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

Matthew Kopec School of Philosophy, Australian National University
Justin P. Bruner Department of Theoretical Philosophy, University of Groningen

Abstract: A driving force behind much of the literature on the non-identity problem is the widely shared intuition that actions or policies that change who comes into existence don't, as a result, lose their morally problematic features. We hypothesize that this intuition isn't entirely shared by the general public, which might have widespread implications concerning how to best motivate public support for large-scale, identity-affecting policies like those involved in climate change mitigation. To test our hypothesis, we ran a behavioural economic experiment, a version of the well-known dictator game, designed to mimic the public's morally loaded behaviour in identity-affecting choice problems. As predicted, we found that the public does seem to behave more selfishly when making identity-affecting choices. We further hypothesised that one possible mechanism involved in this change is the notion of harm that plays a role in the public's normatively loaded decision making. So, during our study, we also solicited subjects' attitudes about harm, in particular about whether the "dictators" had done harm through their choices. The data suggest that substantial portions of the population each employ distinct notions of harm in their normative thinking, which raises some puzzling features about the public's normative thinking that call out for further empirical examination.

1. Introduction

If we want human life to continue on this planet into the foreseeable future, then we need to find a way to motivate the public to care more deeply about the welfare of people who don't yet exist. A sizeable portion of the population seems largely indifferent about what kind of world we leave to those future people, and, as the 2016 US election made clear, this is currently a large enough group to erase the fleeting progress made on matters like climate change mitigation. Although there have been some recent attempts to fashion conservationist arguments that could motivate even Homo economicus (e.g., Broome 2018), the vast majority of environmental economists agree that solving the climate crisis, and conserving our resources more generally, comes at a serious cost to the current generation. So it seems our only viable option, assuming we maintain our democracies and refrain from brainwashing, is *moral* motivation. We need to find a better way to get the public to feel a moral obligation to leave a healthy planet to those people who don't yet exist.

Unfortunately, our obligations on these kinds of problems, where people don't yet exist, are more slippery than most. Our choices between large scale policies will, over a long enough period of time, change which people end up coming into existence. It

thus becomes hard to make the argument that we are leaving an unhealthy planet to *these* people. After all, they wouldn't have existed if we had chosen some other policy. This moral quirk, usually referred to as the non-identity problem (following Parfit 1984), is now widely recognized as a serious philosophical problem—one we must solve if we are to give a proper account of our duties to future generations. And environmental ethicists like Gardiner (2012) tend to see this problem as a key roadblock in generating action on climate change (i.e., as a key contributor to what he calls the "intergenerational storm"). Even the recent IPCC 2014 report mentions the non-identity problem as a key theoretical roadblock to action on climate change (Kolstad et al. 2014, p.216).

But does the non-identity problem really make a practical difference when it comes to motivating the public to support policies which treat future generations responsibly? One might think the obvious answer is "No", since the public doesn't generally know about the problem. And even if they knew the rough details, they might not be able to really grasp the issue. But if we take philosophers' intuitions as a general guide to moral intuitions and judgements more generally, there is a deeper reason for scepticism about the practical relevance of the non-identity problem. A central driving force in the literature on the non-identity problem 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. In other words, an identity-affecting choice *seems* just as wrong as the parallel choice that doesn'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 (see e.g., Boonin 2014). So, the problem doesn't generally move philosophers into thinking we are morally off the hook when our choices affect who comes into existence. Why think the general public should be moved any differently? If the intuitions of the public match the intuitions of these philosophers, the non-identity problem raises no *special* problem for moral motivation after all.

In this project, we set out to see if this is really the case. Our experience from teaching the non-identity problem actually clashed with what is generally accepted in the literature: 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. One of us found it rather difficult to convince the students that the non-identity problem really was a problem, because many students seem to see a substantial moral difference between the cases meant to motivate the problem. Thus, they didn't see why this was a puzzle worth theorizing about in the first place. Because of this, we worry that the optimistic story told above is mistaken, in which case the non-identity problem may indeed pose a special problem for moral motivation. In order to see whether this is the case, we developed a behavioural economic experiment, a version of the well-known dictator game, which was designed to elicit the public's behaviour in identity-affecting choice problems. 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 straightforwardly "harmed". Getting such a study past the university's ethics board would be tricky indeed! That said, what we found should strike many as rather surprising nonetheless.

Not only is a large portion of the population fully able to follow the details of identity-affecting choice problems, the public is also much more selfish when confronted with

such choices. One possible explanation for this change is that a substantial portion of the public employs, in their normative thinking, a version of what has been called the counterfactual comparative account of harm. 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 parallel this feature of identity-affecting choice problems, we were also able to probe the relationship between giving behaviour and judgments of harm. What we found suggests that something like the counterfactual comparative notion of harm does indeed play a role in the public's normative thinking, and the role it plays does seem to have some effect on giving behaviour. That said, there actually seems to be a substantial split within the public over which notion of harm to employ when making normative evaluations, which calls out for further empirical examination.

We should pause to admit, from the outset, that there are a number of limitations of the studies we ran. In particular, we obviously aren't recreating an identity-affecting choice problem per se, and there are a number questions we could have asked that would have allowed us to probe the public's normative thinking more deeply. But we hope that the results we sketch below are striking enough to motivate others to run their own experiments along similar lines. 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.

Our plan is as follows. 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 judgements made by normative ethicists. In Section 3, we lay out the details of our experiment and list the data we found that are 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 the proper understanding of harm. In Section 6, we discuss objections, some of which we must concede will require further study to fully address. We conclude in Section 7.

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. Take the following case:

> Wilma is interested in having a baby. Wilma's doctor tells her that she has a condition such that if she conceives now, any child she conceives will suffer from incurable blindness. However, her doctor also tells her that this result is not unavoidable. If Wilma waits to conceive, and instead takes a pill every day for two months prior to conceiving, then she will conceive a child who is not afflicted with incurable blindness. Had Wilma waited and taken the pill, she would have conceived and given birth to a perfectly sighted boy she would have named 'Rocks.' However, she decides not to take the pill in favor of

> conceiving immediately. As a result, she conceives and gives birth to an incurably blind baby girl. She names this child 'Pebbles.' (Purves 2014)[1]

Intuitively, Wilma does something immoral by choosing not to wait to conceive. The initially plausible reason her act is immoral is because she has done something wrong to Pebbles, who is born blind. And the very natural reason this act seems to wrong Pebbles is because Wilma seems to have harmed Pebbles by causing Pebbles to be born blind. But on closer inspection, it's difficult to really claim that Wilma has harmed Pebbles at all. After all, if Wilma had decided to wait to conceive, Pebbles wouldn't have been better off. Pebbles wouldn't have *been* at all—Rocks would have been born instead. Given the natural thought that you can't harm someone if you haven't made her worse off, Pebbles hasn't been harmed. So, it seems Pebbles hasn't been wronged, and Wilma obviously hasn't wronged anyone else either. So, it seems Wilma hasn't done anything immoral after all. A number of very natural thoughts lead to a clash with the strong intuition that Wilma has acted immorally in this case.

More apropos of our own motivations in undertaking this project, take another kind of case, inspired by one originally sketched by Parfit (1984):

> The US Government in the year 2020 is split between two very different large-scale social policies. Policy C (think "Conservation") involves a range of changes in environmental policy, including: sizeable tax increases on synthetic fertilizers, plastic by-products, carbon emissions, and automobiles and gasoline; substantial spending increases on mass transit and carbon neutral power generation; and zoning changes around population centres to encourage people to live closer to their places of work. Policy D (think "Depletion") involves a business as usual strategy, where resources will be depleted, carbon will be emitted, and other pollutants will be dispersed into the environment at roughly their current levels. The 2020 US Government ends up choosing Policy D, even though Policy C would have required only modest sacrifices to the current generation. In the year 2220, the US population lives in a heavily degraded environment. If the 2020 US Government had instead chosen Policy C, then by 2220 a completely different population would have inhabited an environment roughly similar to the one we enjoy today.

Much like in the previous case, intuitively, the US Government's choice was immoral. The initially plausible reason is that the people within the future population under Policy D has been wronged by that choice. And the natural reason behind this judgement is that the people within that future population have been harmed by Policy D, because they were left with a dirty and depleted environment to live in. But, on closer inspection, this is also a hard case to make. The people within the population under Policy D weren't made worse off, because, if Policy C had been chosen, a completely different set of people would have existed by 2220. So, given the natural thought that you can't harm someone if you don't make them worse off, no particular individual within the future population under Policy D was harmed by the US Government choosing that policy. So, it looks like no individuals in that population

[1] This is Purves's succinct restatement of an example by Boonin (2008), (2014), which was in turn a revised version of the original example given by Parfit (1984).

were wronged. So, it seems that the choice wasn't immoral after all. And 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 judgements 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 were 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 judgements 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 in the way we typically think, e.g., Hanser (1990), Meyer (2003), 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, 2009, 2015)[2] and 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) (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), Heyd (2009), 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, pp.359,363), and their students tend to share the intuition (Boonin 2014, p.25).

As alluded to earlier, we've had a somewhat different experience, at least when it comes to non-philosophers like our students. We have found it somewhat difficult to motivate the non-identity problem as a genuine puzzle worth our concern, because a number of students see an intutive moral difference once they realize different people

[2] See Finneron-Burns (2016) and Gibbs (2016) on somewhat related contractualist solutions to the non-identity problem.

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.[3] 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 Policy C and Policy D. Regardless of what analysis any particular philosopher settles upon concerning that case, we can assume that this philosopher's intuition tells her that a nation that chooses Policy D over Policy C has done something seriously immoral. Perhaps her considered judgement will reject that assessment in the end. Nonetheless, if she shares the intuition that drives the non-identity literature, then she at least initially judges Policy D as the wrong choice. Now, many philosophers working on the non-identity problem see it as having a 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 policies like Policy C over Policy D. But, from a motivational perspective, the solution they seek might be rather irrelevant. If the general public generally shares the intuitions and judgements 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. The public, in short, should be just as motivated to support Policy C over Policy D 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.[4]

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 judgements quickly shift once they grasp that they are dealing with a non-identity case, then we should expect substantial differences in their behaviour. In particular, we should expect that they will be less likely to support policies that involve some sacrifice on their part once they realize that the choice between policies leaves them in a non-identity case. 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. on YouTube, podcasts or popular periodicals, or to thinkers in other fields who might pass it along to the public second hand. (Indeed, it might be a morally bad idea for us to have written this article.) As our data

[3] There is now a substantial empirical literature examining whether the intuitions of the folk come apart from the intuitions of philosophers, possibly in systematic ways. Although the evidence is by no means decisive, Tobia, Buckwalter and Stitch (2013) have found that philosophers and non-philosophers have different moral intuitions. Machery (2017, chapter 2) provides a helpful overview of empirical work done exploring the impact that education and socioeconomic factors have on philosophical judgment. More generally, the thought that philosophers systematically have different intuitions underlies the so-called 'expertise defence' of traditional armchair philosophy (Nado 2014).

[4] This is not to say that there won't be other issues in intergenerational ethics distinct from the non-identity problem that are practically relevant, as opposed to mere theoretical puzzles. See Gardiner (2012, chapter 5) for an argument to this effect.

will make clearer in what follows, it also suggests that it might speak against the seemingly promising strategy of focusing on considerations of harm when attempting to motivate the public to act more environmentally conscientiously , as suggested by Rottman et al. (2015). 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 judgements of philosophers mirror those of the general public?

3. Two Experiments

Our experiments are designed with two main goals in mind. First we aim to determine whether individuals are less likely to make altruistic sacrifices in identity-affecting choice problems. Second, we aim to examine the role that considerations of harm might play in any changes of behaviour when making identity-affecting choices. To satisfy the dual goals of examining the changes in behaviour as well as the role that normative attitudes might have played in such changes, we rely on a mix of traditional survey methods and experimental methods from behavioural economics. In this section, we provide some background on our tools of choice before describing the experimental set-up.

Although philosophers are now largely familiar with the survey methods commonly used in experimental philosophy, experimental methods from behavioural economics have received significantly less use in the philosophical literature.[5] And yet, the experimental methods of economics, since they explicitly were devised to explore behaviour, are particularly promising given our goal of identifying the behavioural consequences of identity-affecting considerations.[6] Experiments in economics tend to proceed by observing how subjects actually behave in a particular scenario, as opposed to merely reporting how subjects *believe* they would behave, as is common in other fields like psychology. In order to observe actual behaviours, it is necessary to construct experiments where the subjects' decisions have real consequences. In typical economics experiments, subjects make decisions with full knowledge that those decisions will have *financial* consequences for themselves and, in some cases, other participants.

We believe that the inclusion of methods from experimental economics carries a number of benefits. First, most theories that have practical implications don't make predictions about how subjects *say* they would behave in some environment, but instead make predictions about how subjects will *actually behave*. As a result, experiments that primarily rely on responses to vignettes or hypothetical scenarios can only be used to assess such theories in very special circumstances (i.e., only in cases

[5] Although, see, e.g., Bicchieri and Xiao (2009), Bicchieri and Chavez (2010), and Bruner et al. (2018).

[6] In what follows, we will often drop the qualifier 'behavioural' in 'behavioural economics', simply to avoid redundancy. This is not to suggest that the only experiments that economists run have to do with overt behaviour, or that behavioural economics is the only sub-field of economics where experiments are run. Such suggestions would be false on both counts.

where what subjects report they would do closely tracks what they actually would do). Second, and more importantly for us, methods from experimental economics already have an established track record for effectively probing the social preferences of individuals. For example, such methods have been used to register the extent to which individuals are driven by an aversion to inequitable outcomes, such as Fehr and Schmidt (1999), and the extent to which individuals act out of self-interest as opposed to acting out of a concern for others.[7] Tracking a subject's predictions about how they would behave in a hypothetical scenario is not a particularly reliable means of registering such preferences, since subjects have little reason not to present themselves as being more caring and altruistic than they really are.

As mentioned, our goal is in part to determine whether individuals behave differently when tasked to make identity-affecting decisions. To best understand whether this will lead to, say, more self-regarding behaviour, we need to place subjects in conditions that approximate the salient features of the actual scenario of interest. 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, as occurs in true non-identity cases. That said, we can recreate strategic scenarios that, we believe, approximate the relevant and salient features of the original scenarios of interest. Behaviour in this proxy condition will provide us with useful insights. We now turn to descriptions of the experiments themselves.

3.1 Experiment 1 (the "Standard Dictator Game")

Our first experiment allows us to register the behaviour and normative attitudes of individuals when confronted by a non-identity-affecting decision. We then compare these to the results of our second experiment to determine how behaviour and normative attitudes are altered when identity-affecting issues become relevant. (We focus on subjects' attitudes regarding harm and fairness in both conditions.)

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 a fixed amount of money (in this case 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 of $1.00 for herself. Or she can choose to share with Player B, by transferring to her counterpart any amount she pleases up to $1.00, in $0.10 increments. Importantly, the allocation chosen by Player A cannot be contested by Player B. Player B is unable to protest or veto the proposed allocation. In this sense, Player A is "the dictator," as their decision is final. (We occasionally refer to Player A as such in what follows.)

If we were to assume that subjects only care about money , we would expect every Player A to pocket the entire bonus, leaving nothing for their corresponding Player Bs. But this is very far from what behavioural economists observe when these games are run in the laboratory. More often than not, Player A transfers some non-zero amount of the bonus to Player B. In fact, transfers can be quite generous. According to a meta-analysis conducted by Christoph Engel (Engel, 2011) that draws on over 100

[7] See Andreoni, Harbaugh and Vesterlund (2008) for an overview of experimental work on altruism.

experiments, those in the role of proposer on average leave a total of 28% of the total 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

We recruited 354 subjects from the US using Amazon Mechanical Turk in the northern winter of 2017.[8] Subjects were provided a fixed participation fee of 0.50 USD and told they could earn up to an additional 1.00 USD. Anonymity was guaranteed as no identifying information was released in the course of the experiment. Finally, the experiment on average took participants over three and a half minutes to complete and subjects were asked some basic demographic questions at the end of the experiment.

The main task of the experiment was the dictator game. Participants were first provided with a description of the dictator game. To ensure comprehension, participants were then asked to complete a quick two question quiz about the dictator game. Those who failed the comprehension check were forced to restart the experiment from the beginning. Subjects were either assigned to the role of dictator (Player A) 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, but this revelation was the only form of contact between the Player As and their respective Player Bs in the experiment.

After Player A chose the allocation, and after the allocation was revealed to Player B, each was given a compulsory exit survey respectively (i.e., completion was a requirement for payment). In addition to basic demographic information, we also asked three questions of philosophical relevance. First, Player A participants were asked whether they thought the allocation they made 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 proposed by Player A was 'fair'. (Call this the "fairness question.")[9]

3.1.2: Experiment 1: Results

We found Player A participants on average transferred a total of $0.238 to their Player B counterpart. As mentioned, a recent meta-analysis of dictator games uncovered an

[8] We limited the subjects to US participants, because using the low stakes that we use in the present study was shown to have little effect on giving behaviour in the dictator game within the US population, but not, for example, within the Indian population. See Raihaini et al. (2013) for details.

[9] The exact wording for all questions is available in the supplementary materials available at [Removed].

average transfer rate of 28% of the pot. This suggests our results are fairly consistent with what has previously been observed in the literature, even if slightly on the more-selfish side. A total of 56 Player A participants (32%) selected to keep the entire allotment for themselves, while 38 (21%) opted for an equal division. Only a small handful (n=9, 5%) of individuals transferred over half of the bonus to their Player B counterpart.

Finally, on the harm and fairness questions, this is what we found. On the generic harm question, where we asked both Player A and Player B participants whether a dictator who transfers zero has in some way harmed their counterpart, 43% of Player A participants and 48% of Player B participants responded that such transfers would, in fact, do harm. So, although subjects are split as to whether exceptionally low offers in the dictator game result in a harm to the recipient, we found that a sizeable portion of subjects thought that such low transfers would 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 would have 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 that Player As and Player Bs were actually split 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. As a further comprehension check, we isolated the Player A's who kept all of the bonus, and noticed that all 56 of them gave the same answer on the specific harm and generic harm questions, as we should expect if they fully understood the task and questions.

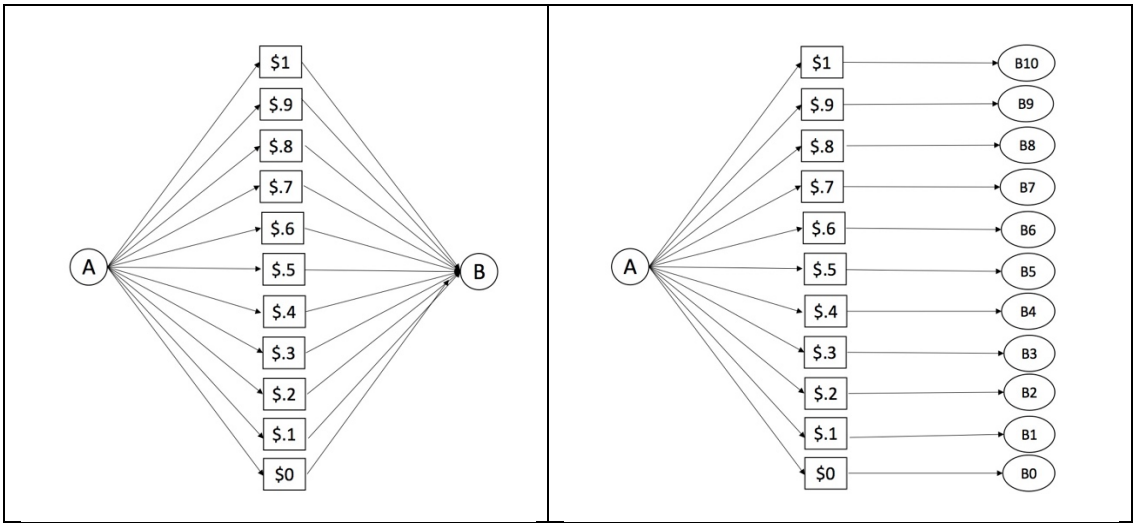
3.2 Experiment 2 (the "Non-Identity Dictator Game")

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 member a label of B0, B1, …, B10. Each of the members 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. And this revelation was the only form of contact between the Player As and any members of their respective Player B Groups in the experiment.

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 behaviour 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 underlying strategic scenario. Importantly, each Player A was explicitly told that those members of the Player B Group not matched to the selected allocation would never find out about Player A's behaviour. 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 in the first place. And, importantly, any member of a Player B Group chosen to receive a bonus through this process, and thus drawn into the strategic scenario, 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.

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".



(a) (b)

Figure 1: Representation of the choice problem faced by Player A in the standard dictator game (a) and non-identity dictator game (b)

As noted earlier, when standard dictator games are run in the lab, subjects don't tend to act as we would expect them to act if we were to assume that subjects are entirely self-interested and care only about maximising their monetary reward. This is indeed what we found in our standard dictator game, and this suggests that our subjects were driven at least partially by other regarding behaviour. The question is whether the identity-affecting nature of our non-identity variant of the dictator game will have a noticeable effect on the behaviour of our subjects. We now turn to the results of our experiment after a brief discussion of the exact experimental set-up. We then compare behaviour in the standard dictator game to that in the non-identity dictator game.

3.2.1 Experiment 2: Set-up

As was the case in experiment 1, we once again used subjects from Amazon Mechanical Turk (and limited our subjects to those using IP addresses in the United States). We recruited a total of 176 Player A participants and 1936 Player B participants. Once again, all interactions took place over the online interface, and all subjects' anonymity was guaranteed. All subjects received a $0.50 participation fee upon

successful completion of the experiment and were told that there was a possibility they could receive up to an additional $1.00. On average, the experiment took the active subjects (i.e., either those selected as Player A or those selected from the Player B group through Player A's choice) an average of just over four minutes to complete.

Player A participants were presented with a detailed description of the non-identity dictator game (described above). 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 before they were allowed to proceed through the rest of the experiment. Those who passed the comprehension test were then allowed to choose their preferred allocation, given a slightly revised set of debriefing question, 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 allocation chosen by the Player A they were paired with (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.[10] 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.[11]

3.2.2 Experiment 2: Results

We found that on average Player As transferred $0.155. A total of 81 Player As (45%) selected to keep the entire endowment for themselves, while only 21 (12%) opted for an equal division. Only a small handful of individuals transferred over 50% of their endowment to their Player B counterpart (n=6, 3%). 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 subjects by and large do not think zero offers in the non-identity dictator game result in any harm to Player B0. On the specific harm question, we found that 92% of Player As thought they did not harm the recipient who would receive the transfer, while 84% of active Player Bs

[10] The information about the experimental set-up given to active Player Bs was different from the information Player As received in two important ways. First, active Player Bs were not initially told exactly which role they were selected to play in the game, leaving open the possibility that they might have been selected as Player A. (Player A's knew their position from the outset.) Second, active Player B's were not told that all members of the Player B group not chosen according to Player A's selection would never be told about their participation in the game. These changes were necessary because of the concern that most of the savvier active Player Bs, if told either explicitly or implicitly about their position, would drop out of the study before completing the comprehension questions, thus biasing that half of our sample.

[11] See supplementary materials for details of the exact wording changes, available at [To Be Posted].

thought the transfer didn't harm them. The latter is rather surprising, given that offers were very low on average. On the fairness question, 68% of player As and 57% of active Player Bs took the transfer to be fair. We did a similar comprehension check as before, and found that only one of the 81 Player As who chose to keep the whole bonus answered the specific and generic harm questions differently, suggesting a very high level of comprehension.

3.3 Comparison of Experiment 1 and Experiment 2

Here is a comparison between the behaviour 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. Figure 2 nicely illustrates this difference. 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 opted to make exceptionally low transfers in the non-identity dictator game than in the standard version. This clearly suggests that, at least in the aggregate, identity-affecting choice problems tend to generate more self-interested behaviour on the part of the dictator.

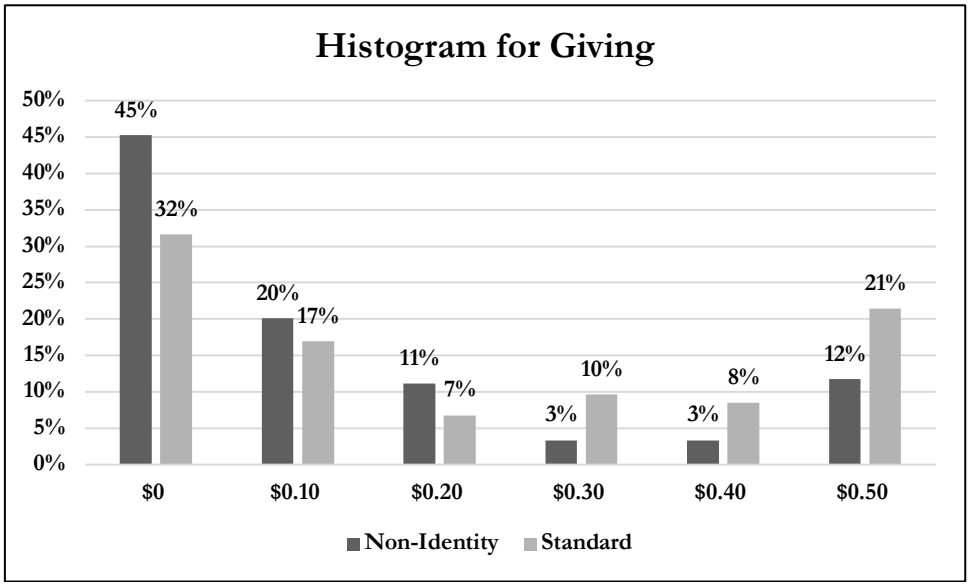


Figure 2: Proportion of player A transfers in standard and non-identity dictator games, with transfers above $0.50 omitted (n=9, 6, respectively)

Attitudes about harm were also substantially different between the two experiments. Both Player As and Player Bs were significantly 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 on the whole much lower than those made in the standard dictator game. As noted, a total of 31% of Player Bs in the standard dictator game felt they were harmed by the specific transfer they received from their Player As. But only 16% of active Player Bs in the non-identity version felt harmed by their corresponding Player As' choice of allotment. Likewise, substantially more Player As in the standard dictator game felt their behaviour 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, in these kind of low transfers, 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.

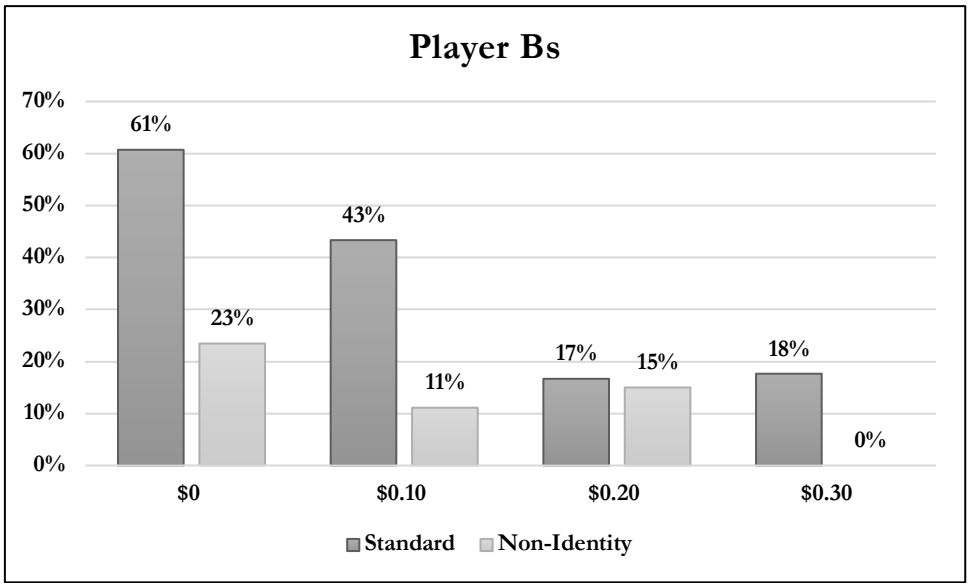


Figure 3: Proportion of Player Bs receiving a low transfer who believed the transfer did them harm

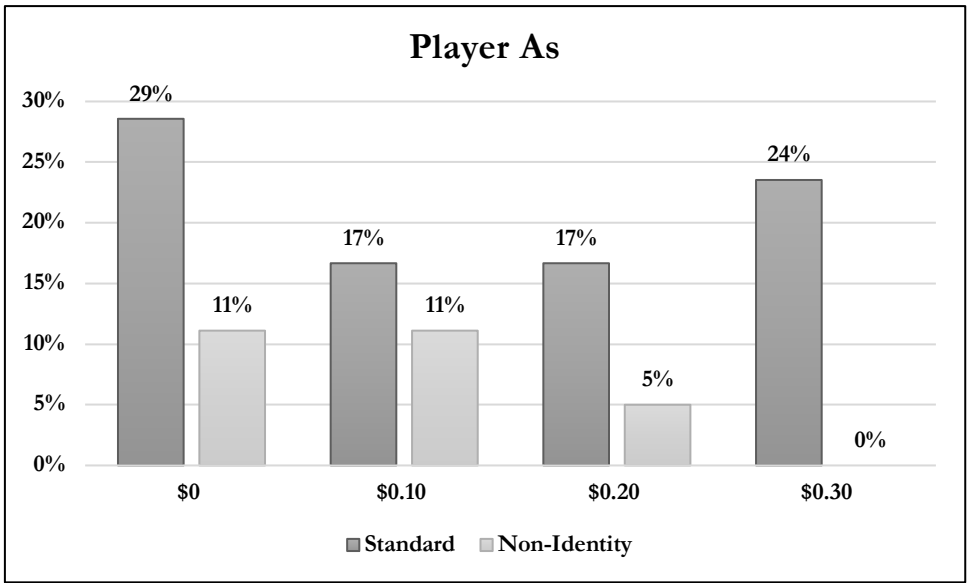


Figure 4: Proportion of Player As giving a low transfer who believed the transfer did the recipient harm

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%).

4. Practical Implications for the Non-identity Problem

These studies were ultimately intended to probe the practical importance of the non-identity problem. Recall that various authors claim that the non-identity problem is of great practical importance, and yet this claim is slightly puzzling, given other claims made in the literature. In particular, a widespread assumption made in the literature is that the paradigm identity-affecting actions seem to carry most of the same intuitively immoral features as their parallel same-person actions. Even authors who ultimately argue that such intuitions are mistaken also admit to sharing the intuition. But 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. And even if the public's intuitions are slightly different from those of philosophers, there is another reason for caution: 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 behaviour.

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 the public generally was 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 the 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). This suggests that while there 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 behaviour. And the substantial portion who make it through the comprehension question are clearly 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. A substantial portion of the population is fully able to track the morally relevant features of identity-affecting choice problems.

Do the data suggest that the identity-affecting nature of a choice problem has an effect on the public's behaviour? We believe they clearly do. First, there was a rather pronounced drop in giving behaviour, from a mean of $0.238 to $0.155. Although a

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 took the whole pie for themselves and a 7 percentage point increase in those who took the vast majority of the pie (i.e., 80 or 90 percent of the bonus). Even offers, i.e., Player As that kept exactly half of the pie for themselves, decreased by 9 percentage points (from 21% in the standard dictator to 12% in the Non-ID version). These results suggest that a substantial portion of the population does see a moral difference when facing an identity-affecting choice problem. In particular, when placed in an identity-affecting choice problem, they are less willing to make altruistic sacrifices for the sake of those on the other side.

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. Nonetheless, we do think they give us some good reasons to be careful with 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 will be less willing to make even small sacrifices when faced with such a problem. In practical terms, this could easily equate to an unwillingness to accept small tax increases, or small price increases on certain products, etc. And although the increased selfishness was far from uniform across our subjects, it was certainly enough to make a policy difference in a population like the US or Australia, where support for these kinds of policies comes in at close to an even split among the voting population. In such a context, broadcasting the identity-affecting nature of our various choice problems, like those involved in climate policy, is likely to have negative consequences. And the fact that the increases in selfish behaviour 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.[12] Such motivational efforts are likely to be especially foolhardy if philosophers continue to broadcast the identity-affecting nature of such policies. What we've uncovered through this study suggests caution is due on both fronts.

5. Implications for Normative Reasoning on Harm

Much of the literature on the non-identity problem is entangled with the literature on harm. In particular, when Parfit (1984) set the stage, the non-identity problem was deemed a problem because of the common-sense notion of harm that we tend to employ in our moral theorizing. This common-sense notion of harm, often referred to as the counterfactual comparative account, holds that to harm someone is to make them worse off than they would have been otherwise. And in a non-identity case, it's not true that you've made the relevant person or people worse off than they would

[12] This is not to say that we should also be cautious when using shorter term harm considerations to motivate the population toward conservationist policies. There are many green policy choices where a failure to act now causes harm to individuals who will exist no matter which policy is enacted, since they work on a much shorter timeframe. Our continued failure to curb our use of coal-burning power generation is a clear example, since our inaction will harm many individuals over the coming decades due simply to respiratory disease.

have been otherwise. After all, if you had made the other available choice, that person, or those people, wouldn't have existed. 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 harmed after all. So various authors have seen the non-identity problem 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).

We admit that our non-identity dictator game doesn't place our subjects into a non-identity case in the exact same sense as the individuals involved in the cases used in the above debates. After all, the members of the Player B group don't fail to come into existence when they aren't chosen to receive a benefit. We think this disanalogy, strictly speaking, doesn't impede our ability to make predictions about how the public's behaviour could change if the non-identity problem were made salient, but we can imagine others might not be so optimistic. Regardless of how crucial one believes the strict disanalogy with the non-identity problem to be, the choice problem we've generated in our game can nonetheless offer some insight into which notions of harm members of the public are employing in their normative reasoning. 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 behaviour.

First, it will be helpful to have a precise statement of the 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.[13]

[13] To avoid confusion, we should point out that one of the main authors who is typically associated with non-comparative accounts of harm in the literature, namely Harman (2004, 2009), would likely not endorse the formulation we give here. For her to offer a solution to the non-identity problem, she only requires that placing someone into an intrinsically bad state is sufficient for the act to count as doing harm, as should be clear from what we say below. She doesn't require that it is also necessary. And others who endorse non-comparative accounts might be similarly hesitant. But nonetheless, the public might be employing a notion of harm where an act wouldn't count as doing harm unless the act causes those affected to be in an intrinsically bad state. Just as an aside, we note that if one endorses a non-comparative account of harm on which placing another into an intrinsically bad state is sufficient by not necessary for harm, such a view would fail to bypass some of the drawbacks of counterfactual comparative accounts that have been pointed out in the literature. For example, it would allow that failing to benefit might still count as a harm. (We discuss this supposed drawback below). We thank David Boonin for insisting we make it clear that we are not intending this account to represent Harman's view of harm.

To give an example of how these two accounts of harm can come apart, think back to the example of Wilma, Pebbles, and Rocks from earlier. In that scenario, if Wilma decided to conceive a child immediately, then she'd have Pebbles, and Pebbles would be born incurably blind. But if Wilma decided to delay conceiving and receive treatment for a few months, she'd conceive Rocks who would be unencumbered by that disability. But Wilma decided to conceive early anyway, bringing Pebbles into existence. On the counterfactual comparative account, Wilma hasn't harmed Pebbles with her decision, since it's not the case that Pebbles would have been better off if Wilma had decided otherwise. If she had waited, Rocks would have been born instead. But the non-comparative account has a very different verdict. By choosing to conceive early, as opposed to waiting for the treatment to kick in, she has caused a child to be in an intrinsically bad state.[14] So Pebbles is indeed harmed by Wilma's decision to conceive early.

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 defence). But since we are interested in which notions of harm play a role in the normative reasoning of the public, the existence of philosophical counterexamples doesn't give us a reason to toss out the concept. It's well known 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 in 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 akin to a counterfactual comparative notion of harm in their normative thinking.[15]

---

[14] As suggested earlier, it's assumed in the literature that incurable blindness is an intrinsically bad thing for someone to have. If the reader disagrees, the original example could have been modified, swapping blindness with an impairment the reader accepts is intrinsically bad. It's likely that such a reader wouldn't have seen the force of the non-identity case in the first place, and so we would have had to make such a modification at that earlier stage anyway.

[15] We admit that there is a possible wrinkle with this inference. It could be that when the subjects say that a 0 transfer does harm, it might be because they are making a prediction about how the Player B in such a scenario will react to receiving nothing. If they think such a result would make that Player B angry, and if they think being angry is an intrinsically bad state, they might register that it's a harm even

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 2017 for critique of Bradley's argument.) It turns out that a 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 we don't intend to take a stand on which notion of harm is the "correct" one here.

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 account of harm. Recall that a much smaller proportion of Player As believe they have harmed the member of the Player B Group with their offer, even though the offers were on the whole much lower. And we also saw a sizeable 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.[16]

But we think the data also suggest that an even larger portion of the population employs some other notion of harm. 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 although our study suggests some limits on what portion of the population could be using these notions of harm, our data don't help us distinguish which of the alternative accounts this portion of the population employs. 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.[17] In future

---

if they employ a non-comparative notion of harm. We think it's rather unlikely that this is the main explanation of the responses, given that all participants are aware that even the subjects given a 0 transfer will still take away a show up fee, which we set at a higher rate than the average show up fee for MTurk studies. But we are devising ways to tease out and test this possibility in future studies. We thank David Boonin for raising this suggestion.

[16] The possible wrinkle mentioned in the previous note might apply here as well. It could be that subjects who are employing an anger-based non-comparative account, if you will, are predicting that someone given a 0 transfer wouldn't be angry in the Non-identity version, perhaps because such a Player B should realise they couldn't have done any better than a 0 transfer. Like before, we think this kind of story is unlikely to be the right explanation of the shift, but we're looking into ways to examine the possibility further. Thanks, again, to Boonin for pointing out this possibility.

[17] David Boonin has pointed out to us that there may be another competitor, which he tells us he sees hints of in class discussions on related topics. This competitor is a kind of morally laden counterfactual comparative account, which diagnoses an act as doing harm if and only if the act *wrongfully* makes an individual worse off than she would have been otherwise. If we take this as another competitor, we

work, we are devising questions to allow us to tease apart which of these competitors this substantial portion of the population are using. Additionally, a non-negligible portion of our subjects in the non-identity dictator 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 the possible contender mentioned in note 17). 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. Further study is needed to tell if this is the case, or if, instead, this group we found amounts to merely noise in the data.

6. Objections and Replies

In this section, we address some objections that raise questions about the relevance of our results to the philosophical debates we are engaging with. First, one might object that our study isn't as relevant to the non-identity problem in the context that originally motivated our enquiry, namely the context of climate change policy. In particular, a number of philosophers working on the non-identity problem have noted the difference between what we might call "personal" and "impersonal" identity-affecting cases (e.g., Boonin 2014, Weinberg 2014). In personal cases, like the case of Wilma discussed in Section 2, the actor can easily conceive of a concrete individual who predictably will be affected by the identity affecting choice. On the other hand, in impersonal cases, like the choice between Policy C and Policy D, the agent at issue is considering the effects the choice predictably will have upon an large and amorphous body of possible people, each of which would be difficult to conceive of in any concrete way. And as previous authors have pointed out (e.g. Weinberg 2014), our intuitions may justifiably be quite different in these two kinds of identity-affecting cases. If so, this could give us a reason to think that the population will also exhibit markedly different behaviour in these two kinds of cases. This would, in turn, threaten to undermine the relevance of our study, understood as a personal identity-affecting case, to the impersonal identity-affecting case of climate change policy.

In reply, although we admit that there is an intuitive difference between personal and impersonal identity-affecting cases, we note that it is important to properly track the directional shift of the intuitions at issue. Most authors who raise the personal/impersonal distinction do so in order to argue that our intuitions are on firmer ground in the personal case. As the story goes, we tend to have a stronger intuition that Wilma has done something immoral when she chooses to conceive of blind Pebbles than that the US has done something wrong when choosing Policy D (i.e. "Depletion"). But notice that the shift here goes from a stronger intuition of having done wrong in a personal identity-affecting case than an impersonal identity-affecting case. So while the objector above may have rightly found a trace of disanalogy, it turns out to be largely beside the point. If our subjects' intuitions or wrongness are supposed to be *stronger* in the personal identity-affecting case than an impersonal one, then claiming that our study is more parallel to a personal identity-affecting case actually strengthens our argument. Since we saw a marked change in

should note that our data likely also don't distinguish between this version of a counterfactual account and the temporal or non-comparative account. After all, our subject 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*.

behaviour in our study, if we follow the objector here in thinking it's more like a personal case, we should expect an even larger effect when the other "players in the game" are amorphous, unidentifiable, and inconceivable future beings.

A second, more challenging objection one might raise against the relevance of our study to the non-identity problem in the context of climate change policy would be to point out a possibly relevant difference between the harms involved, which could cause a corresponding difference in behaviour. As we admitted earlier in Section 5, many philosophers will think that there simply isn't any harm done in either of our dictator games, since they believe there is a drastic difference between failing to benefit someone and actually harming her. On such a picture, the subjects in our study that think harm can be done in the dictator game are simply confused. But the objector we have in mind here takes our subjects at their word and then tries to show that the behaviour we saw in our study shouldn't be expected to scale up to the case of climate change policy. In particular, if not being given any slice of a very small monetary pie can be rightly seen as a harm, it certainly isn't much of a harm. And when we move from the standard dictator game over to the identity-affecting version, the specific details of the change might be enough to wash out the tiny amount of harm that was originally registered by our participants. On the other hand, being left with an environment that is severely degraded and badly depleted of its resources would seem like a much more serious harm. So, although we saw a marked change in attitude and behaviour between our two studies, with subjects exhibiting much more selfishness in the identity-affecting version, it could be that the seriousness of the perceived harms to future generations erases this change in bad behaviour. This objection attempts to diffuse our worry that the non-identity problem will make action on climate change more difficult by motivating the hopeful thought that the public will still want to avoid leaving a depleted and degraded planet for future generations, even after they fully grasp the puzzle.

Our response to this objection is to admit that this is certainly a possibility, and it's one that would need to be empirically tested before we could fully adjudicate the matter. It is true that even if we accept the subjects at their word in thinking that harm can be done in the dictator game, the harms are surely minor. If we wanted to test whether the kind of behaviour we found in our study is likely to scale up, we would need to substantially raise the stakes. One possibility would be to run a version of the studies where bonus money is endowed on both sides of the player gap, and where Player A must take the bonus money away from Player B (or the Player B group) if she wants to get the "full" bonus. (For this version of the dictator game, see List 2007.) A second would be to simply up the stakes in the game, from say $0.50 to $100. If we were to see more similar behaviour between the standard and identity-affecting games under either of these variants than we saw in our versions, then this might speak in favour of the objector here.[18] Finally, we could retreat from the behavioural games and simply poll subjects about whether they see a relevant moral difference between identity-affecting cases with low stakes (like those involving money) and those with higher stakes (like those involving general well-being). As we noted earlier, we probably shouldn't conclude anything too strong from the answers we receive from such questionnaires, since our interest is ultimately in behaviour, and what subjects predict

[18] We are currently devising variants of the study with somewhat higher stakes, although resource limitations admittedly make variants with very high stakes somewhat infeasible.

about their own behaviour can drastically come apart from how they actually behave. But, all the same, this indirect method might be the best available way to further probe this question, given the legitimate moral constraints against doing harm in the lab. But regardless of how we adjudicate the quality of this objection, it remains an empirical question left to be tested.

7. Conclusion

In this paper, we have tried to show that the non-identity problem 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 behaviour 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 behaviour 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 behaviour 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.[19]

REFERENCES

Andreoni, James, William Harbaugh and Lise Vesterlund. 2008. "Altruism in experiments." in Durlauf and Blume (eds.), *The New Palgrave Dictionary of Economics*. Palgrave.

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.

[19] We would like to thank the audiences at the 6th Australasian Workshop in Moral Philosophy, Australian National University, University of Colorado – Boulder, National University of Singapore, and Northeastern University for all their helpful comments, questions, challenges, and suggestions which have greatly improved the paper. We would especially like to thank: Daniel Cohen and Duncan Purves for helpful discussion and guidance; David Boonin for his extensive and helpful comments on an earlier draft; and our colleagues on a successor project, Ben Grodeck and Toby Handfield, for ideas and insights that have improved the present project. Research on this project was supported by the School of Politics and International Relations, the School of Philosophy, and the Research School of Social sciences at Australian National University, as well as Kopec's ARC DECRA Grant: DE180101119.

Bruner, Justin, O'Connor, Cailin, Rubin, Hannah and Simon Huttegger. 2018. "David Lewis in the lab: Experimental results on the emergence of meaning." *Synthese* 195: 603-621.
Buchanan, Allen, Dan Brock, Norman Daniels, and Daniel Wikler. 2000. *From Chance to Choice: Genetics and Justice.* Cambridge University Press.
Engel, Christoph. 2011. "Dictator games: A meta study." *Experimental Economics* 14: 583-610.
Fehr, Ernst and Klaus Schmidt. 1999. "A theory of fairness, competition and cooperation." *The Quarterly Journal of Economics* 114: 817-868.
Feit, Neil. 2017. "Harming by failing to benefit." *Ethical Theory and Moral Practice.* (online first) https://doi-org.virtual.anu.edu.au/10.1007
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.
Gibb, Michael. 2016. "Relational Contractualism and Future Persons." *Journal of Moral Philosophy* 13: 135–60.
Gardiner, Stephen. 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.
Hanser, Matthew. 1990. "Harming future people." *Philosophy & Public Affairs* 19: 47–70
Hanser, Matthew. 2008. "The metaphysics of harm." *Philosophy and Phenomenological Research* 77: 421-450.
Harman, Elizabeth. 2004. "Can we harm and benefit in creating?" *Philosophical Perspectives* 18: 89-113.
Harman, Elizabeth. 2009. "Harming as causing harm." In M. Roberts and D. Wasserman (eds.), *Harming Future Persons: Ethics, Genetics and the Nonidentity Problem.* Springer: 137-154.
Heyd, David. 2009. "The intractability of the nonidentity problem." In M. Roberts and D. Wasserman (eds.), *Harming Future Persons: Ethics, Genetics and the Nonidentity Problem*, Springer: 3–25.
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 O. Edenhofer et al. (eds.), *Climate Change 2014: Mitigation of Climate Change.* Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change . Cambridge: 207-282.
Kumar, Rahul. 2003. "Who can be wronged?" *Philosophy and Public Affairs* 31: 99-118.
Kumar, Rahul. 2009. "Wronging future people: A contractualist proposal." In A. Gosseries and L. Meyer (eds.), *Intergenerational Justice.* Oxford University Press: 251-272.
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: 482-493.
Machery, Eduard. 2017. *Philosophy within its proper bounds.* Oxford University Press.

Meyer, Lukas. 2003. "Past and future: the case for a threshold notion of harm." In L. Meyer et al. (eds.), *Rights, Culture, and the Law: Themes from the Legal and Political Philosophy of Joseph Raz*. Oxford University Press: 143-159.

Nado, Jennifer. 2014. Philosophical Expertise. *Philosophy Compass* 9: 631-641.

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

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

Purves, Duncan. 2014. "The Non-Identity Problem." In N. Nobis et al. (eds.), *1000-Word Philosophy: An Introductory Anthology*. Available at:. https://1000wordphilosophy.com/2014/02/27/non-identity-problem/

Raihani, Nichola, 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: 134-144.

Shiffrin, Seana. 2009. "Wrongful life, procreative responsibility, and the significance of harm." *Legal Theory*. 5: 117-148.

Steinbock, Bonnie. 2009. "Wrongful life in procreative decisions." in M. Roberts and D. Wasserman (eds.), *Harming Future Persons: Ethics Genetics and the Non-identity Problem*. Springer: 155-178.

Thomson, Judith Jarvis. 2011. "More on the metaphysics of harm." *Philosophy and Phenomenological Research* 82: 436-458.

Tobia, Kevin, Wesley Buckwalter and Stephen Stitch. 2013. "Moral intuitions: are philosophers experts?" *Philosophical Psychology* 26: 629-638.

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.