental Philosophy & Normativity at ANU for helpful comments, questions, and suggestions. We thank Matthew Lindauer for help with ethics approval and Lachlan Walmsley for the conversation that generated the idea for the project and for some early contributions. Financial support was provided by ANU's School of Politics and International Relations, School of Philosophy, and College of Arts and Social Sciences, and Kopec's ARC DECRA Grant (DE180101119). We apologize to anyone we've inadvertently omitted.

REFERENCES

Andreoni, James, William T. Harbaugh, and Lise Vesterlund. (2008) 'Altruism in experiments'. in Steven N. Durlauf and Lawrence Blume (eds.), *The New Palgrave Dictionary of Economics*. Palgrave Macmillin.

Bicchieri, Cristina, and Erte Xiao. (2009) 'Do the right thing: But only if others do so'. *Journal of Behavioral Decision Making*, 22, 191–208.

Boonin, David. (2008) 'How to solve the non-identity problem'. *Public Affairs Quarterly*, 22, 129–59.

Boonin, David. (2014) *The Non-Identity Problem and the Ethics of Future People*. Oxford University Press.

Bradley, Ben. (2012) 'Doing away with harm'. *Philosophy and Phenomenological Research*, 85, 390-412.

Broome, John. (2018) 'Efficiency and future generations'. *Economics & Philosophy*, 34, 221–41.

Bruner, Justin P., Cailin O'Connor, Hannah Rubin, and Simon Huttegger. (2018) 'David Lewis in the lab: Experimental results on the emergence of meaning'. *Synthese*, 195, 603-21.

Buchanan, Allen, Dan W. Brock, Norman Daniels, and Daniel Wikler. (2000) *From Chance to Choice: Genetics and Justice.* Cambridge University Press.

Butts, Marcus M., Devin C. Lunt, Traci L. Freling, and Allison S. Gabriel. (2019) 'Helping One or Helping Many? A Theoretical Integration and Meta-Analytic Review of the Compassion Fade Literature'. *Organizational Behavior and Human Decision Processes*, 151, 16–33.

Crockett, Molly J. et al. (2014) 'Harm to others outweighs harm to self in moral decision making'. *Proceedings of the National Academy of Sciences*, 111, 17320–25.

Engel, Christoph. (2011) 'Dictator games: A meta study'. *Experimental Economics*, 14, 583-610.

Fehr, Ernst and Klaus M. Schmidt. (1999) 'A theory of fairness, competition and cooperation'. *The Quarterly Journal of Economics*, 114, 817-68.

FeldmanHall, Oriel et al. (2012) 'What We Say and What We Do: The Relationship between Real and Hypothetical Moral Choices'. *Cognition*, 123, 434–41.

Feit, Neil. (2019) 'Harming by failing to benefit'. *Ethical Theory and Moral Practice*, 22, 809-23.

Finneron-Burns, Elizabeth. (2016) 'Contractualism and the Non-Identity Problem'. *Ethical Theory and Moral Practice*, 19, 1151–63.

Foddy, Bennett. (2014) 'In defense of a temporal account of harm and benefit'. *American Philosophical Quarterly*, 51, 155–65.

Gardiner, Stephen M. 2012. *The Perfect Moral Storm*. Oxford University Press.

Gardner, Molly. (2015) 'A harm based solution to the non-identity problem'. *Ergo*, 2, 427-444.

Gibb, Michael. (2016) 'Relational Contractualism and Future Persons'. *Journal of Moral Philosophy*, 13, 135–60.

Grodeck, Ben, Toby Handfield, and Matthew Kopec. (ms.) 'Framing Effects and Selfish Choices: An Examination of Compassion Fade and Excuse Seeking in Multiplayer Dictator Games'.

Hanser, Matthew. (1990) 'Harming future people'. *Philosophy & Public Affairs*, 19, 47–70

Hanser, Matthew. (2008) 'The metaphysics of harm'. *Philosophy and Phenomenological Research*, 77, 421-50.

Harman, Elizabeth. (2004) 'Can we harm and benefit in creating?'. *Philosophical Perspectives*, 18, 89-113.

Harman, Elizabeth. (2009) 'Harming as causing harm'. In Melinda A. Roberts and David T. Wasserman (eds.), *Harming Future Persons: Ethics, Genetics and the Nonidentity Problem* (Springer), pp. 137-54.

Hurley, Paul and Rivka Weinberg. (2015) 'Whose problem is non-identity?'. *Journal of Moral Philosophy*, 12, 699-730.

Kolstad, Charles et al. 2014. 'Chapter 3: Social, economic and ethical concepts and methods'. In Ottmar Edenhofer (ed.), *Climate Change 2014: Mitigation of Climate Change (Vol. 3),* (Cambridge University Press), pp. 207-82.

Kumar, Rahul. (2003) 'Who can be wronged?'. *Philosophy and Public Affairs*, 31, 99-118.

Kumar, Rahul. (2015) 'Risking and wronging'. *Philosophy and Public Affairs*, 43, 27-51.

List, John. (2007) 'On the interpretation of dictator giving'. *Journal of Political Economy*, 115, 482-93.

Machery, Eduard. (2017) *Philosophy within its proper bounds*. Oxford University Press.

Parfit, Derek. (1984) *Reasons and Persons.* Oxford University Press.

Parfit, Derek. (2011) *On What Matters, vol. 2.* Oxford University Press.

Parfit, Derek. (2017) 'Future People, the Non-Identity Problem, and Person-Affecting Principles'. *Philosophy and Public Affairs,* 45, 118-57.

Raihani, Nichola J., Ruth Mace and Shakti Lamba. (2013) 'The effects of $1, $5 and $10 stakes in an online dictator game'. *PLOS One*, 8, 1-6.

Rivera-López, Eduardo. (2009) 'Individual procreative responsibility and the non-identity problem'. *Pacific Philosophical Quarterly*, 90, 336–63.

Rottman, Joshua, Deborah Keleman and Liane Young. (2015) 'Hindering harm and preserving purity: How can moral psychology save the planet?'. *Philosophy Compass,* 10, 134-44.

Shiffrin, Seana V. (2009) 'Wrongful life, procreative responsibility, and the significance of harm'. *Legal Theory*, 5, 117-48.

Steinbock, Bonnie. (2009) 'Wrongful life in procreative decisions'. In Melinda A. Roberts and David T. Wasserman (eds.), *Harming Future Persons: Ethics Genetics and the Non-identity Problem.* (Springer), pp. 155-78.

Thomson, Judith Jarvis. (2011) 'More on the metaphysics of harm'. *Philosophy and Phenomenological Research*, 82, 436-58.

Tobia, Kevin, Wesley Buckwalter and Stephen Stitch. (2013) 'Moral intuitions: are philosophers experts?'. *Philosophical Psychology*, 26, 629-38.

Vlaev, Ivo. (2012) 'How Different Are Real and Hypothetical Decisions? Overestimation, Contrast and Assimilation in Social Interaction'. *Journal of Economic Psychology*, 33, 963–72.

Weinberg, Justin. (2014) 'Non-identity matters, sometimes'. *Utilitas*, 26, 23-33.

Wollard, Fiona. (2012) 'Have we solved the non-identity problem?'. *Ethical Theory and Moral Practice*, 15, 677-690.