Concrete Over Abstract: Experimental Evidence of Reflective Equilibrium in Population Ethics

Philipp Schoenegger and Ben Grodeck[1]*

Abstract

One central method of ethics is narrow reflective equilibrium, which relates to the conflict between intuitions about general moral principles and intuitions about concrete cases. In these conflicts, general principles are refined, or judgements in concrete cases are changed until no more conflicts exist. In this paper, we present empirical data on this method in the context of population ethics. We conduct an online experiment (n=543) on Prolific where participants can endorse moral principles relevant to population ethics. They also judge concrete population ethical cases that may conflict with their endorsed principles. When conflicts arise, they can choose to revoke the principle, revise their intuition about a case, or continue without having resolved the conflict. We find that when there is conflict, participants are significantly more likely to revoke their endorsements of general principles, than their judgements about concrete cases. This evidence suggests that for a lay population, case judgements play a central revisionary role in reflective equilibrium reasoning in the context of population ethics.

Introduction[2]

General moral principles typically assert what is the right way to act. These general principles then have direct implications on how to act across all concrete moral cases. Conflicts are present when the choice one makes in these concrete cases differs from the choices prescribed by the general moral principle. For one to remain consistent and to have a coherent set of beliefs, one must resolve this conflict by either altering their choice in the concrete case to bring it in line with the general principle, to revoke the principle entirely, or both—until no more conflicts are present. This method of reasoning is known as reflective equilibrium (Rawls 1971; Scanlon 2002; Daniels 2016).

---

[1] Both authors contributed equally to this paper.

[2] We thank for helpful comments and suggestions especially Kirby Nielsen as well as Theron Pummer, Stefan Schubert, Lucius Caviola, Neele Engelmann, Raimund Pils, and Stijn Bruers.

This work was supported by a grant from the Forethought Foundation and the Centre for Effective Altruism.

* Contact: Philipp Schoenegger, University of St Andrews, School of Philosophical, Anthropological and Film Studies & School of Economics and Finance (ps234@st-andrews.ac.uk)
Ben Grodeck, Monash University, Monash Business School, Department of Economics (ben.grodeck@gmail.com)

In this deliberative process, a central distinction is that between abstract intuitions and concrete intuitions.[3] A common way of distinguishing these types of intuitions is offered by Huemer (2008), who divides intuitions into three distinct categories: concrete intuitions (e.g. about cases and thought experiments), abstract theoretical intuitions (e.g. about very general principles), and mid-level intuitions (e.g. about principles with an intermediate degree of generality) (Huemer 2008, 383). In this paper we focus on the two end poles of this spectrum: concrete intuitions about specific cases and abstract intuitions about general principles. The interrelation between these two types of intuitions is important as both play unique and central roles in philosophy, and in ethics specifically. Both are also fundamental to reflective equilibrium, a preeminent method in ethics (Tersman 2018, 1; Paulo 2020, 334), or, as Scanlon, put it, "the only defensible method" (Scanlon 2002, 149) in ethics.

For Huemer, the entire history of analytic philosophy is one in which generalisations that appear rather plausible often lose their appeal once a fitting counterexample case is presented (Huemer 2008, 385). On such a picture of a simplified reflective equilibrium, conflicting intuitions are weighed against each other, often resulting in concrete intuitions having a revisionary function as they produce relatively strong moral intuitions leading to a revision of abstract theoretical principles[4].

The current literature in experimental philosophy more generally is largely preoccupied with showing the frailty and predictable variability of philosophical intuitions. These findings are sometimes used to fundamentally challenge reflective equilibrium as a method. For example, Paulo (2020) argues that if it was shown that moral intuitions were unreliable in a specific (and systematic) way, this would discredit the method of reflective equilibrium altogether (cf. also Brun 2014). Much of the previous research in (the negative programme of) experimental philosophy could then be plausibly seen as contributing to this question (e.g., Buckwalter & Stich 2013; Machery et al. 2017).

In this paper, we present empirical evidence of how conflicts between abstract and concrete intuitions play out in an online experiment. While there has been ample discussion of how these conflicts ought to be resolved normatively, there has been no evidence of how these conflicts are actually resolved. This paper is directly relevant to academic theorizing about the concept of reflective equilibrium, in which conflicts between abstract principles and concrete cases are central. It is also potentially useful for those interested in shaping moral discourse more generally, be it as part of a conceptual engineering framework aiming to change concepts (Cappelen & Plunket 2020)

[3] For the purposes of this paper, we do not specify what exactly we understand an intuition to be. Our results and arguments generalize across different interpretations of intuitions and their role in philosophy and beyond.
[4] Huemer does not himself endorse this picture and favors one in which the strength of intuitions is based on their propensity to bias (Huemer 2008, 391). He argues for the claim that concrete and mid-level intuitions are more prone to errors like biases than abstract theoretical intuitions, which are themselves susceptible to what he calls 'overgeneralisation'.

or for those thinking about preparations for a period of long reflection at the end of which moral uncertainty might be significantly reduced (Ord 2020) among others. While research like this may also help us better understand public controversies, there are numerous potential challenges that make direct translation of any finding to public debate extremely difficult.

Some of the previous literature investigates a similar point to that discussed in this paper, namely the role concreteness and abstractness play in intuitions and judgements. For example, experimental research on free will and moral responsibility finds that, generally, participants tend to hold determinism and moral responsibility as inconsistent if the questions are stated in the abstract, but when confronted with a concrete case, participants are more likely to report determinism as being consistent with moral responsibility (Nichols & Knobe 2007; De Brigard, Mandelbaum, & Ripley 2009; Nichols 2011; cf. also Nahmias & Murray 2011; Murray & Nahmias 2014). Moreover, findings of the abstract/concrete divide on the question of personal identity follow a similar picture, see Nichols & Bruno (2010). Further, Caviola, Schubert, & Mogensen (2021) find differences in intuitions along the rule-case dimension in the context of hypothetical rescue prioritisations. However, above and beyond these strands of research, not much empirical analysis of the concrete/abstract question has been published in experimental philosophy (or cognate fields) (though cf. Carlsmith, Darley, & Robinson 2002), and importantly, no analysis of conflict between abstract and concrete intuitions has been offered so far.

We choose to study reflective equilibrium in the context of population ethics due to the rich environment in which all general moral principles that may be prima-facie morally compelling run into concrete cases where general intuitions about 'what is best' will create conflict with a general principle. The impossibility theorems that arise in population axiology (Arrhenius 2000; Greaves 2017) allow us to investigate which concrete intuitions people are most willing to give up—if any—to have consistent general moral principles.

Our aim is to provide data on how exactly these confrontations between abstract and concrete intuitions play out in a lay population in this relevant context. Prima facie, there are three patterns that the data might take. First, one might think that abstract intuitions are more likely to be revised to resolve conflict. This would be the '*Primacy of the Concrete'*. Second, one might think that the opposite picture holds. This would be the *'Primacy of the Abstract'*, where it is the concrete intuitions that are typically revised. Third, a possible pattern of data is that there is no general tendency to favor one type of intuition over another in cases of conflict.

To test which of these three patterns best describes reflective equilibrium reasoning in a lay population, we conduct an online experiment (n=543) on Prolific. Drawing heavily on the methodology introduced by Nielsen and Rehbeck (2020), we present participants with abstract general principles drawn from the population ethics literature. We also present them with several

population ethics cases in which worlds differ along the dimensions of the number of persons who exist, and their utility. Our participants evaluate these worlds judging which one is better in each scenario. In the first part of the experiment, we ask participants to carefully consider and endorse those principles that seem right to them. In the second part, we ask them to judge which world (out of two) they think is better. We present some participants with these two parts reversed. In the reverse-order treatment, participants first judge which worlds are better and only then are asked to endorse principles that seem right to them. The abstract principles and cases are chosen such that the judgements about some worlds conflict with some of the abstract principles that they could have chosen. In a third part, participants are informed about any potential conflict between their endorsed principles and their choices in the concrete cases. In response, participants can either revoke the principle they endorsed at the beginning, change their choice in the concrete case, or leave the conflict unresolved. Their choice in the last part of the experiment is the variable of primary interest as it enables us to identify which of the three data patterns is present. This research helps answer this question and provide first descriptive data of lay reasoning in cases of conflict in the example of population ethics.

We find that participants prefer to revoke their previously endorsed principles at a significantly higher rate than their judgements about concrete cases across all conflicts. Our results thus provide support for the *Primacy of the Concrete* hypothesis in that abstract intuitions are typically revised in favor of the concrete. This provides the first empirical results of folk moral reasoning and gives us some insight into how lay populations engage in reflective equilibrium reasoning in the context of population ethics.

Methods

We collected our sample from Prolific, an online participant recruitment platform. We recruited a total of 600 participants from the United Kingdom who were paid £2 for completion of the study. Further, they could earn £0.10 for answering each comprehension question correctly. There were a total of five comprehension questions, resulting in a maximum additional reward of £0.50. Overall, we excluded 55 participants from analysis because of their performance on the comprehension questions, i.e., they correctly answered only two or fewer comprehension questions[5]. This research was pre-registered[6] and has received ethics approval.[7]

[5] This was outlined in our pre-registration: However, our main results are insensitive to the inclusion of these observations.

[6] For our pre-registration on the Open Science Framework following the AsPredicted form, see https://osf.io/hu5xk.

[7] Ethics approval has been received from the University of St Andrews (approval code: SA15172).

This main study consists of four parts and has two orders of appearing to participants[8]. Figure 1 outlines the experimental procedure, where we indicate the number of participants, the randomization procedure, and the general structure that the experiment takes. Our methodology is directly adapted from Nielsen & Rehbeck (2020) who studied choice axioms in expected utility theory and decisions in lotteries. The mechanism Nielsen & Rehbeck (2020) introduced allows for the analysis of participant's intuitions about general principles, concrete cases, and most importantly their response to the eventuality that the former two conflict.

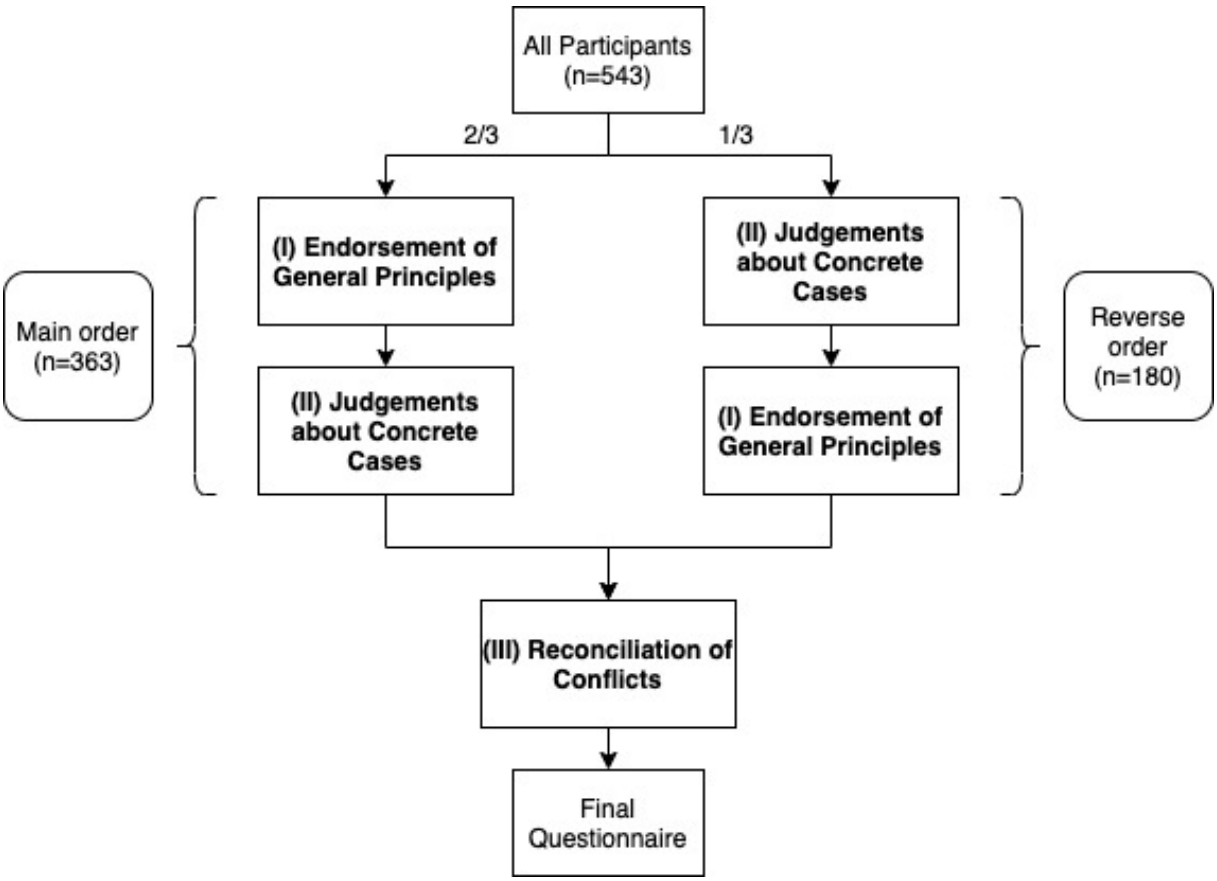

Figure 1. Experimental Procedure

Overview of experimental procedure (total sample size, levels of randomization, and structure of experiment)

In the first part of the experiment[9], participants are presented with five general principles. These are four population ethics principles—Totalism, Averagism, Person Affectism,—along with the more general Pareto principle. For each principle, participants had to choose whether to endorse or not endorse the principle. Figure 2 shows how we presented these principles to participants. We phrased these principles in a manner that was accessible to a lay audience. We also provided an

[8] The full instructions are available in Appendix A.
[9] When referring to the order of tasks, we will always be referring to the main order unless specifically stated otherwise.

introductory paragraph that introduced key conceptions such as 'population' and 'happiness levels' to ensure adequate understanding of the concepts. These introductory paragraphs were available for consultation throughout the study. We told participants to click a link that directs them to a website that portrayed these paragraphs. They were encouraged to keep this website open for the remainder of the study should they wish to consult it. Further, we included a control principle, i.e., a principle that states the opposite of what a specific normal principle states (i.e., we included a control principle for totalism). The purpose of this control principle is to ensure that people are not simply endorsing principles randomly, thus controlling for acquiescence bias. For full texts of all five principles, see Figure 2. In our reverse order treatment, we give 1/3rd of participants the concrete population ethics cases first, and only then are they given a chance to endorse the general (population) ethics principles. Since we were primarily interested in data from the first treatment ex-ante, we decided on a 2/3 vs 1/3 split, instead of ½ vs ½.

Averagism
Society X is better than Society Y **if and only if the average happiness level in X is higher than the average happiness level in Y**. If these average happiness levels are equal, then X and Y are equally good.

Totalism
Society X is better than Society Y **if and only if the total sum of happiness levels in X is higher than the total sum of happiness levels in Y**. If these total sums are equal, then X and Y are equally good.

Control Totalism
Society X is better than Society Y **if and only if the total sum of happiness levels in X is lower than the total sum of happiness levels in Y**. If these total sums are equal, then X and Y are equally good.

Person Affectism
Society X is better than Society Y **if and only if changing from Y to X makes people already existing happier, not if it brings new people into existence that would be happy**. X is not better than Y if changing from Y to X does not make already existing people happier.

Pareto
Society X is better than Society Y **if and only if changing from Y to X makes no people that already exist worse off, and at least one person that already exists better off**. If some people are made better off and some worse off, then changing from Y to X is not better.

Figure 2. Abstract Principles

List of all five principles used throughout the experiment

In the second part of the experiment, participants are presented with three concrete population ethical cases. In these cases, participants judge which of two worlds is better. All populations are described in numerical terms for total number of people, total utility, and average utility. Further, participants are also provided a graphical depiction of these populations to aid understanding. In our last two cases, we told participants that we start in Society I and the world changes to Society II. Figure 3 presents the three cases in detail. Our first case (Repugnance) represents the repugnant conclusion (Zuber et al. 2021), where Society 2 has greater total utility, despite people only having barely happy lives. This case was included to conflict with Totalism. Our second case (New Happy People) represents the example where new happy people are brought into the world at a slightly lower level of happiness than the original population. This case was included to conflict with Averagism and Person Affectism. Our third case (Trade-offs) represents a situation where one population is made a lot better off, with the other population only being made slightly worse off. This case was included to present a conflict with the Pareto principle. For the set of all three cases, see Figure 3.

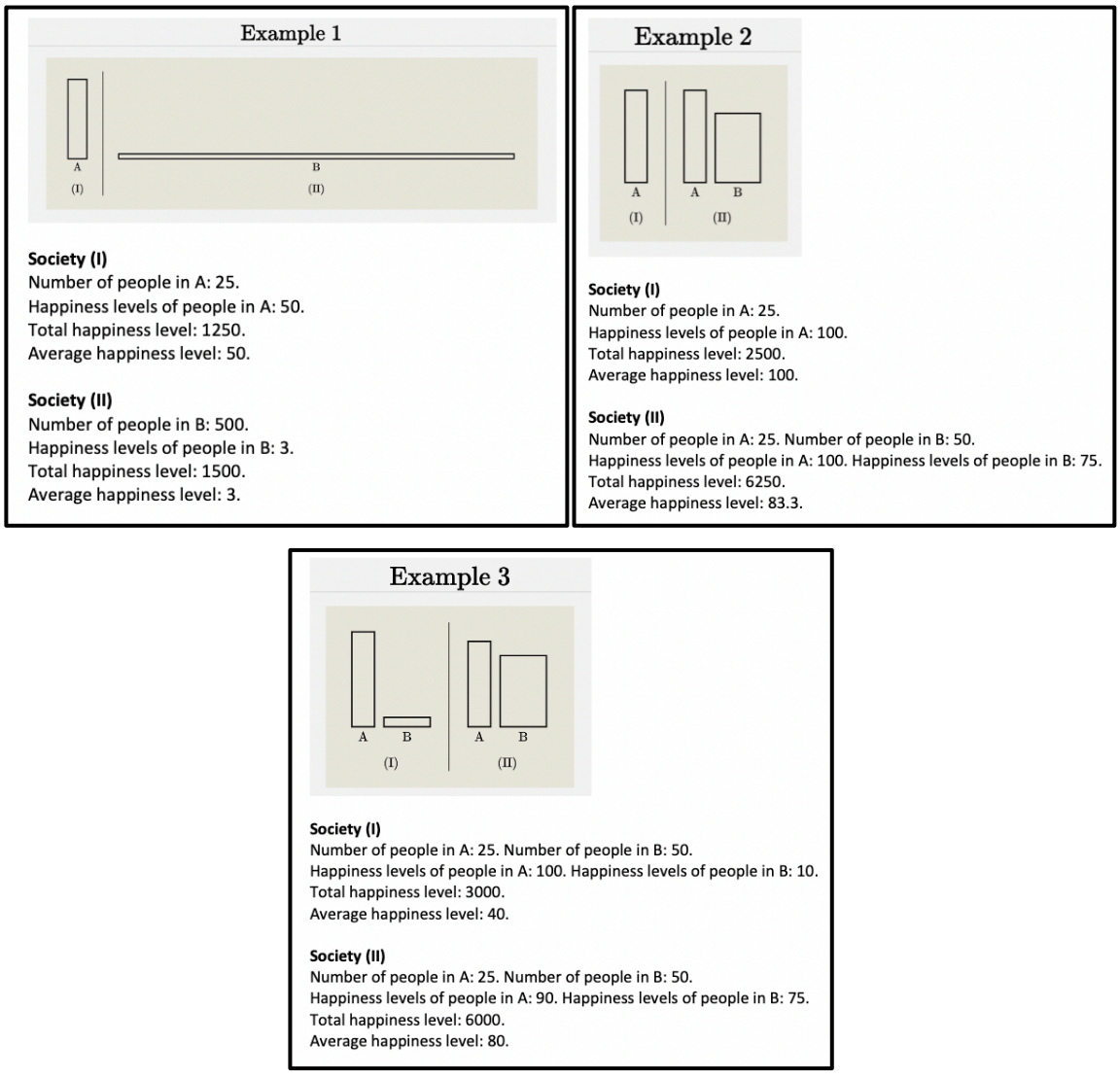

Figure 3. Concrete Population Ethical Cases

Set of all three cases including graphical depiction and numerical values

In the third part of the experiment, conditional on participants choices in Task 1 and Task 2, participants are finally presented with all the violations between their endorsed principles and choices in the concrete judgements (if any). This involved reminding them of the principle they endorsed, their choice in the specific concrete case, and explaining why their choice violates the general principle. For all violations, participants are presented with three options to respond: If they wish to resolve the inconsistency, they can either revoke the principle that they endorsed previously or change their choice in the concrete case. Further, they can also decide to do neither and retain the inconsistency. See Figure 4 for the decision presented to participants, including all three options.

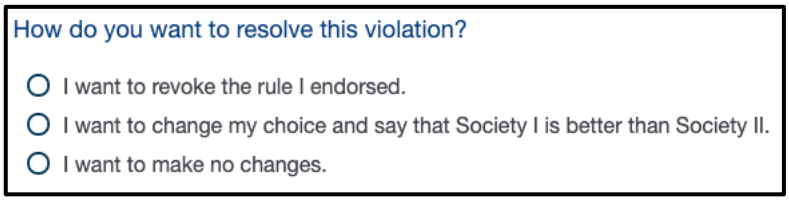

Figure 4. Resolution of Conflict

Resolution screen with three options presented to participants in case of conflict

In the last part of the experiment, participants are asked to fill out a demographic survey.[10]

Results

Part 1: Endorsing General Principles:

We find that the two utilitarian principles (Averagism and Totalism) are the most popular with 85% of participants endorsing Average Utilitarianism and 74% endorsing Total Utilitarianism, with participants significantly more likely to endorse the former (McNemar test, $p<.001$). Both utilitarian general principles are considerably more popular than Person Affectism (59%) and the Pareto Principle (61%), with both utilitarian axioms being endorsed at a significantly higher frequency (Average Utilitarianism vs Person Affectism, McNemar test, $p<.001$; Average Utilitarianism vs Pareto, McNemar test, $p<.001$; Total Utilitarianism vs Person Affectism, McNemar test, $p<.001$; Total Utilitarianism vs Pareto, McNemar test, $p<.001$). Lastly, we find no significant difference between the proportion of participants who endorse Person Affectism compared to Pareto (McNemar test, $p=.845$).

[10] We provide a summary of our sample's demographic characteristics in Appendix B.

We find that 23% of participants endorse our Control principle (Control Total Utilitarianism). This control axiom is endorsed at a much lower rate than the other four general principles. More importantly, Total Utilitarianism is endorsed at a much higher rate than the Control principle. However, out of the 23% of participants who endorsed the Control, we find that 87% also endorse the contradictory axiom Total Utilitarianism. This occurs at a similar rate to the endorsement of Control principles in Nielsen and Rehbeck (2020) and suggests a similar level of acquiescence to the literature.

Table 1 breaks down the number of general principles that participants endorsed. We find that 25% of participants endorse all four of the main general principles[11], while 12% endorse only the two utilitarian principles and only 4% endorse both Pareto and Person Affectism. Given the high rate of endorsement of our general principles, and the relatively lower rate of endorsement of our Control principle, this suggests that subjects on average understand the decision rules to a reasonable extent.

Number of endorsed principles	**Including Control Axiom**	**Excluding those who selected Control Axiom**
0	2%	2%
1	5%	6%
2	21%	26%
3	42%	44%
4	25%	21%
5	6%	
n	543	419

Table 1. Frequency of endorsed principles.

Overall, we conclude that these general principles are considered as desirable by participants. While participants violate these principles in the concrete cases, they at least show a preference for these rules in the abstract.

> **Result 1:** *Nearly all individuals have a preference for at least some of the general principles, with significantly more participants endorsing the utilitarian rules. Participants also endorse the general principle Total Utilitarianism at a much higher frequency than its opposite (Control Total Utilitarianism).*

[11] Excluding Control Total Utilitarianism

Lastly, we investigate if participants endorse general principles at a different rate when they see the concrete cases before making their endorsement decisions. Table 2 shows the percentage of participants endorsing each axiom when doing Task 1 (endorsing general principles) first, or Task 1 second. We find that on average participants are more likely to endorse more general principles overall when they see Task 1 first compared to seeing Task 2 first (3.1 vs 2.7, Mann-Whitney test, $p<.001$).

We find that there is no significant difference between the rates of endorsement between the two treatment orderings for Pareto, Person Affectism, and the Control principle respectively. However, we do find that participants endorse the two utilitarian principles at significantly lower rates after completing the concrete cases task, compared to before. This is most prominent for Total Utilitarianism where we observe participants endorsing this principle significantly less of the time, by 25.7 percentage points. This result suggests that once people realize some of the implications of utilitarian principles, their rate of endorsement of these principles decreases. This further lends support to the idea, that in the abstract, utilitarian principles seem reasonable, but once participants are faced with the implications of these principles, the principles become less intuitively appealing.

	Task 1 first	**Task 1 second**	**Order test p-value**
Average Utilitarianism	87%	79%	.021
Total Utilitarianism	83%	57%	.001
Control	25%	19%	.123
Person Affectism	59%	57%	.655
Pareto	61%	61%	.942
n	363	180	

Table 2. Differences in principle endorsement depending on whether the principle endorsement (Task 1) was presented first or second. The p-values are based on a Chi-squared test.

> ***Result 2:*** *Participants are significantly less likely to endorse utilitarian principles after observing the three concrete cases, compared to the reverse order. This is most prominent for Total Utilitarianism.*

Part 2: Concrete Cases

Table 3 below shows the percentage of participants who chose Option 1 in each of the three concrete cases[12]. See Figure 3 for all three cases and the corresponding choices available. We find

[12] We find no statistically significant differences as a result of presenting concrete choices after or before endorsing the general principles. Case 1 (Chi-squared test, p=.412), Case 2 (Chi-squared test, p=.549), and Case 3 (Chi-squared test, p=.170).

that almost all of the participants (94%) chose to avoid the repugnant conclusion in Case 1 (Repugnance). We also find that almost all participants (98%) prefer to slightly harm a better well-off group, to increase welfare of the lesser off group (increasing total and average welfare) in Case 3 (Trade-offs). We find some disagreement in Case 2 (New Happy People): Surprisingly, we find 69% of participants prefer a society where more happy people are not added to the world (that do not affect the happiness levels of the original population). Interestingly, several participants cited inequality concerns as a reason for not judging Society 2 as better[13]. Even though population B is at a high level of happiness, because their happiness level is relatively lower than population A—creating inequality—participants often viewed this as undesirable.[14]

	Option 1 Choice %	**Option 2 Choice %**
Case 1 (Repugnance)	94%	6%
Case 2 (New Happy People)	69%	31%
Case 3 (Trade-offs)	2%	98%

Table 3. Endorsement frequency of each individual concrete case.

> ***Result 3:*** *The majority of participants avoid the Repugnant conclusion in Case 1(Repugnance), and maximize total and average happiness in case 3 (Trade-offs), but not in case 2 (New Happy People). They prefer not to create new happy people.*

Part 3: Reconciling Choices

Given that in Part 1, we find most participants endorse at least one general principle, it remains to be seen how many of them violate these principles with their choices in the concrete scenarios and how they respond to these violations. Among those who endorsed the respective general principle, we find that in their concrete judgements, 33% of individuals violated Average Utilitarianism, 99% violated Total Utilitarianism, 42% violated Control Total Utilitarianism, 99% violated Person Affectism, and 97% violated Pareto[15]. These results suggest that for the majority of participants, their initial intuitions about which principles to endorse and cases to choose are not internally

[13] This statement draws its evidence from a free response section at the end of the experiment, where participants explained qualitatively why they made certain choices.
[14] We find no statistically significant effects of the order at which the three cases were presented and on which moral principle was presented first, all p>.05.
[15] We count an individual as violating a principle if they choose a society in any of the three cases that conflicts with a principle they endorsed.

consistent. In other words, the majority of participants' choices are not consistent with their endorsed principles, before they have a chance to reconcile any conflicts.

Given the surprising result that in Case 2 (New Happy People) the majority of participants chose Society 1 as being better, we observe a much lower frequency in violations of Average Utilitarianism compared to the other general principles. It is less surprising to see a lower rate of violations for Control Total Utilitarianism, since intuitive choices (Case 1 (Repugnance) = Society 1, Case 2 (New Happy People) = Society 2, and Case 3 (Trade-offs) = Society 2) do not violate this not-axiom, even though following it would actually mean choosing the following (Case 1 (Repugnance) = Society 1, Case 2 (New Happy People) = Society 1, and Case 3 (Trade-offs) = Society 1).

Our main focus will be on the reconciliation frequencies and patterns between the different general principles, and the Control principle. Figure 5 shows choices in the conflict resolution block. We find that a non-trivial percentage of conflicts are not reconciled (31%).[16] However, when participants do decide to reconcile the conflict, we find that it is more common for participants to revoke the general principle they endorsed, rather than change their choice of which society they consider better. We find that participants revoke the general principle 71%[17] of the time. Furthermore, excluding the Control principle this rate is 70%, which is significantly different from a 50-50 split between revoking principles and changing choices ($p<.001$)[18]. We find that the majority of participants always decide to revoke the general principle (58%), while only 20% of participants always choose to change their choice to be consistent with the general principle. These results suggest that in general, participants prefer to revoke general principles that conflict with the choices they make in concrete scenarios.

> ***Result 4****: When conflict arises, participants prefer to revoke general principles rather than change their judgements in concrete cases to align with the principle.*

Table 4 below examines revisions by general principle across the three cases. Taking the total violations across all three cases, we report the percentage of instances where no changes were made (column 2), the general principle was revoked (column 3), and the choice of which society is better changed to be consistent with the general principle (column 4). Note that the number of

[16] We calculate the average at the individual level and treat each individual as an independent observation. In this case the proportion of participants not reconciling their conflict is calculated as $\sum^{n}_{t=1}$No Reconciliation/(Conflicts they can reconcile) for each participant.

[17] We calculate the average revocation rate at the individual level and treat each individual as an independent observation. In this case the proportion of participants revoking a general principle endorsed is calculated as $\sum^{n}_{t=1}$Revoking the principle/(Revoking the Principle + Changing the Concrete Choice) for each participant.

[18] We find no significant difference in reconciliation rates between the two ordered treatments (0.67 vs 0.74, t-test, p=.168).

observations in each general principle varies due to some principles being selected more than others, and some principles having more questions than others[19].

	n	**No Change**	**Principle Revocation**	**Case Revocation**
Average Utilitarianism	149	28%	40%	32%
Total Utilitarianism	669	34%	48%	17%
Control	171	38%	45%	17%
Person Affectism	91	34%	43%	23%
Pareto	321	29%	55%	17%

Table 4. Reconciliation choice frequency by principle for cumulative cases.

We find some heterogeneity in participants' decisions to resolve conflicts. The general principle of Average Utilitarianism is the least likely to be revoked, and participants change their choice to be consistent with this principle at the highest rate. On the other hand, both Pareto and Total Utilitarianism have the lowest rate of participants changing their choice to be consistent with these axioms. These two principles have the highest rate of participants deciding to revoke the principle at 55% and 48% respectively.[20] For a full overview of all participant decisions at reconciliation for each principle, see Figure 5.

[19] We only had one case where Pareto and Person Affectism could be violated. Whereas we had three cases where Average, Total, and Control Utilitarianism could be violated.

[20] We do not report results for the reconciliation between Averagism and Case 1 due to a coding error which made the data unusable. However, since the majority of participants avoid the repugnant conclusion in Case 1, this should not have a significant effect on our results.

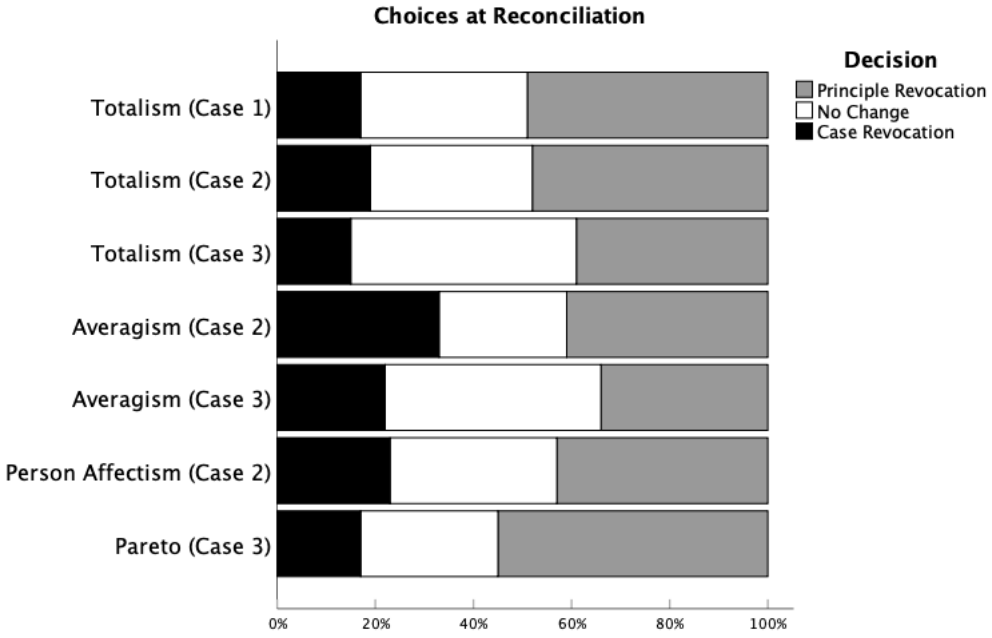

Figure 5. Choices at Reconciliation Task

Graph depicting frequency of decisions at the reconciliation phase with regard to each principle

On an individual level, we investigate how many participant decisions were consistent after reconciliation for each axiom. Table 5 below reports the survival rate of each general principle. We define the survival rate as endorsed principles that are not in conflict at the end of the experiment. In other words, how many participants chose to neither revoke the general principle they endorsed, nor let it remain inconsistent with a choice at the end of the study by changing their choice. We find that 69% of participants who originally chose to endorse Average Utilitarianism, either choose societies that are consistent with this principle, or change their choices to be consistent with this principle. We observe that this is the only general principle that has a high survival rate from final inconsistencies. All other general principles have a lower survival rate of inconsistencies. However, we still find some percentage of participants who are willing to change all the relevant choices so that the general principles they endorsed remain consistent. Unsurprisingly though, this is quite low for the principles that have concrete cases where one needs to bite the bullet on intuitively undesirable outcomes.[21]

	n	**Survival rate of endorsed principles**

[21] We find no statistically significant effects of gender on reconciliation rates, with men revoking general principles at 73.6% and women at 67.8% (p=.191).

Average Utilitarianism	459	69%
Total Utilitarianism	404	15%
Control	116	14%
Person Affectism	91	23%
Pareto	321	17%

Table 5. Survival rate of endorsed principles.

Result 5: *Average Utilitarianism is the only general principle with a high survival rate. All other principles have a low survival rate.*

Discussion

Our main result is that when participants' choices result in a conflict between their endorsed abstract principles and their judgements on concrete cases, they prefer to revoke their previously endorsed principle rather than changing or revoking their judgement regarding the concrete population ethical case. Our findings are directly relevant to theorizing about reflective equilibrium. Specifically, we take these results to indicate that for lay moral reasoning, case judgements do play a major revisionary role, which is an empirical finding that we take to be relevant for further theory building. While we find that some participants want to maintain consistency with the abstract principles, the evidence shows that participants do put more weight on their concrete choices. This data thus provides direct evidence in favor of the structure of reflective equilibrium reasoning as being such that concrete judgements play a revisionary role.

Returning to our potential patterns of results introduced earlier, the presented data clearly support *Primacy of the Concrete*, i.e. the claim that abstract intuitions are typically revised in favor of concrete ones. This pattern of results fits well with standard theorizing of professional philosophers, where any given ethical theory is often directly criticized by an appeal to a certain concrete case that stands in conflict with that theory. As Huemer (2008) suggests, often "our intuitions about the cases are stronger than our initial inclination to accept the moral theory" (Huemer 2008, 391). This should then result in concrete cases (like thought experiments) playing a revisionary function in philosophical discourse. While we are not able to show whether this is a feature of professional philosophers' theorizing, we present evidence for this pattern in lay moral theorizing. This suggests that the primacy of the concrete may be a more general feature and that intuitions about concrete cases play a strong revisionary function in folk morality.

Our design does not allow for an exact answer on why exactly this might be so. Answering this question will require further empirical work untangling the different mechanisms and additional theorizing on the matter. However, we suspect that a possible reason may be that concrete intuitions are taken to be stronger intuitions overall (and that stronger intuitions ought to revise weaker intuitions). Alternatively, it may be the case that participants are not yet aware of most of the philosophical implications of endorsing general rules and as a result are more likely to give them up after observing these implications in the concrete cases (similar to a learning effect).

Regarding what types of choices are available in situations of such conflict, Sinnott-Armstrong (2008) raises the possibility that that we may just "need to learn to live with conflicting intuitions" (Sinnott-Armstrong 2008, 226). Drawing on Pyrrhonians and their "feel[ing] the pull of both sides [while] reconigiz[ing] that they cannot have it both ways" (Sinnot-Armstrong 2008, 226), he claims that any resolution in favor of either the concrete or the abstract resolutions might not be possible as one might not be able to "imagine how any resolution would succeed or even proceed". Our data also provide some evidence for this being a descriptively correct picture of lay behaviour, as we find that across all relevant conflicts, approximately 30% of participants neither revoked their endorsed principle nor their judgement in the concrete case. Rather, they indicated that they would like to continue in the experiment with the conflict in place, which would be in line with the possibility raised by Sinnott-Armstrong (2008).

Importantly, the present study only reports data from one area of moral philosophy, namely population ethics. Any generalisations based on these data alone ought to be taken with an appropriate amount of caution. And it may be the case that the *Primacy of the Concrete* does not hold in other domains of philosophy. Given the nature of some of the population ethics cases—such as the repugnant conclusion—judgements may be stronger about these concrete cases than in other domains. Having said that, we do argue that despite the limited scope of topic, the present picture does present data on lay moral reasoning that is both interesting in itself and opens up further avenues of research.

Following Nielsen & Rehbeck (2020), we also do not remove those participants who endorsed the Control principle (which states that one society is better if and only if the total sum of happiness is lower than in the other). Given that the Control was endorsed at a significantly lower rate we take this as overall evidence that, at least by and large, most participants showed an adequate understanding of the tasks at hand and that our results are not primarily the result of acquiescence bias.

As a secondary interest, we also tested whether presenting participants with the abstract principles first and then the concrete cases or the reverse changes their endorsement rates of these principles. We found no statistically significant effects in the Control, Person Affectism, or Pareto,

though for both versions of Utilitarianism we did find order effects. This drop in endorsement rates provides further evidence for the case above that once participants are presented with some concrete cases that they can form judgements on, they are less likely to endorse the principles (and if they already endorsed them, more likely to revoke their endorsement). This adds to both the literature on order effects in social psychology and experimental philosophy, as well as to our understanding of folk utilitarian morality. Further, this also strengthens the above claim that perhaps it is participant's inability to foresee the implications of their endorsed principles that make them discard the principles when confronted with their judgements on concrete cases, which would be consistent with a type of learning effect. However, since this occurs only for the Utilitarianism principles and not the Pareto principle—even though Pareto was revoked at the highest rate—suggests that maybe the implications of the Utilitarian principles were made the most salient by our cases. Alternatively, another explanation is that since the endorsement of Utilitarian principles was much higher than Pareto, it left more room for participants to realize the implications of these principles, thus not endorsing it in the reverse order treatment.

As an additional tertiary result, we also report survival rates of endorsed principles. Interestingly, Average Utilitarianism showed a survival rate of 69% compared to all others who had rates ranging from 15% to 23%. While we do think that this is a potentially interesting finding possibly pointing to the importance that is ascribed to equality when making case judgements in population ethics (perhaps best exemplified by 69% of participants choosing (I) in Case 2 (New Happy People)), it is important to point out that we did not select principles and cases to test specifically for this effect. In other words, it may be the case that this survival rate difference is an artefact based on the choices of our cases. For example, one might argue that Case 1 (Repugnance) is a much stronger counterexample to Total Utilitarianism than any of the other cases would be for Average Utilitarianism. While we thought Case 2 (New Happy People) would be a good counterexample to Average Utilitarianism, participants stated a preference to avoid inequality between populations in potential worlds. This finding is similar to Caviola et al., (2021) in that people are intuitively attracted to judgements that are consistent with Average Utilitarianism even though most philosophers would not expect this. As a result, conflict between choices in Case 2 (New Happy People) and those who endorsed Average Utilitarianism occurred at a much lower rate. Instead, it is quite plausible that if we provided a utility monster example, we might have found a much higher rate of participants revoking Average Utilitarianism. Alternatively, a case that focuses on harms instead of happiness may also create conflicts for those who endorsed Average Utilitarianism. It stands to reason that if we had cases like this, the difference in survival rate might not be as stark and one ought to be reasonably sceptical of this result until such data are provided in further research.

Interestingly, our results differ to Nielsen & Rehbeck (2020), who find that people are significantly more likely to revise their choices to be consistent with axioms in Expected Utility Theory. Our results may differ due to consistency with axioms being seen as less desirable in our context, or due to the nature of the impossibility theorems that arise in population ethics, but not in lottery cases. Outside of our results relating to Average Utilitarianism, the low survival rate of the other axioms—which did not occur at a higher rate than the control—provides empirical evidence for how difficult it is to have a consistent population axiology and that impossibility theorems create real constraints for lay-people's judgements and decisions.

Conclusion

In this paper, we provide novel evidence for lay reflective equilibrium reasoning in the context of population ethics. We conduct an experiment that investigates what happens when a lay population's endorsed abstract principles conflict with their intuitions about concrete cases. We find that participants prefer to revoke their abstract principles rather than their judgements about concrete cases. In other words, intuitions about concrete population ethical cases play a larger revisionary function with regard to intuitions about abstract principles. Further, we also find an order effect in that participants are less likely to endorse Average and Total Utilitarianism if they are first presented with specific population ethical cases compared to when they are first shown the abstract principles. Our results create a platform for future research. This includes extending the understanding of reflective equilibrium within population ethics, as well as in other domains of philosophy.

References

Arrhenius, G. (2000). An Impossibility Theorem for Welfarist Axiologies. *Economics & Philosophy*, *16*(2), 247-266.

Brun, G. (2014). Reflective Equilibrium Without Intuitions? *Ethical Theory and Moral Practice*, *17*(2), 237-252.

Buckwalter, W., & Stich, S. (2014). Gender and Philosophical Intuition. *Experimental Philosophy*, *2*, 307-346.

Cappelen, H., & Plunkett, D. (2020). Introduction: A Guided Tour of Conceptual Engineering and Conceptual Ethics. In *Conceptual engineering and conceptual ethics* (pp. 1-34). Oxford University Press.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we Punish? Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology*, *83*(2), 284-299.

Caviola, L., Althaus, D., Mogensen, A., & Goodwin, G. (2021). Population Ethical Intuitions. *Cognition*.

Caviola, L., Schubert, S., & Mogensen, A. (2021). Should you Save the More Useful? The Effect of Generality on Moral Judgments about Rescue and Indirect Effects. *Cognition*, *206*, 104501.

Daniels, N. (2016). *Reflective Equilibrium.* The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/reflective-equilibrium/

De Brigard, F., Mandelbaum, E., & Ripley, D. (2009). Responsibility and the Brain Sciences. *Ethical Theory and Moral Practice*, *12*(5), 511.

Greaves, H. (2017). Population Axiology. *Philosophy Compass*, *12*(11), e12442.

Huemer, M. (2008). Revisionary Intuitionism. *Social Philosophy and Policy, 25*(1), 368-392.
Machery, E., Stich, S., Rose, D., Alai, M., Angelucci, A., Berniūnas, R., ... & Cohnitz, D. (2017). The Gettier Intuition from South America to Asia. *Journal of Indian Council of Philosophical Research*, *34*(3), 517-541.
McPherson, T. (2015). The Methodological Irrelevance of Reflective Equilibrium. In C. Daly (Ed.) *The Palgrave Handbook of Philosophical Methods* (pp. 652-674). Palgrave Macmillan, London.
Murray, D., & Nahmias, E. (2014). Explaining Away Incompatibilist Intuitions. *Philosophy and Phenomenological Research*, *88*(2), 434-467.
Nahmias, E., & Murray, D. (2011). Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions. In *New Waves in Philosophy of Action* (pp. 189-216). Palgrave Macmillan, London.
Nichols, S. (2011). Experimental Philosophy and the Problem of Free Will. S*cience*, *331*(6023), 1401-1403.
Nichols, S., & Bruno, M. (2010). Intuitions about Personal Identity: An Empirical Study. *Philosophical Psychology*, *23*(3), 293-312.
Nichols, S., & Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Nous*, *41*(4), 663-685.
Nielsen, K., & Rehbeck, J. (2020). When Choices Are Mistakes. *Working Paper*.
Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.
Paulo, N. (2020). The Unreliable Intuitions Objection Against Reflective Equilibrium. *The Journal of Ethics*, 1-21.
Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.
Scanlon, T. (2002). Rawls on Justification. In S. Freeman (Ed.) *The Cambridge Companion to Rawls* (pp. 139-167). Cambridge University Press.
Sinnott-Armstrong, W. (2008). Abstract + Concrete = Paradox. In J. Knobe & S. Nichols (Eds.), *Experimental Philosophy* (p. 209–230). Oxford University Press.
Tersman, F. (2018). Recent Work on Reflective Equilibrium and Method in Ethics. *Philosophy Compass*, *13*(6), 1-10.
Zuber, S., Venkatesh, N., Tännsjö, T., Tarsney, C., Stefánsson, H. O., Steele, K., ... & Asheim, G. B. (2021). What Should We Agree on About the Repugnant Conclusion? *Utilitas*, 1-5.

Appendix – Experimental Material

Appendix A- Instructions

1. Introduction

On the next page you will be asked some comprehension questions about the information on this page.

For each question you answer correctly you will receive a bonus of £0.1

Please read the following carefully:

In this survey we will talk about the concepts 'hypothetical societies' and 'happiness levels.' Please read carefully while we explain these two concepts.

Hypothetical Societies:

You can think of a society as a world. Simply put, a society is a number of individual humans taken together. In our examples, different hypothetical societies cannot interact with each other: Think of them as inhabiting different planets without the ability to travel between them.
We can call a group of people within one society a 'population', and a society might consist of several populations. For example a society could have populations 'A' and 'B.' **If there are two societies 'I' and 'II', and both have a population called 'A,' you can understand these two populations 'A' as consisting of the exact same people**. Also, if a society 'I' has two populations, 'A' and 'B,' these two populations include all people living in this society.

Happiness Levels:

You can think of a happiness level as a measure of how well-off a person is overall. This includes the satisfaction of basic needs like food and shelter but also things like friendship and meaning, all being summarised in a single number.

That is, if we say that someone has negative happiness/quality of life (below zero), you can think of this as a person that is miserable (does not have shelter, goes hungry, is in pain, etc). If we say that someone has positive happiness/quality of life (above zero), you can think of this as a person that is flourishing (has shelter and plentiful food, as well as friends and further entertainment). If we say that someone has neutral happiness (zero), you can think of a person whose life is neither bad nor good.

Differences in these happiness levels are there to compare two persons. **For example, if person P1 has a happiness level of 10, and person P2 has a happiness level 20, then person P2's life is, all things considered, much better than person P1's.**

Changing Societies

If both societies I and II have the same population (A) then they are made up of the exact same people. If we start in Society I, and the world changes to Society II, the happiness levels of Population A, will either increase, decrease, or stay the same depending on if it is higher, lower, or the same in Society II compared to Society I.

2. **Rule Task**

Task:

On the next page you will find a number of rules.

Please read each rule carefully:

You will be asked to either 'endorse' or 'not endorse' the rule. These rules relate to how good a hypothetical society is. If you think that a rule is true and you agree with it, please press 'endorse.' If you think that a rule is false and you disagree with it, please press 'not endorse.' If you are unsure, please also press 'not endorse.'

Please click here to access the instructions and definitions. You can leave it open for as long as you need.

3. **Case Task**

You will now be presented with a number of examples. As shown at the beginning of the survey, you will be presented with two societies in (I) and (II) that differ from each other on a couple of dimensions.

All examples are represented in a graphic and are also accompanied by numerical information. **You are asked to decide which of the two societies (I) or (II) is better, all things considered.**

4. **Resolution Task**

Final Task:

On the next page you will be asked some comprehension questions about the information on this page. For each question you answer correctly you will receive a bonus of £0.1

Some of the choices you have made about which society is better contradict a Rule that you endorsed. This means that your choices between which of the two societies are better violates a rule that you endorsed earlier.

In this section you will now be able to address these violations in a variety of ways.

For each violation you will have three options :

1) You can revoke your endorsement of a previously endorsed Rule

2) You can change your choice about which society is better

3) You can decide to keep your endorsed Rule and keep your original choice and continue in the study

Below you will find the violations that occurred for each of the choices you made.

Appendix B- Demographics Summary.

Demographics	**n**	**Percentage**
Gender		
Female	377	69.4%
Male	161	39.7%
Other	5	0.9%
Age		
18-29	170	31.3%
30-39	170	31.3%
40-49	123	22.7%
≥50	80	14.7%
Education Level		
Some High School	12	2.2%
High School	192	35.4%
Bachelor's Degree	212	39.0%
Master's Degree	116	21.4%
PhD or Higher	11	2.0%
Philosophy Background		
None	505	93.0%
Some Philosophy (Uni level)	33	6.1%
Bachelor's Degree in Philosophy	5	0.9%
Total n	543	