Cognition 218 (2022) 104941

Contents lists available at ScienceDirect

Cognition

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

Full Length Article

Population ethical intuitions

Lucius Caviola [a,*], David Althaus [b], Andreas L. Mogensen [c], Geoffrey P. Goodwin [d]

[a] Department of Psychology, Harvard University, Cambridge, MA, United States of America
[b] Center on Long-Term Risk, London, United Kingdom
[c] Global Priorities Institute, University of Oxford, Oxford, United Kingdom
[d] Department of Psychology, University of Pennsylvania, Philadelphia, PA, United States of America

ARTICLE INFO

Keywords:
Happiness
Suffering
Moral judgment
Population ethics
Axiology

ABSTRACT

Is humanity's existence worthwhile? If so, where should the human species be headed in the future? In part, the answers to these questions require us to morally evaluate the (potential) human population in terms of its size and aggregate welfare. This assessment lies at the heart of *population ethics*. Our investigation across nine experiments ($N = 5776$) aimed to answer three questions about how people aggregate welfare across individuals: (1) Do they weigh happiness and suffering symmetrically?; (2) Do they focus more on the average or total welfare of a given population?; and (3) Do they account only for currently existing lives, or also lives that could yet exist? We found that, first, participants believed that more happy than unhappy people were needed in order for the whole population to be net positive (Studies 1a-c). Second, participants had a preference both for populations with greater total welfare and populations with greater average welfare (Study 3a-d). Their focus on average welfare even led them (remarkably) to judge it preferable to add new suffering people to an already miserable world, as long as this increased average welfare. But, when prompted to reflect, participants' preference for the population with the better total welfare became stronger. Third, participants did not consider the creation of new people as morally neutral. Instead, they viewed it as good to create new happy people and as bad to create new unhappy people (Studies 2a-b). Our findings have implications for moral psychology, philosophy and global priority setting.

Imagine a world containing 2.5 billion people, out of which 1.5 billion live unhappy lives and 1 billion live happy lives.[1] Now imagine that forty years later this world contains over 5 billion people, out of which 2 billion live unhappy lives and 3 billion live happy lives. Focusing just on the number of happy and unhappy people and setting aside other considerations, has the world improved or not? This is a question of population ethics. In this paper, we investigate lay people's population ethical intuitions.[2]

1. Approaches to population ethics

Population ethics deals with questions that arise when the composition or size of the population varies across different outcomes, e.g., when one or more additional people (with different identities) could be born (Greaves, 2017). The current paradigm in population ethics was established by the philosopher Derek Parfit in his 1984 book *Reasons and Persons*. Even though population ethics is a relatively young field, it is

* Corresponding author.
E-mail address: lucius.caviola@gmail.com (L. Caviola).

[1] These numbers roughly reflect the overall human population sizes and the number of people living in extreme global poverty vs. not living in extreme global poverty in 1950 and 1990 respectively (Roser & Ortiz-Ospina, 2020). Of course, it is by no means the case that all people living in extreme poverty live unhappy lives and all people not living in extreme poverty live happy lives (e.g., Biswas-Diener & Diener, 2009; Diener & Diener, 1996; Diener, Oishi, & Tay, 2018). Note also that since 1990 the number of people living in global poverty has more than halved to 734 million and the total human population size has increased to 7.8 billion in 2020.

[2] Philosophers disagree on the nature and role of intuitions in philosophy. In one conception, intuitions in philosophy are just beliefs. Lewis (1983, p. x) famously writes: "Our 'intuitions' are simply opinions; our philosophical theories are the same." However, many philosophers wish to understand intuitions as something different from and prior to belief (Bealer, 1998; Huemer, 2005; Chudnoff, 2013). Others, like Cappellen (2012), are skeptical of the significance of 'intuition'-talk in philosophy. By and large, our studies merely ask subjects to report their beliefs. It seems to us reasonable to suppose that the beliefs reported will typically have arisen unreflectively as a result of certain claims appearing to our subjects as correct, but the studies we report cannot speak to views that distinguish between people's beliefs and their intuitions, with the plausible exception of the thinking-style manipulations in 3c and 3d.

https://doi.org/10.1016/j.cognition.2021.104941
Received 10 April 2021; Received in revised form 9 October 2021; Accepted 11 October 2021
Available online 5 November 2021

highly relevant today and has direct implications for decision-making. It is especially relevant to policy questions regarding humanity's intermediate and long-term future, including those related to population growth and control, as well as to factors that could increase or decrease the risk of human extinction (Bostrom, 2013; Greaves et al., 2020). Also relevant are questions of individual decision making, such as whether to have children or not.

While population ethics is a broad field, in this paper we will focus on how people evaluate world states that differ in the number of happy and unhappy people they contain — that is, in terms of the welfare of their constituent members. To evaluate a pair of world states (or populations), one needs to order them in terms of their aggregate well-being, in order to determine which world has the higher aggregate well-being. However, this is by no means a straightforward task. In fact, it generates a series of deeply complex subsidiary questions that lie at the heart of population ethics, and that comprise the topic of our present research.

One question is whether, in aggregating welfare, happiness and suffering should be weighed equally (or symmetrically). A second question is whether one should consider the average or total level of well-being of the population. And a third question is whether the overall value of a population should be thought about only in terms of people who actually exist, or whether it should also account for people who could exist. Though distinct, these questions all trace back to the same central theme — how we should aggregate welfare (or well-being) when deciding whether a given population is better or worse than another.

2. Weighing happiness against suffering

Our first research question is how people weigh and trade happiness (i.e., positive welfare) and suffering (i.e., negative welfare) in evaluating populations. Do people think that a greater number of happy people is required to outweigh a certain amount of unhappy people in order for a population to be net positive? Although at first blush, it may seem intuitive to weigh happiness and suffering equally, there is also a strong pull towards weighing suffering more than happiness, as we describe presently.

In addressing this issue, we focus on whether people's intuitions align with the verdicts of different *utilitarian* theories (or utilitarian *population axiologies*). We focus on theories that rank variable population outcomes in virtue of the happiness and suffering experienced by the people in those outcomes (i.e., hedonistic welfarist theories). This point of departure seemed reasonable to us, because even if some (or many) people are not utilitarians, most people presumably value certain states of the world more than others and do so in ways that take into account the happiness and suffering contained in these world states when evaluating them. For example, a world filled with happiness is presumably considered as better than a world filled with suffering. In that sense, many people plausibly hold certain broadly utilitarian axiological intuitions.[3]

Utilitarian theories differ in how they weigh happiness and suffering in populations with varying sizes. *Classical utilitarianism* is symmetrical in that it weighs happiness and suffering the same. With respect to the example we introduced at the beginning of this paper, this suggests that the world is better 40 years later when 2 billion unhappy lives and 3 billion happy lives exist because the total amount of happiness outweighs the total amount of suffering. *Negative utilitarianism*, in contrast, weighs suffering more than happiness. Strict negative utilitarianism is the view on which one population is better than another if and only if it contains a smaller total of suffering, without taking into account the amount of happiness. This means that according to strict negative utilitarianism the world in our example becomes worse when the number of people with unhappy lives increases to 2 billion. To the best of our knowledge, no existing research has examined whether ordinary individuals' views align more closely with classical or negative utilitarian theories. Accordingly, our first aim was to address this question for populations that were described as already existing.

3. Averagism vs. totalism

A second key normative question concerns whether one should evaluate populations based on their average level of happiness or their total sum of happiness (Sidgwick, 1981, p. xxxvi), or both. The first view is referred to as *averagism*. The second view is referred to as *totalism*. The contrast between these views represents an important schism within moral philosophy. According to averagism, a world is better if it contains a small population of people that are on average extremely happy than a much larger population that is on average very slightly less happy. According to totalism, the second world would be better.

Resolving which of these theories is correct has proved to be an especially difficult philosophical problem because both lead to conclusions that many would consider unacceptable. For example, according to averagism, a world with 10 extremely unhappy people could be improved by adding an 11th unhappy person, as long as this person is slightly less unhappy than the current average, thereby increasing the average happiness level of the population. Because of these and other counterintuitive implications (cf. Greaves, 2017) moral philosophers tend to reject averagism. Despite this, averagism has been endorsed by several prominent researchers, such as Hardin (1968) in 'The Tragedy of the Commons'.

Totalism can also lead to conclusions that many consider counterintuitive. The most famous example is the so-called Repugnant Conclusion formulated by Parfit (1984): according to totalism, a population of perfectly happy people would be less good than a world with an enormous number of people whose lives are barely worth living and thereby just slightly positive. If the number of people in the second world is large enough, the total amount of happiness will be larger than that of the first world.

How to make philosophical progress on this issue is not obvious. Indeed, it may be that no way exists to jointly preserve the strong intuitions that underlie each of these approaches (Greaves, 2017). Yet, this does not absolve us of the need to make judgments of the relative values of different populations, nor of the need to make policy choices between them. The second aim of our investigation was to determine whether ordinary individuals' intuitions more typically align with averagist or totalist theories, or some blend of the two.

4. The intuition of neutrality and the asymmetry

One consequence of totalist views is that adding a new person to a world is a good thing, as long as the new person's well-being is higher than neutral (Greaves, 2017). And, since adding a new happy person is a good thing, then it might seem to follow that we have a moral obligation to create new happy people. This same consequence is also implied by averagist theories in cases where the new person's happiness exceeds the population average. The reason for this is that the respective utilitarian theories (totalist and averagist) take into account the happiness and suffering of future people who do not exist yet but could exist in the future.

However, this consequence has not appealed to many philosophers.

[3] This of course does not mean that people are utilitarians. First, many people are likely not pure welfarists. They also value other things in addition to happiness and suffering (or other factors that contribute to good or bad lives), such as knowledge, beauty or complexity. But we limit our research to welfarist aspects. Second, we know from previous research that in their ethical decision-making (i.e., *deontics* as opposed to *axiology*) people often follow deontological constraints that prohibit them from bringing about the utilitarian outcome (e.g., the footbridge trolley dilemma, Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). That is, even if people consider one world state better than another, they might still not consider it morally required or permissible to actively choose the option that leads to that world state.

Narveson (1973: 80) famously writes: "We are in favor of making people happy, but neutral about making happy people." The latter part of this slogan has been called the *intuition of neutrality* (Broome, 2005). While the view that creating additional happy lives is neutral is generally considered intuitive by moral philosophers, the same cannot be said for the view that there is nothing bad about adding lives in which suffering predominates. According to the *asymmetry*, the addition of people with happy lives to the population is intrinsically neutral, whereas the addition of people with unhappy lives to the population is intrinsically bad (McMahan, 1981).[4] Our interest was in whether people accept the intuition of neutrality, and whether their views align with the asymmetry. Do they regard the addition of suffering people as worse to a greater extent than the addition of happy people is good?

Because this last question relates to our first question (on the weighing of happiness and suffering), we order the sequence of our investigation as follows: In our first studies, we examine how people weigh happiness and suffering for existing populations — do they do so symmetrically, or asymmetrically? In our next studies, we examine how people evaluate the addition of new people. Do they regard the addition of happy people as good and the addition of suffering people as bad? If so, do they do so symmetrically, or asymmetrically? Our final set of studies examine whether people typically hold totalist or averagist intuitions. In addressing these questions, we hoped to make systematic progress in revealing the way people think about the aggregation of population well-being, and thereby, the way they think about population ethics.

5. Previous psychological research

To our knowledge there exists almost no published psychological research on population ethical intuitions. In most previous research on moral judgment and decision making, population sizes were usually not systematically varied. However, psychological research on related topics, as well as research in other fields (e.g., economics), is informative for our research questions.

As for our first research question of how people trade happiness against suffering, existing research on negativity bias could be relevant (for reviews, see Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001). Negativity bias refers to the phenomenon that negative events (e.g., losing money or having bad social interactions) have a greater impact on our behavior, cognition, and emotions than do positive events (e.g., gaining money or having good social interactions). Negativity bias has been observed in numerous domains, including financial decision-making (Kahneman & Tversky, 1979, 1984), impression formation (e.g., Peeters & Czapinski, 1990), affective processing (e.g., Ito, Larsen, Smith, & Cacioppo, 1998), and social interaction (e.g., Fiore, Becker, & Coppel, 1983; Manne, Taylor, Dougherty, & Kemeny, 1997). However, there exists less research on negativity bias in the moral domain. One exception is the finding that immoral behaviors have a greater influence on moral character evaluation than do moral behaviors (Birnbaum, 1972, 1973; Riskey & Birnbaum, 1974). For example, a person who has committed two highly immoral acts but many more highly moral acts is still considered immoral (Riskey & Birnbaum, 1974).

Other research provides mixed evidence regarding the application of negativity bias to moral judgments. In a variation of the one-shot dictator game, Tappin and Capraro (2018) found that the preference to "do good" was as strong as the preference to "avoid bad", suggesting that negativity bias does not affect moral judgment in that context. Most directly relevant to our research question is a pilot study ($N = 14$) by Rozin and Royzman (2001, p. 307), in which participants were asked to imagine that they could press a button that would give one minute of "intense pain" to one person and, in version A, 10 min of "similarly intense pleasure to another person or, in version B, one minute of "similarly intense pleasure" to 10 other people. Every participant refused to press the button. The participants were then asked to state the minimum number of minutes (version A) or people (version B) that would make them press the button. In version A, the lowest number was 800 min, and in version B, all participants except one stated that no number would be large enough. These findings suggest that people consider suffering to be much worse than happiness is good. This study, however, comprised only 14 participants, and it also concerned people's hypothetical behavioral intentions—what they themselves would be willing to do—whereas our focus is on more basic value judgments about which states of the world are better or worse (referred to in philosophy as "axiological judgments"), regardless of people's willingness to bring about those states of the world.

The economist Spears (2019) examined people's intuitions regarding the asymmetry. Participants were presented with a scenario about a couple considering whether to have another child. Participants were asked to rate whether the hypothetical facts that the child's life would either be full of suffering or full of happiness were morally relevant to the couple's decision. Nearly 75% of participants thought that both facts are morally relevant. Only 15% of participants supported a strict asymmetry and thought that only the fact about suffering but not the fact about happiness was morally relevant. This could be seen as evidence that most people do not endorse the asymmetry, at least in its strict form. However, the study's approach was rather indirect since it did not directly ask participants how good or bad they would consider it if a happy or unhappy child were to be born, so it is possible that participants would have endorsed a more moderate version of the asymmetry, according to which suffering is worse than happiness is good.

As for our second research question of whether people focus on the average or total amount of happiness, there is no directly relevant research published in psychological journals. Nonetheless, some relevant research exists.

One investigation was conducted by Starmans & Bloom, 2015. In one study, the authors presented participants with a "happiness scale" ranging from 1 (very sad) to 5 (very happy). They found that participants' normative judgments about distributions of happiness across different people in a population (i.e., inter-individual) were similar to their normative judgments about distributions of happiness within the lifetime of one individual (i.e., intra-individual). The authors also found that the vast majority of participants preferred one person on happiness level 5 to 15 people on happiness level 1. It is unclear, though, whether this finding provides evidence for the claim that participants followed averagism—a possible explanation considered by the authors. Instead, it is plausible that participants interpreted level 1 (very sad) as negative welfare in which case their preferences would be consistent with both totalism and averagism.

Another relevant set of studies was conducted by Spears (2017). In the first three studies, participants chose between smaller populations with higher average incomes and larger populations with lower average incomes. Participants valued both increased population size as well as increased average income, suggesting a focus on both total and average welfare. One important limitation of these studies is that they focused on different levels of *income* and not *happiness* directly. It is unclear how happy or unhappy participants perceived the people with different

[4] One way to derive the intuition of neutrality is by assuming a *person-affecting axiology* (cf. Bader, 2021; Parfit, 1984; Greaves, 2017; Beckstead, 2013; McMahan, 1981; Arrhenius, 2000), according to which one outcome is better (worse) than another only if it is better (worse) for someone. If we deny *existence comparativism* and maintain that one outcome is better (worse) than another for someone only if that person exists in both outcomes (Broome, 2004), then the intuition of neutrality follows. It is possible to derive the asymmetry given a person-affecting axiology if one accepts *asymmetric comparativism*, which says that a person cannot be better off as a result of being brought into existence with a good life, as opposed to never existing, but can be worse off as a result of being brought into existence with a bad life, as opposed to never existing. As demonstrated by Nebel (2019), asymmetric comparativism is significantly more defensible than may be immediately apparent.

income levels to be.

In another study within the same economics paper, participants ranked six different populations consisting of either "10 billion" or "many more" people, leading either "bad", "good" or "excellent lives". Participants were allowed to be indifferent between options. 46% of participants thought that a larger population living excellent lives is better than a smaller population living excellent lives. And only 35% thought that a larger population living good lives is better than a smaller population living good lives. By contrast, 69% thought that a smaller population living bad lives is better than a larger population living bad lives. That is, responses tended to be more in line with totalism in cases of negative lives compared to positive lives. 12% of participants were indifferent between smaller and larger populations with the same level of happiness, demonstrating that only a minority of responses were completely in line with averagism. Overall, these results show that participants' responses were neither consistently in line with totalism nor averagism. One limitation of this study is that the number "many more" and the welfare difference between "good" and "excellent" lives were not further quantified. As a result, it is not clear if a strict totalist and a strict averagist would of necessity always give different responses. Another limitation is that the study did not directly examine the relative strength of participants' preferences. Finally, it is likely that the results were confounded by participants' concerns about overpopulation, as expressed in some of their comments.

Two other psychological findings may also be relevant to the question of whether people favor average or total welfare, because they both reveal cases in which people do not seem to focus on overall magnitudes (i.e., totals). One such finding is the proportion dominance effect, which refers to people's greater sensitivity to proportions compared to absolute numbers (e.g., Baron, 1997; Fetherstonhaugh, Slovic, Johnson, & Friedrich, 1997). For example, in one study by Bartels (2006), participants preferred saving 225 out of 300 people over saving 230 out of 900 people. That is, people focus on relative savings, sometimes even at the expense of absolute savings. This response seems to be partly driven by erroneous deliberative thinking (Bartels, 2006; Mata, 2016) and diminishes upon reflection (Bartels, 2006). Because the proportion dominance effect involves deprioritizing total magnitudes, it seems possible that some of the psychological mechanisms that drive it also incline people to focus on the average rather than total level of happiness when making population ethical judgments.

In a similar vein, another potentially relevant psychological phenomenon is scope insensitivity (Frederick & Fischhoff, 1998)—people's tendency to neglect the size of a problem when evaluating it. For example, research has shown that, at least in separate evaluation, people are willing to pay roughly the same amount of money to help either 2000 birds, 20,000 birds or 200,000 birds (Desvousges et al., 1992). It is possible, therefore, that scope insensitivity weakens people's preferences for larger over smaller populations.

As this review indicates, while past research has yielded suggestive evidence, no past psychological research has directly answered the questions we began with. We therefore sought to undertake more direct tests of ordinary people's population ethics, with a particular focus on the positive-negative asymmetry for existing populations as well as for the addition of new people, and the trade-off between totalism and averagism—three central issues within the moral philosophy of this topic.

6. The present research

In this paper, we present nine experiments which test our three overarching research questions. First, how do people trade unhappiness against happiness when assessing the value of a population (Studies 1a-c)? Do people value the addition of new people as morally neutral or not (Studies 2a-b)? Third, are people focused on the average or total happiness level of a population (Studies 3a-d)?

Note that we are primarily interested in people's *axiological* population ethical judgments, i.e., their moral evaluation of populations. We measure axiological judgments in different ways. One way is to ask participants to directly assess the overall value of a population in absolute terms, where a population has a net positive value if one thinks it's better for it to exist rather than not exist (Studies 1a-c). A slightly different way is to ask participants about the relative value of two populations. This can be done by asking i.) which out of two populations is better or worse (Studies 2a-b), ii) which it would be better to have come into existence (Studies 3c-d), or iii) which of the two one would prefer to come into existence (Studies 3a-b). For simplicity, we do not strictly differentiate between these different assessment techniques when discussing our findings (cf. the Limitations section of this paper).

6.1. Participant recruitment

For all reported studies apart from Study 2b, we recruited participants from Amazon MechanicalTurk (MTurk). We collected the data through the platform Positly, which is a front-end platform that recruits MTurk participants. Positly includes additional proprietary quality metrics (https://www.positly.com/participants/). Concretely, Positly by default blocks duplicate and suspicious IP addresses, requires an approval rate of above 96% and at least 500 HITs, and requires participants to consistently pass attention checks. For Study 2b, we recruited participants through Prolific (US nationals, at least 98% approval rate, at least 100 prior Prolific submissions).

6.2. Open science

Reports of all measures, manipulations, and exclusions, as well as all data, analysis code, and experimental materials are available for download at https://osf.io/qt65w/.

6.3. Ethics statement

For all studies, relevant ethical guidelines were followed and the research was approved through University of Oxford's Central University Research Ethics Committee, with the reference number MS-IDREC-R56657/RE002, and Harvard University's Internal Review Board, with the reference number IRB14-3025.

7. Study 1a: trading happiness against suffering

In Study 1a, we investigated how people trade happiness against suffering, both across people in a population (i.e., inter-individual) and within an individual's lifetime (i.e., intra-individual). For a population with a given size, what percentage of happy vs. unhappy people is required for people to believe that the population is overall positive rather than negative (i.e., better to exist rather than not to exist)? Or, for an individual, what percentage of happy vs. unhappy moments must a person experience during their whole life for people to regard their life as overall positive rather than negative? Based on previous research on negativity dominance Baumeister et al., 2001; Rozin & Royzman, 2001) outlined in the introduction, we hypothesized that participants believe that more happiness is needed to outweigh a given amount of suffering (i.e., unhappiness). Our study was pre-registered at https://aspredicted.org/9ch28.pdf and had two between-subjects conditions: inter-individual vs. intra-individual.

7.1. Method

7.1.1. Participants

We recruited 529 US American participants online via MTurk ($0.35 payment per participant). 55 were excluded, leaving a final sample of 474 people (230 female, $M_{age} = 38.49$, $SD_{age} = 11.99$). A priori power analysis showed that 505 participants were required to detect an effect size of $d = 0.25$, α of 0.05, power of 0.8, two-tailed. The effect size was

estimated based on previous pilot studies. We aimed to recruit 520 participants to account for any exclusions. Sample size was determined before data collection.

7.1.2. Procedure and materials

Participants in the inter-individual condition were asked to imagine a world that contains 1000 people. "*People in this world comprise one of two types: they are either extremely happy or they are extremely unhappy. The people who are happy consistently have extremely positive experiences, similar to the feeling of falling in love. The people who are unhappy consistently have extremely negative experiences, similar to the feeling of being tortured.*" Participants in the intra-individual condition read: "*All people in this world experience a mixture of both extreme happiness and extreme unhappiness, meaning that they are extremely happy some of the time, and extremely unhappy some of the time. When they are happy, they have extremely positive experiences, similar to the feeling of falling in love. When they are unhappy, they have extremely negative experiences, similar to the feeling of being tortured. Furthermore, everyone in this world experiences the exact same ratio of happiness to unhappiness.*"

Participants were then asked the following question: "*Given this information, what percentage of happy and unhappy people[time]would there have to be for you to think that this world is overall positive rather than negative (i.e., so that it would be better for the world to exist rather than not exist)? In my view, the percentage of happy vs. unhappy people [time] would need to be as follows: _ % happy people [time] (experiencing positive feelings similar to falling in love) _ % unhappy people [time] (experiencing negative feelings similar to being tortured)*".

Afterwards, participants were asked the following questions, which we included for exploratory purposes: "How happy do you feel right now at this present moment?", "How happy do you feel in general in life?", "Would you be willing to relive the worst day of your life if doing so would enable you to relive the happiest day of your life?". Finally, participants responded to demographic questions.

7.2. Results

The reported analyses include only participants ($N = 474$) who correctly responded to the check questions. 55 were excluded because they failed at least one of the attention and manipulation check questions.

In the population condition, the mean response for the percentage of required happiness was 75.62 ($SD = 16.99$) and in the individual condition it was 74.98 ($SD = 16.84$). An independent t-test revealed that there were no significant differences between the two conditions, $t(468) = 0.41$, $p = .68$, $d = 0.04$ (Fig. 1). Across both conditions, the percentage for happiness was significantly above the midpoint of 50%, $t(473) = 32.58$, $p < .001$, $d = 1.50$.

For the following analyses we collapsed responses across the two conditions. The less willing participants were to relive the worst day of their lives in order to relive the best day of their lives, the higher their stated percentage of required happiness, $r(472) = -0.11$, $p = .02$. Women stated a higher percentage of required happiness ($M = 78.4$, $SD = 15.42$) than men ($M = 72.35$, $SD = 17.71$), $t(469) = 3.97$, $p < .001$, $d = 0.36$. Women were also less willing to relive the worst day of their lives in order to relive the best day of their lives ($M = 3.23$, $SD = 1.93$) than men ($M = 3.96$, $SD = 1.95$), $t(471) = 4.11$, $p < .001$, $d = 0.38$. Liberals stated a higher percentage of required happiness than conservatives, $r(472) = -0.17$, $p < .001$. There were no correlations between the percentage of required happiness and participants' stated level of current or general happiness. Each of the correlational analyses described above remained significant when controlling for condition (and there was no significant effect of condition for any of them).

7.3. Discussion

In Study 1a participants believed that a preponderance of extreme happiness (e.g. like the feeling of falling in love) over extreme unhappiness (e.g. like the feeling of being tortured) is required in order to make a population or an individual's life net positive. Participants' intuitions about the required proportion of happy and unhappy people in a given population were the same as their intuitions about the required proportion of happy and unhappy moments in people's lifetimes. This is in line with Starmans & Bloom, 2015 findings that showed that judgments about the distribution of happiness within a population closely match judgments about the distribution of happiness within a person's lifetime.

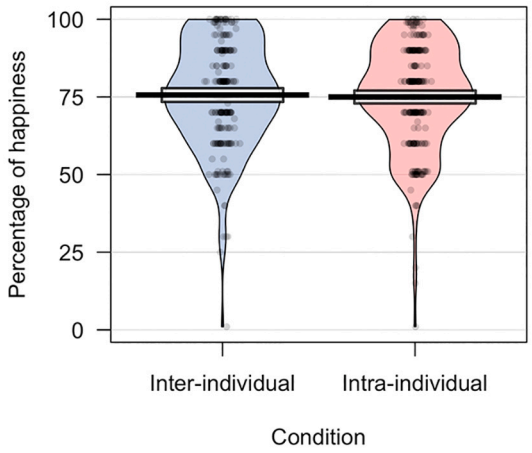

Fig. 1. On average, participants in Study 1a believed that ca. 75% of the people in a population must be happy (25% unhappy) for it still to be better to exist than not exist. Similarly, they believed that ca. 75% of people's lives must be composed of happy experiences (25% unhappy) for it still to be better for those people to exist than not exist.

Given the particular descriptions of the happiness and unhappiness experiences in question in this study, participants on average believed that roughly three times as much happiness is needed per a given amount of suffering for a population or a life to be overall positive. However, it is important to note that this number may depend on the context, the framing of the question, and the descriptions of happiness and suffering. In particular, it is not clear whether participants perceived the two types of experience as similar in magnitude.

As a result, the study leaves open the question of why participants believed more happiness was required to outweigh suffering. One possibility is that people weigh a given suffering unit more heavily than the equivalent unit of happiness when making their moral evaluations. According to this view, people are neither following strict classical utilitarianism (weighing happiness and suffering identically) nor are they following strict negative utilitarianism (only weighing suffering). Instead, their intuitions lie somewhere in between these views. That is, their intuitions take into account both happiness and suffering but they weigh suffering somewhat more than happiness.

An alternative explanation is that participants simply perceived the suffering dimension referenced in the instructions as more intense than the corresponding happiness dimension. According to this view, the asymmetry in the observed judgments is not a result of people's normative evaluation of two equivalent units of happiness and suffering. In the context of our studies, this may have resulted from our initial exampless of "extreme" happiness (falling love) and "extreme" suffering (torture) — perhaps participants simply felt that torture involves more by way of suffering than falling in love involves by way of happiness, which impacted the later judgments they provided. If this is true, then it remains possible that participants followed strict classical utilitarianism, but simply assumed in this case that suffering was more intense than happiness.

8. Study 1b: trading happiness against suffering with varying intensities

In Study 1a, the happiness and suffering were described as extreme. In Study 1b, we aimed to investigate whether people's intuitions about trading happiness against suffering are sensitive to the respective intensity levels of those experiences. We did this by manipulating whether the happiness and suffering amounts were either mild or extreme. We also introduced a symmetrical and linear happiness scale, ranging from −100 (extreme suffering), to −1 (mild suffering), to 0 (neutral), to +1 (mild happiness), to +100 (extreme happiness). We assumed that this could, at least partly, clarify that each happiness unit has an equivalent suffering unit.

We had two hypotheses. Our first hypothesis was that, overall, participants would believe that more happiness is needed to outweigh suffering, replicating Study 1. Our second hypothesis was that this asymmetry would be more pronounced when both happiness and suffering were extreme as opposed to both mild. We based this hypothesis on findings from previous research on the negativity effect, according to which the negativity of negative events increases at a faster rate compared to the positivity of positive events (e.g., Rozin & Royzman, 2001). Rozin and Royzman (2001) call this the principle of greater steepness of negative gradients. Note, however, that to our knowledge this effect hasn't been demonstrated yet in the context of evaluating the moral value of different outcomes. As a sanity check on the data, we also hypothesized that participants would believe that a greater percentage of happiness is required to outweigh suffering as the happiness became less intense and the suffering more intense. The study was pre-registered at https://aspredicted.org/b9wk4.pdf and had a 2 happiness (extreme vs. mild) x 2 suffering (extreme vs. mild) between-subjects design.

8.1. Method

8.1.1. Participants

We recruited 431 US American participants online via MTurk ($0.35 payment per participant). 75 were excluded, leaving a final sample of 356 people (170 female, M_{age} = 38.46, SD_{age} = 11.20). A priori power analysis showed that 351 participants were required to detect an effect size of f = 0.15, α of 0.05, power of 0.8, two-tailed, and four groups. The effect size was estimated based on previous pilot studies. We aimed to recruit at least 400 participants to account for any exclusions. Sample size was determined before data collection.

8.1.2. Procedure and materials

Participants were again presented with the same vignette as in the population condition of Study 1a. They were asked to envisage a world containing 1000 people, with these people comprising one of two types, happy or unhappy. In addition, they were presented with the following happiness scale: *"Let's assume a happiness scale ranging from -100 (extreme unhappiness) to 0 (neutral) to +100 (extreme happiness). Someone on level 0 is in a neutral state that feels neither good nor bad. Someone on level -1 experiences a very mild form of unhappiness, only slightly worse than being in a neutral state. Someone on level +1 experiences a very mild form of happiness, only slightly better than being in a neutral state. Someone on level -100 experiences the absolute worst form of suffering imaginable. Someone on level +100 experiences the absolute best form of bliss imaginable."* Depending on the condition, participants were told that the happy people were either extremely or mildly happy and that the unhappy people were either extremely or mildly unhappy. They were then asked the following question: "*Given this information, what percentage of extremely [mildly] happy people vs. extremely [mildly] unhappy people would there have to be for you to think that this world is overall positive rather than negative (i.e., so that it would be better for the world to exist rather than not exist)? In my view, the percentage of extremely [mildly] happy vs. extremely [mildly] unhappy people would need to be as follows: X% extremely [mildly] happy people; Y% extremely [mildly] unhappy people*".

8.2. Results

The reported analyses include only participants (N = 356) who correctly responded to the check questions. 75 were excluded because they failed at least one of the attention and manipulation check questions.

Overall, in accordance with our first hypothesis, participants believed that more happy than unhappy people were required to make it better for the world to exist than not to (see Fig. 2)—the required percentage of happy people was greater than 50% in all four conditions (extreme happiness vs. extreme suffering: $t(90) = 10.60$, $p < .001$, $d = 1.11$; mild happiness vs. extreme suffering: $t(85) = 14.41$, $p < .001$, $d = 1.55$; extreme happiness vs. mild suffering: $t(81) = 4.21$, $p < .001$, $d = 0.46$; mild happiness vs. mild suffering: $t(96) = 9.81$, $p < .001$, $d = 1.0$). A two-way ANOVA revealed two main effects but no significant interaction effect. That is, the preference for a preponderance of happy people was more pronounced both when suffering was extreme as compared to mild, $F(1, 352) = 41.41$, $p < .001$, $\eta_p^2 = 0.11$, as well as when happiness was mild as compared to extreme, $F(1, 352) = 6.54$, $p = .01$, $\eta_p^2 = 0.02$. The interaction between happiness and suffering levels was not significant, $F(1, 352) = 2.85$, $p = .09$, $\eta_p^2 = 0.008$.

Tukey HSD post-hoc tests revealed that participants believed that a greater percentage of happy people was needed to outweigh extreme suffering when the happiness experienced by those people was mild (M = 81.97, SD = 20.57) rather than extreme (M = 72.03, SD = 19.84), p = .006, d = 0.49. By contrast, the percentage of happy people required to outweigh mild suffering did not differ significantly as a function of whether that happiness was mild (M = 64.56, SD = 14.61) or extreme (M = 61.9, SD = 25.62), p = .82, d = 0.13. In accordance with our second hypothesis, participants believed that a greater percentage of happy people was needed to outweigh suffering when both happiness and suffering were extreme rather than mild. This difference was statistically significant with an independent t-test, $t(164) = 2.93$, p = .004, d = 0.43, but not with the Tukey HSD post-hoc test that adjusts for multiple comparisons, p = .06.

As in the previous study, women required a higher percentage of

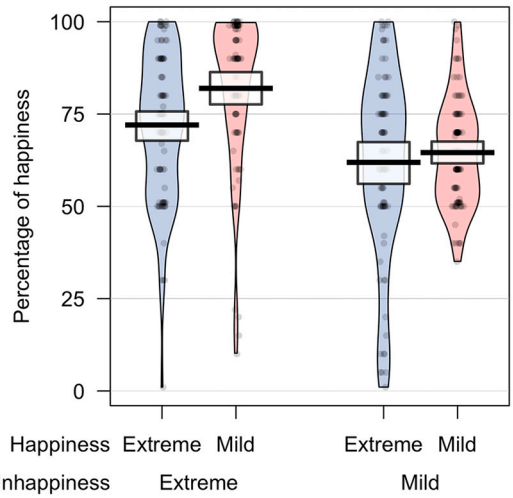

Fig. 2. Participants believed that more happy than unhappy people were required to make it better for the world to exist than not exist—the required percentage of happy people was greater than 50% in all four conditions in Study 1b. Participants were sensitive to the intensity levels of happiness and suffering. For example, they believed that a greater preponderance of mildly happy people were required to outweigh the complementary proportion of extremely unhappy people than to outweigh the complementary proportion of mildly unhappy people. Moreover, the observed asymmetry was more pronounced when both happiness and suffering were extreme as opposed to both being mild.

happy individuals ($M = 71.95$, $SD = 20.32$) than did men ($M = 68.34$, $SD = 22.67$) (pooled across all conditions). The difference, however, was not statistically significant, $t(354) = 1.58$, $p = .11$, $d = 0.17$. Neither was there a significant interaction effect between gender and experimental condition on the dependent variable. There were no further noteworthy correlations between the required percentages of happy individuals and demographic variables.

8.3. Discussion

The results of Study 1b replicate and extend our findings from Study 1a, chiefly that people require substantially more happy than unhappy people to exist in order to believe that a population is worth existing. They also confirm our hypothesis that people are sensitive to the intensity levels of happiness and suffering. Of most importance, and as we had hypothesized, the observed asymmetry was more pronounced when both happiness and suffering were extreme as opposed to both being mild. Participants believed that a greater preponderance of happy people is required to outweigh the complementary proportion of unhappy people when the happiness and suffering were both extreme compared to when they were both mild. This effect is in line with the principle of greater steepness of negative gradients, according to which the negativity of negative events increases at a faster rate compared to the positivity of positive events (e.g., Rozin & Royzman, 2001; compare Hurka, 2010).

One reason we included the explicit symmetrical happiness scale is because we believed this would make it more clear that a given happiness unit (e.g. +1) has an equivalent suffering unit (e.g. -1) with equally strong intensity. However, since the materials did not explicitly equalize the intensity levels of happiness and suffering, it remains unclear how participants interpreted the scale. It cannot be ruled out that they still perceived each suffering unit to be more intense than its numerically equivalent happiness unit, even when the two are presented on a common symmetrical linear scale. Therefore, although this study improves upon the method in Study 1, it cannot definitively rule out that the observed asymmetry in judgments is driven by an asymmetrical perception of the intensity of happiness and suffering and not by an asymmetrical normative evaluation of happiness and suffering (cf. Cacioppo & Berntson, 1994).

9. Study 1c: trading happiness against suffering with equal intensities

In the previous two studies we found that people believe that more happiness than suffering is needed to make it worthwhile for a population to exist. This could be because people believe that, given equal intensities, suffering is worse than happiness is good (i.e., asymmetric normative evaluation), or it could be because people generally perceive suffering to be more intense than happiness (i.e., asymmetrical perception). In Study 1c, we aimed to investigate whether people continue to believe that a preponderance of happiness is required even when it is explicitly stated that the two experiences are exactly equally intense. If this is the case, that would be evidence that the observed asymmetry is at least partly driven by an asymmetric normative evaluation of happiness and suffering.

Similar to Study 1a, the study had two conditions: inter-individual vs. intra-individual. Our first hypothesis, which was pre-registered at https://aspredicted.org/xa8ew.pdf, was that in both conditions, participants would believe that more than 50% of happiness is needed. Our second hypothesis was that, similar to Study 1a, there would be no significant difference between the inter-individual and intra-individual conditions.

9.1. Method

9.1.1. Participants

We recruited 283 US American participants online via MTurk ($0.35 payment per participant). 5 were excluded, leaving a final sample of 278 people (112 female, $M_{age} = 37.60$, $SD_{age} = 11.94$). An a priori power analysis showed that 128 participants were required to detect an effect size of $d = 0.25$, α of 0.05, power of 0.8, two-tailed, for a one-sample *t*-test. The effect size was estimated based on previous pilot studies. Since there were two conditions, we multiplied that number by two to obtain a minimum desired sample size of 256. We aimed to recruit at least 280 participants to account for any exclusions. Sample size was determined before data collection.

9.1.2. Procedure and materials

The materials in both conditions were similar to those in Study 1a with a few changes. The happiness and suffering levels were described as mild (as opposed to extreme in Study 1a). This is because we assumed that it is easier for participants to imagine a form of happiness that is equally intense as a form of suffering if both of these are described as mild than if both are described as extreme.

Participants were informed that the happy and unhappy experiences of the described people were exactly equally intense. Further, they were told that a typical person—including every person from this particular population in question—would weigh the two types of experiences equally strongly. For example, it was said that *"[i]f a person were to experience one hour of the positive experiences and then one hour of the negative experiences, they would consider this whole experience as exactly neutral, i.e., as neither good nor bad. If they were to experience the positive experiences very slightly longer than the negative experiences, they would consider this whole experience a positive one (and better than nothing). And if they were to experience the negative experiences very slightly longer than the positive experiences, they would consider this whole experience a negative one (and worse than nothing)."* Participants had to indicate whether they accepted these assumptions or not. Note that participants were asked whether they accept the provided information that people from this population perceive the positive experiences as being equal in intensity to the negative experiences; they were not at this point asked whether they themselves morally weigh a unit of suffering as much as a unit of happiness (since this is essentially what our dependent measure ascertains). As mentioned above, four participants rejected these assumptions and were therefore excluded from the analysis.

Next, as in Study 1a, participants were asked about the percentage of happy people (or time) they personally believe is required to make the population (or lives) net positive. After the main task, participants were asked whether, when answering the question, they had assumed that the described negative and positive experiences were equally intense or not. Participants were not excluded from the analysis based on their responses to this question, except in the case of particular follow-up analyses described below. Finally, participants responded to demographic questions.

9.2. Results

The reported analyses include only participants ($N = 278$) who correctly responded to the check questions. Four were excluded for not accepting the stated assumption that happiness and suffering were equally intense. One was excluded for not having a valid MTurk participant ID.

In the inter-individual condition, participants on average believed that 61.84% ($SD = 12.70$) of people would have to be happy in order to make the population overall net positive, which was significantly greater than 50%, $t(138) = 10.99$, $p < .001$, $d = 0.93$. In the intra-individual condition, participants on average believed that people would need to be happy 60.37% ($SD = 13.49$) of the time to make their lives overall net positive, again significantly greater than 50%, $t(138) =$

9.06, $p < .001$, $d = 0.77$. Thus, in both conditions the average was significantly above the midpoint of 50%. There was no significant difference between the two conditions, $t(275) = 0.93$, $p = .35$, $d = 0.11$. However, Fig. 3 shows that while the average response (both mean and median) was above the midpoint, the modal response in both conditions was 51.

83% of participants stated that when they answered the question, they assumed that the described negative and positive experiences were equally intense. 9% stated that they assumed that the positive experiences were more intense than the negative experiences and 8% assumed that the negative experiences were more intense than the positive experiences. We conducted the same analyses described above with only the subset of participants who stated that they assumed both experiences were equally intense and received the same results. These results are reported in the Supplementary Materials.

As in the previous studies, women required a higher percentage of happy individuals ($M = 62.35$, $SD = 12.9$) than did men ($M = 60.26$, $SD = 13.2$) (pooled across the two conditions). The difference, however, was not statistically significant, $t(242) = 1.31$, $p = .19$, $d = 0.16$. Neither was there a significant interaction effect between gender and condition on the dependent variable. There were no further noteworthy correlations between the required percentages of happy individuals and demographic variables.

9.3. Discussion

The results of Study 1c demonstrate that on average participants tended to evaluate suffering as worse than happiness is good, even when both are exactly equally intense. Participants were informed about and accepted the information that the happiness and suffering described were exactly equally intense. They also accepted the provided information that people from the population in question themselves perceive happiness and suffering as being equally strong, i.e., that they would consider one hour of happiness and one hour of suffering in combination as neutral overall. However, despite the fact that people accepted these assumptions, on average they continued to believe that ca. 1.5 times as much mild happiness than suffering is needed to make a given population or individual life net positive.

It is worth noting that it is not a logical contradiction for participants to believe that more happiness than suffering is needed to make a population or a life morally net positive, while at the same time believing that the very people in question weigh happiness and suffering equally. Similarly, there are many other cases where people's moral judgments about what is morally good deviates from what the affected people in question would want. For example, a strict negative utilitarian would consider a life filled with 95% happiness and 5% suffering as morally net negative, even if the very person in question would judge it as a life worth living. And a hedonic utilitarian would force someone against their own will into a hypothetical happiness experience machine (Nozick, 1974).

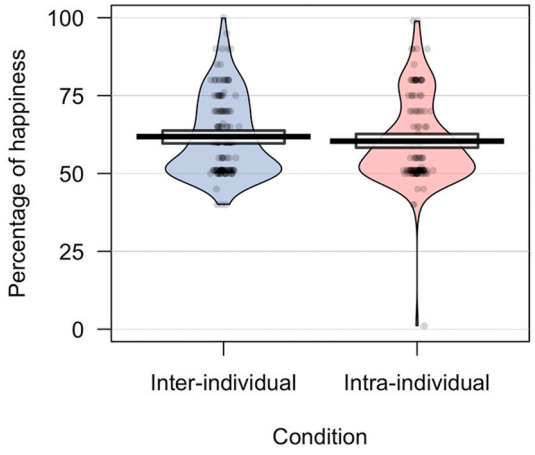

Fig. 3. Even when happiness and suffering were said to be equally intense, participants in Study 1c on average believed that more happiness than suffering is needed to make a given population or individual life net positive. A substantial proportion of participants, however, weighed happiness and suffering equally strongly, believing that only a very slightly greater amount of happiness was required to tip the balance towards net positivity.

Overall, these findings suggest that the asymmetry we found in the previous two studies is indeed, at least partly, driven by an asymmetric normative evaluation of happiness and suffering, and not purely by an asymmetric perception of the intensity levels of happiness and suffering. Yet, it should be noted that 37.0% of participants believed that only 50% or 51% of happiness is sufficient to outweigh the complementary amount of suffering. By contrast, in Study 1b only 19.6% believed that 50% or 51% of mild happiness is sufficient to outweigh mild suffering, with the large majority requiring a greater proportion of happiness. This may provide some evidence that the effect found in Studies 1a and 1b was partly driven by asymmetric perceptions of the intensity levels of happiness and suffering, with suffering being perceived as more intense than happiness—since when these intensity levels are more strictly equalized, less polarized distributions of happiness and suffering were judged net positive. Thus, it is possible that the average person's population ethical intuitions are driven both by asymmetric perceptions of the intensity level of happiness and suffering as well as by asymmetric normative evaluations of happiness and suffering.

10. Study 2a: adding a new happy or unhappy person

In the previous three studies, participants made judgments about the value of existing populations of a constant size. In Study 2a, we examined judgments about cases in which additional people could be introduced. In particular, we investigated how participants value the addition of a new individual who is either happy or unhappy and who otherwise would not exist. This allowed us to test whether people accept the *intuition of neutrality*, according to which the addition of a life worth living to the population is morally neutral, all else being equal, as well as a related view, called the *asymmetry*, according to which it is bad to create a new unhappy person but neutral to create a new happy person (McMahan, 1981).

Whether one endorses the intuition of neutrality can have dramatic implications for moral priority setting. For example, if one believes there is value in creating new happy people, it could be a priority to focus one's efforts on reducing the chances of human extinction and positively shaping the long term future of humanity due to the vast number of potential people who could exist in the future (Ord, 2020; Schubert, Caviola & Faber, 2019). Conversely, if one believes there is no value in creating new happy people, other priorities may follow, such as improving the lives of currently existing people or reducing the suffering of future generations. It has even been claimed that the asymmetry should lead us to favor human extinction (Beckstead, 2013; Holtug, 2004), given that it is virtually certain that many more people will be born with lives in which suffering predominates and that there is nothing to be said for bringing into existence lives in which happiness predominates, according to the asymmetry. However, this argument is disputed (see Frick, 2014; Nebel, 2019).

As noted, the intuition of neutrality is supposed to capture the thought that morality is about “making people happy” and not about “making happy people” (Narveson, 1973). To our knowledge, there exists no psychological research so far that directly examines whether people actually accept the intuition of neutrality. Do people indeed believe that there is no value in creating a new happy person who would not have been created otherwise? And do people—in line with the asymmetry—find a world with an additional unhappy person worse but a world with an additional happy person not better?

Based on our findings from Studies 1a-c, we hypothesized that people

would find a world with an additional unhappy person worse to a greater extent than they find a world with an additional happy person to be better. However, we were not sure whether people would fully endorse the asymmetry, such that they would consider a world with an additional unhappy person worse but a world with an additional happy person not better *at all*.

10.1. Method

10.1.1. Participants

We recruited 162 U.S. participants online via MTurk ($0.40 payment per participant). Five were excluded, leaving a final sample of 157 people (66 female, M_{age} = 38.86, SD_{age} = 11.29). We aimed to recruit 150 participants. The sample size was set in advance based on rough approximations of what would be needed to comfortably detect the smallest effect sizes of interest; but they were not based on precise power analyses. Note that this study was not pre-registered.

10.1.2. Procedure and materials

Participants were first presented with the happiness scale already used in Study 1b. Then, participants were asked two simple questions testing whether they accepted the assumptions of the scale — they were asked whether they accepted that a person at level − 100 experienced the "absolute worst form of suffering" and also whether a person at level + 100 experienced the "absolute best form of happiness."

Subsequently, in the first task, participants were asked to imagine a world with one million neutral people (on level 0). They were asked to imagine that a new person on level 0 could be added. They were told that this *"person's life would be exactly neutral with respect to the overall happiness and suffering they experience"*. Participants were asked: *"In terms of its overall value, how much better or worse would this world (containing this additional person) be compared to before?"* and responded on a 7-point scale (1 = *Much worse*, 4 = *Equally good*, 7 = *Much better*).

Next, in the main task, participants were again asked to imagine a world with one million neutral people into which a new person could be added. This new person was said to be either extremely happy (+100) or extremely unhappy (−100). The design was within-subjects such that each participant was presented with both questions (happy person and unhappy person) in randomized order on a separate page. The questions were as follows: "*One new person could be added to this world. This person would be extremely unhappy and live a life full of suffering and misery, on level -100 on the scale."* or *"One new person could be added to this world. This person would be extremely happy and live a life full of bliss and joy, on level +100 on the scale."* Again, they were asked how much better or worse the world would be containing the new person compared to the previous world. Finally, participants responded to demographic questions.

10.2. Results

The reported analyses include only participants (N = 157) who correctly responded to the check questions. Five were excluded because they failed at least one of these two questions.

One-sample *t*-tests against the midpoint 4 revealed that participants on average judged a world with an additional happy person to be better (M = 5.06, SD = 0.99), $t(156) = 13.44$, $p < .001$, $d = 1.07$, and a world with an additional unhappy person to be worse (M = 3.08, SD = 1.24), $t(156) = -9.39$, $p < .001$, $d = 0.75$ (Fig. 4). Next, we reversed the judgment scores in the unhappy condition, by subtracting them from 8 (M = 4.92, SD = 1.23) and compared them to the judgment scores in the happy condition. Surprisingly, a *t*-test revealed no significant differences in the strength of these two preferences, $t(156) = 1.23$, $p = .22$, $d = 0.13$. There were no significant order effects, depending on which question participants answered first.

Next, a one-sample t-test against the midpoint 4 revealed that participants on average judged it as an improvement to add one *neutral* person into the world, (M = 4.23, SD = 0.67), $t(156) = 4.40$, $p < .001$, $d = 0.35$. This suggests the existence of a weak general preference to create a new person, even if their happiness level is neutral.

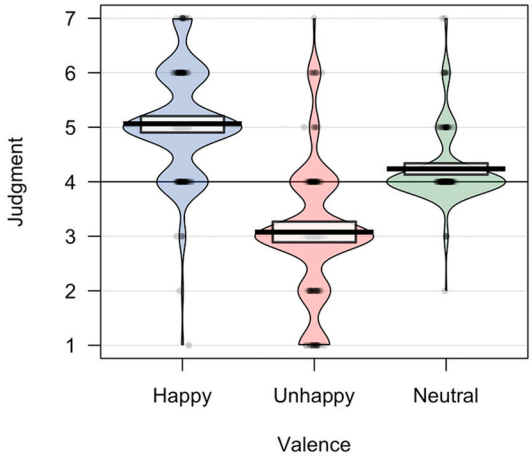

Fig. 4. Participants in Study 2a considered a world with an additional happy person to be better and a world with an additional unhappy person to be worse. These judgments were symmetrical, even after factoring out a weak general preference to add a new (neutral) person. 1 indicates 'Much worse', 4 indicates 'Equally good, 7 indicates 'Much better'.

One possibility is that people's judgments about adding a new happy or unhappy person are only symmetrical because they have a general preference to add new people. To test this, we conducted a repeated-measures ANOVA with ratings of the happy and unhappy person as the paired outcome variables (reverse scored in the unhappy condition) and with the rating of adding a neutral person as the covariate. The results revealed that there were no significant differences in the ratings even when the covariate was included in the analysis, $F(1, 156) = 1.52$, $p = .22$, $\eta_p^2 < 0.001$. This means that participants' judgments about the goodness of adding a new happy person and the badness of adding a new unhappy person still were not asymmetrical, even when the rating of adding the neutral person was statistically controlled for. There were no noteworthy associations between the dependent variables and demographic measures.

10.3. Discussion

The results show that participants did not consider the addition of new people morally neutral. Participants considered a world containing an additional happy person better and a world containing an additional unhappy person worse than a world with one million neutral people. In conflict with our prediction set out at the beginning of this study, participants did not find a world with an additional unhappy person to be more bad than they found a world with an additional happy person to be good. Instead, participants' judgments were symmetrical. Their judgments even remained symmetrical after factoring out their weak general preference for adding a new (neutral) person.

This suggests that people's axiological judgments about adding a new person are roughly in line with classical utilitarianism, according to which it is good to create a new happy person and (equally) bad to create a new unhappy person. At least when presented with the question we used in our study, participants' axiological judgments did not reflect the so-called asymmetry, according to which the creation of an unhappy person is bad but the creation of a happy person is only neutral.

This finding is surprising, given our findings from Studies 1a-c. In these studies, we found that participants weighed suffering more than happiness. We therefore predicted that they would also weigh the suffering of the newly added unhappy person more than the happiness of the newly added happy person. But that was not the case.

One possibility is that people have a particularly strong preference to add a new person to a world that only contains one million people, but that they might have a weaker preference to add a new person to a world that already contains a much larger population. The opposite is possible too: perhaps people's judgments become more asymmetrical if the pre-existing world contains no people at all. In Study 2b, we test this possibility.

11. Study 2b: adding a new person to an empty or full world

In Study 2a, we found that participants considered it good to add a new happy person to an existing world and bad to add a new unhappy person to an existing world. In contrast to our hypothesis, we found that their judgments were symmetrical. Study 2b had three aims.

First, we aimed to replicate the finding that people's moral judgments about adding a new happy or unhappy person are symmetrical in a between-subjects study design. By contrast, Study 2a featured a within-subjects study design. Another difference from Study 2a is that in Study 2b, participants were asked about the value of adding a new neutral person only after, and not before, they were asked about the value of adding a new happy or unhappy person.

Second, we explored whether participants are sensitive to the initial population size. It is possible, for example, that people have stronger preferences to add a new person to an empty world than to a world which is already filled with a lot of people (cf. *variablevaluetheories*, Hurka, 1983). It's also possible that people's judgments tend to be less symmetrical if the initial population size is very small or very big.

Third, we tested whether participants consider the questions nonsensical or feel that they don't know how to answer them. In all other studies reported in this paper, participants were forced to give a response to each question. However, it is possible that some participants may consider the questions too difficult or nonsensical and would therefore prefer to not give an answer. In this study, we tested this possibility.

11.1. Method

11.1.1. Participants

We recruited 402 US American participants online via Prolific ($0.40 payment per participant). Thirty-three were excluded, leaving a final sample of 369 people (193 female, M_{age} = 33.25, SD_{age} = 11.3). We aimed to recruit 400 participants. The sample size was set in advance based on rough approximations of what would be needed to comfortably detect the smallest effect sizes of interest; but they were not based on precise power analyses. Note that this study was not pre-registered.

11.1.2. Procedure and materials

The study had a 2 (world: empty vs full) x 2 (valence: happy vs unhappy) between-subjects design. As in Study 2a, participants were first presented with the happiness scale and they were asked the same assumption acceptance questions as in that study. Next, participants in the empty world condition were asked to imagine an empty world. Participants in the full world condition were asked to imagine a world filled with 10 billion neutral people (on level 0). By contrast, in Study 2a, participants were asked to imagine a world filled with 10 million neutral people. Depending on the condition, participants were then asked to imagine that a new happy or unhappy person was added into this world. Using the same question as in Study 2a, participants rated whether this change made the world better or worse than before. In contrast to Study 2a, participants had an additional option they could choose from, which was labelled as "This question doesn't make sense to me / I don't know". Next, participants were asked to imagine a new scenario, in which they considered the addition of a neutral person to either an empty world or a world filled with 10 billion neutral people (with the size of this world corresponding to the size of the world in the earlier main task). Finally, participants responded to an attention check and demographic questions.

11.2. Results

Twenty-six participants were excluded for rejecting the assumptions or for failing the attention check. Seven additional participants were excluded because they responded that the question regarding the addition of a happy, unhappy, or neutral person did not make sense to them or they did not know how to answer it. This means that over 98% of participants considered this type of population ethical question sensible. The reported analyses include only those participants who agreed with the assumptions, passed the attention checks, and considered the questions sensible (N = 369).

One-sample t-tests against the midpoint 4 revealed that participants on average judged a previously empty world with an additional happy person to be better (M = 5.64, SD = 1.16), $t(84) = 12.96$, $p < .001$, d = 1.41, and a previously empty world with an additional unhappy person to be worse (M = 2.38, SD = 1.07), $t(94) = -14.72$, $p < .001$, $d = 1.51$ (Fig. 5). Similarly, they judged a previously full world with an additional happy person to be better (M = 4.74, SD = 1.08), $t(89) = 6.56$, $p < .001$, $d = 0.69$, and a previously full world with an additional unhappy person to be worse (M = 3.14, SD = 0.99), $t(98) = -8.63$, $p < .001$, $d = 0.87$.

Next, we reversed the judgment scores in the unhappy conditions (by subtracting them from 8) and conducted a two-way ANOVA with world size and valence as the two factors. There was no main effect for the valence factor, suggesting that judgments were symmetrical, $F(1, 365) = 0.24$, $p = .63$, $\eta_p^2 < 0.001$. There was a main effect for the world size factor, $F(1, 365) = 54.16$, $p < .001$, $\eta_p^2 = 0.13$, indicating that participants considered it more good to add a happy person to an empty world than to a full world, and accordingly more bad to add an unhappy person to an empty world than to a full world. There was no interaction effect, $F(1, 365) = 0.33$, $p = .57$, $\eta_p^2 < 0.001$. As in Study 2a, even after factoring out participants' ratings of adding a new neutral person into the world there was no effect of valence, $F(1, 364) = 0.24$, $p = .62$, $\eta_p^2 < 0.001$.

Next, we conducted an analogous ANOVA with the ratings of the added neutral person as the outcome measure. Again, there was a main effect for the world factor, $F(1, 365) = 15.82$, $p < .001$, $\eta_p^2 = 0.04$, indicating that participants considered it better to add a new neutral person to an empty world (M = 4.25, SD = 0.65) than to a full world (M = 4.04, SD = 0.33). There was no main effect for the valence (happy vs unhappy) of the person considered in the preceding task, $F(1, 365) = 0.29$, $p = .59$, $\eta_p^2 < 0.001$, nor was there an interaction effect, $F(1, 365) = 0.16$, $p = .68$, $\eta_p^2 < 0.001$. There were no noteworthy correlations with demographic variables.

11.3. Discussion

The findings of Study 2b replicate our findings from Study 2a, using a between-subjects design and a systematic manipulation of the initial population size. Participants believed a world becomes better if a new happy person is added to this world and worse if a new unhappy person is added to this world. Again, these judgments were symmetrical, both when the initial population size was zero and when it was 10 billion. These results therefore contrast sharply with those of Studies 1a-1c, which showed that, for existing populations, people required a preponderance of happiness over suffering in order for the world to be net positive. What explains this discrepancy?

One possibility is that the nature of the question in Studies 2a-b, which involved adding new individuals, prompted different ethical considerations than the question about evaluating the acceptable happiness-to-suffering ratio within a given (or hypothetical) population. When people evaluate entire populations, as in Studies 1a-1c, they must globally assess the entire balance of happiness to suffering, and determine a minimum acceptable ratio of the two states. This question asks them to consider the final state of a given population. In contrast, when people evaluate the addition of new people, they are asked to evaluate a

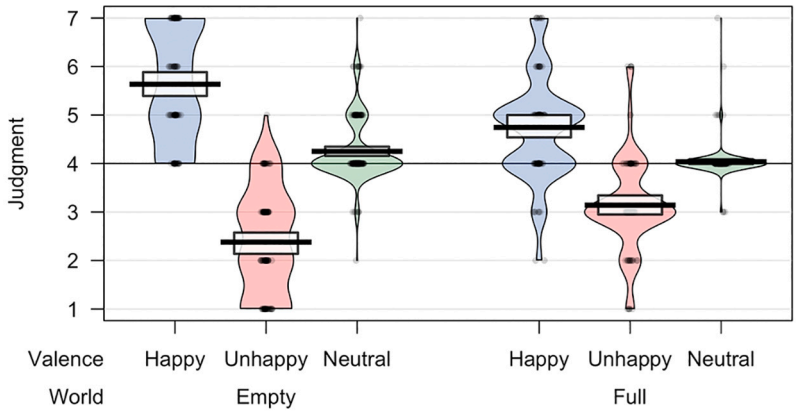

Fig. 5. Participants in Study 2b considered a world with an additional happy person to be better and a world with an additional unhappy person to be worse. These judgments were symmetrical, both when the pre-existing world contained no people (empty world) or when it contained 10 billion neutral people (full world). Ratings were more polarized for the empty than the full world. 1 indicates 'Much worse', 4 indicates 'Equally good, 7 indicates 'Much better'.

change to a population, not a final state. It might be that people simply think that, all else equal, it is a good thing to add a happy person, and an equivalently bad thing to add a person experiencing a commensurate level of unhappiness. Such a view is not inconsistent with also believing that considerably more happiness than suffering is required to make a population as a whole net positive. However, one important feature of our studies was that the to-be-added happy people experienced a happiness level that far exceeded the existing population's average (which was neutral); similarly, the to-be-added unhappy person experienced a level of unhappiness that was far worse than the existing population's average unhappiness (again, neutral). In both cases therefore, there was a large contrast between the new person's hedonic experience and the existing population's experience, which presumably made the assessment of the goodness or badness of adding the new person more straightforward. Accordingly, we cannot say on the basis of these results whether adding a new person whose happiness level matches or falls short of the existing population's average would produce similar results. Indeed, the question of how much people take into account a population's average level of happiness (or suffering) is one we turn to in the ensuing studies.

We found that participants had a stronger preference when the initial population size was zero than when it was 10 billion. That is, participants consider it better to add a new person to an empty world than a full world and worse to add an unhappy person to an empty world than a full world. One possible explanation is that this effect is driven by a preference for populations with better average (un)happiness levels — when an additional person is added to an empty world, the population average immediately becomes that person's happiness level, whereas when an additional person is added to a full world, the population's average shifts only minutely from the existing average (which in our studies, was neutrality). Thus, the question of whether people take into account the average happiness or unhappiness of a population appears here in another guise. Accordingly, in the studies that follow, we examine this paper's second research question of whether, in considering the overall value of a population, people focus on the population's average level of happiness (i.e., averagism), or its total amount of happiness (i.e., totalism), or some blend of the two (cf. *variable valuetheories*, Hurka, 1983; Ng, 1989).

12. Study 3a: populations with fixed average levels and varying total levels of happiness

In Study 3a, we tested whether people strictly follow averagism by varying the size of populations that consist of either happy or unhappy people. If people strictly follow averagism, they should be indifferent between choosing a small population with just 1000 people or a large population with 1 billion people, as long as the average happiness or unhappiness levels across each population are equal. By contrast, if people at least partially follow totalism, they should have a preference for larger over smaller happy populations and for smaller over larger unhappy populations.

Our first hypothesis, which we pre-registered at https://aspredicted.org/e2jp6.pdf, was that people would prefer smaller over larger unhappy populations and larger over smaller happy populations. Thus, we hypothesized that people would not strictly follow averagism but would at least partly follow totalism. Our second hypothesis was that their preference for smaller over larger unhappy populations would be stronger than their preference for larger over smaller happy populations. This second prediction reflects the tendency found in our prior studies for people to weigh unhappiness more strongly than happiness, which in this context, should amplify the subjective importance of differences in unhappiness as compared with equivalent differences in happiness.

12.1. Method

12.1.1. Participants

We recruited 868 US American participants online via MTurk ($0.45 payment per participant). 79 were excluded, leaving a final sample of 789 people (358 female, $M_{age} = 40.11$, $SD_{age} = 12.05$). A priori power analysis showed that 788 participants were required to detect an effect size of $d = 0.2$, α of 0.05, power of 0.8, two-tailed. The effect size was estimated based on previous pilot studies. We aimed to recruit at least 850 participants to account for any exclusions. Sample size was determined before data collection.

12.1.2. Procedure and materials

Participants were randomly assigned to one of two conditions: happy populations and unhappy populations. First, they were informed that they would be presented with three different scenarios in which they would be asked to consider two civilizations that last for millions of years and that differ only in their size. Depending on the condition, participants were told that all inhabitants of these civilizations are happy or unhappy. For reasons of simplicity, all inhabitants of a civilization were said to have the same happiness level. It was made clear that these civilizations would exist for millions of years and that their population size would remain constant. Further, participants were informed that these civilizations would have no issues with resource depletion, environmental degradation or overpopulation and that they would have multiple Earth-like planets available to them. At this point, participants had to indicate whether they accepted the provided information or not. It was not specified where these civilizations would exist and whether there would be other human populations in this world.

Next, participants responded to three questions presented on separate pages which presented three different comparisons of population sizes, in randomized order (1000 vs. 10,000; 1 million vs. 10 million; 1 billion vs. 10 billion). Note that the ratio of the small to the large population was constant, i.e. a 10-fold increase. The 1 billion vs. 10 billion scenario of the happy condition, for example, read as follows: "*In the first civilization, there are one billion (1B) people living at any one time. In the second civilization, there are ten billion (10B) people living at any one time. The inhabitants of both civilizations are equally happy and lead lives filled with bliss and joy. Every single person's life is well worth living. If only one civilization could come into existence, which would you prefer?*" Participants responded on a 7-point scale (1 = *Strongly prefer 1B civilization*, 4 = *No preference*, 7 = *Strongly prefer 10B civilization*).

After the main task, participants were asked to state in an open text field what the ideal population size for such a civilization would be. They then indicated whether they assumed that the civilizations had no issues with resource depletion, environmental degradation or overpopulation, and whether they assumed that the two civilizations were identical apart from their size. Next, they responded to an attention check question asking how the inhabitants of the civilizations were described, for which the correct answer was either happy or unhappy.

Next, participants were presented with a short task that we included in order to statistically control for any general preferences participants may have had regarding populations of particular sizes. Participants were informed that they would be presented with another set of three scenarios similar to the main task, with the only difference being that rather than the inhabitants being either happy or unhappy, their lives are instead exactly neutral with respect to the overall happiness they experience. Apart from this aspect, everything else was the same as in the main task and participants responded to the same three scenarios in randomized order.

Finally, participants responded to demographic questions, including a question about their religiosity and belief in the afterlife.

12.2. Results

The reported analyses include only participants ($N = 789$) who correctly responded to the check questions. 11 failed to accept the information provided at the beginning of the study, 38 stated after the task that they did not assume that there were no issues regarding resource depletion, environmental degradation, and overpopulation, 23 stated after the task that they did not assume that the two populations had the same happiness or unhappiness levels, and 35 failed a simple attention check.

One sample *t*-tests against the mid-point (4) revealed that participants preferred larger over smaller happy populations ($M = 4.39$, $SD = 1.65$, aggregated, $\alpha = 0.82$), $t(401) = 4.69$, $p < .001$, $d = 0.23$, and smaller over larger unhappy populations ($M = 5.36$, $SD = 1.50$, aggregated and reverse coded, $\alpha = 0.84$), $t(386) = 17.78$, $p < .001$, $d = 0.91$. As predicted, participants' preference for larger over smaller happy populations was weaker than their preference for smaller over larger unhappy populations, $t(785) = 8.67$, $p < .001$, $d = 0.62$.

Paired sample *t*-tests showed that as population size increased, participants' preferences for larger over smaller happy populations decreased (thousands vs millions: $t(401) = 10.63$, $p < .001$, $d = 0.42$; millions vs billions: $t(401) = 8.24$, $p < .001$, $d = 0.31$; Table 1, Fig. 6). By contrast, as population size increased, participants' preferences for smaller over larger unhappy populations persisted (thousands vs millions: $t(386) = 6.51$, $p < .001$, $d = 0.26$; millions vs billions: $t(386) = 4.09$, $p < .001$, $d = 0.15$). Preferences for larger over smaller happy populations decreased with population size more than preferences for smaller over larger unhappy populations increased. We tested this by comparing the absolute difference between the two extreme dilemmas (1000 vs 10,000 and 1 billion v 10 billion) across the two conditions, $t(773) = 5.05$, $p < .001$, $d = 0.36$. It is noteworthy that participants' preference for larger over smaller happy populations decreased so much with larger population sizes that their preference reversed once it reached a billion: a one sample t-test revealed that participants preferred a population of one billion over ten billion happy people, $t(401) = -2.90$, $p = .004$, $d = 0.14$.

On average, there was no clear preference for or against larger over smaller neutral populations ($M = 4.04$, $SD = 1.57$; aggregated, $\alpha = 0.83$), $t(788) = 0.63$, $p = .53$, $d = 0.02$. However, *t*-tests showed that as population size increased, preferences for larger over smaller neutral populations decreased(thousand vs million: $t(788) = 12.64$, $p < .001$, $d = 0.37$; million vs billion: $t(788) = 10.65$, $p < .001$, $d = 0.27$; Table 1). The tendency in the aggregate to prioritize the larger over the smaller neutral population correlated positively with the tendency in the aggregate to prioritize the larger over the smaller happy population, $r(400) = 0.70$, $p < .001$, and negatively with the aggregate tendency to prioritize the smaller over the larger unhappy population, $r(385) = -0.29$, $p < .001$. Using linear regression, we found that participants' aggregate preference for larger over smaller happy populations was still weaker than their aggregate preference for smaller over larger unhappy populations even after controlling for their aggregate preference (or lack thereof) for larger over smaller neutral populations, $t(786) = 8.24$, $p < .001$, $d = 0.57$, $b = 0.27$.

The median response for the ideal population size was 1.5 million for the happy civilization and 100 for the unhappy civilization. There were no correlations between these responses and demographic variables, including religiosity and belief in an afterlife.

Table 1
Mean ratings in Study 3a for happy, unhappy, and neutral population comparisons of different population sizes. 1 indicates a strong preference for the smaller population, 4 indicates no preference, 7 indicates a strong preference for the larger population. Note that these are raw values (not reversed).

	1000 vs 10,000	1 million vs 10 million	1 billion v 10 billion
Happy	5.12 (1.84)	4.33 (1.93)	3.71 (1.99)
Unhappy	3.04 (1.99)	2.56 (1.66)	2.32 (1.49)
Neutral	4.65 (1.82)	3.98 (1.84)	3.48 (1.81)

12.3. Discussion

The results of Study 3a confirm our hypothesis that participants would not strictly follow averagism. If they had done so, they would have been indifferent between each of the different population comparisons presented in this study. Instead, participants at least partly followed totalism. They had a preference for larger over smaller happy populations and for smaller over larger unhappy populations.

We also found that participants' preference for smaller over larger unhappy populations was stronger than their preference for larger over smaller happy populations. This effect persisted even when controlling for a potential preference for larger or smaller neutral populations. Thus, people show "asymmetric scope sensitivity" with respect to happy and unhappy population sizes. And this asymmetric scope sensitivity was more pronounced the larger the population sizes got.

It is possible that this asymmetric scope sensitivity can be explained by the fact that, normatively speaking, people consider suffering more bad than they consider (equivalently intense) happiness to be good, and they are therefore particularly focused on minimizing the extent of suffering rather than maximizing the extent of happiness. An alternative hypothesis is that people consider the described suffering as more intense than the described happiness. Study 3a does not allow us to distinguish between these two hypotheses. In Study 1c we found that the average participant's population ethical intuitions remained asymmetrical even when they believed happiness and suffering to be equally intense. We therefore believe it is plausible that even when participants believe happiness and suffering to be equally intense, they would still demonstrate a similar asymmetric scope sensitivity effect.

Participants did not consistently prefer the happy populations to be

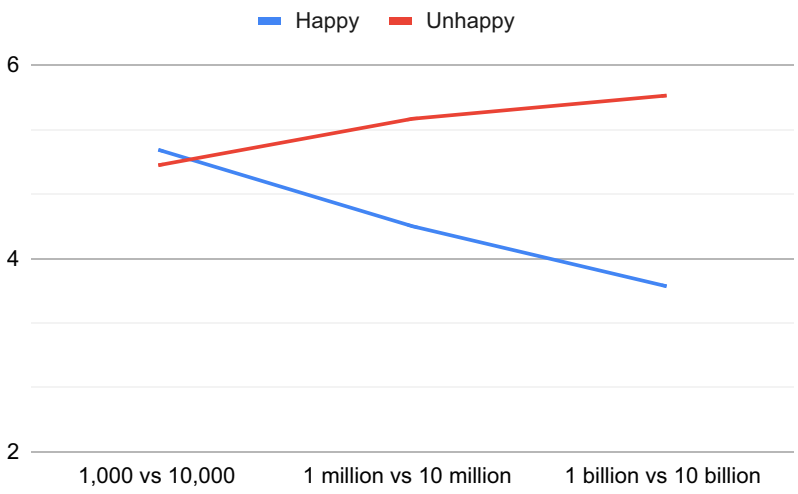

Fig. 6. Absolute preference strength with increasing population sizes in Study 3a. Note that the values in the unhappy condition were reverse coded to better demonstrate the asymmetric trend across the two conditions. Preferences for larger over smaller happy populations and preferences for smaller over larger unhappy populations were similarly strong when the comparison was between populations of 1000 and 10,000 respectively. However, the greater the populations became (e.g. 1 million vs. 10 million or 1 billion vs. 10 billion), the stronger the preference for smaller over larger unhappy populations became and the weaker the preference for larger over smaller happy populations became.

as large as possible. Instead, in our main task we found that participants preferred the happy civilization to be smaller than 10 billion, but larger than 1 million. And when directly asked to state their ideal population size for a happy population, participants' median response was 1 million. When interpreting these results, it is important to note that participants did not necessarily assume that these populations would encompass all of humanity. Our instructions left open the possibility that there could be other human populations in the world.

It is not entirely clear why participants preferred the happy civilizations to be smaller than 1 billion. A similar pattern was found in the neutral happiness task. Even though on average participants had no statistically significant preference for larger or smaller neutral civilizations, descriptively we found that they tended to prefer neutral civilizations with a population size between 1 million and 10 million. One interpretation is that people have a general preference for civilizations with a specific population size that they deem neither too large nor too small. Perhaps people are concerned that if a civilization contains too many people this could lead to negative consequences, such as overcrowdedness. While we tried to rule out such concerns by making clear that the civilizations would have multiple Earth-like planets available and that they would have no issues with overpopulation, and even excluded participants who did not accept these assumptions, it is still possible that such concerns were partly driving participants' intuitions. Several participants mentioned in their comments that they were concerned that overly large civilizations could cause overcrowdedness.

Another surprising finding was that when directly asked to state their ideal population size for an unhappy population, participants' median response was 100. We do not know why participants did not believe that the ideal population size was 0. Their response appears also inconsistent with our findings from Studies 2a-b in which participants considered it wrong to create new unhappy people. One possibility is that participants were driven by concerns for certain non-welfarist goods, such as an intrinsic preference for the continued existence of humanity.

13. Study 3b: populations with fixed total levels and varying average levels of happiness

In Study 3a, we found that people have totalist intuitions in cases where populations differ in their total levels but have constant average levels. However, this finding does not rule out that people also have averagist intuitions in addition. In Study 3b, we aimed to investigate whether people have a preference for populations with greater average levels of happiness or smaller average levels of suffering if the total amount of happiness or suffering is the same across populations.

Our first hypothesis, which we pre-registered at https://aspredicted.org/3we4j.pdf, was that people have a preference for populations with a higher average level of happiness and a preference for populations with a lower average level of suffering when the total amount of happiness or suffering between contrasting populations is held constant. In Study 3a, we found that people's totalist preference was stronger for suffering than for happiness. Based on this, our second hypothesis was that people's averagist preference would also be stronger for suffering than for happiness.

13.1. Method

13.1.1. Participants

We recruited 866 US American participants online via MTurk ($0.45 payment per participant). 41 were excluded, leaving a final sample of 825 people (358 female, $M_{age} = 39.54$, $SD_{age} = 12.35$). A priori power analysis showed that 788 participants were required to detect an effect size of $d = 0.2$, α of 0.05, power of 0.8, two-tailed. The effect size was estimated based on previous pilot studies. We aimed to recruit at least 850 participants to account for any exclusions. Sample size was determined before data collection.

13.1.2. Procedure and materials

Participants were randomly assigned to one of two conditions: happy or unhappy populations. The instructions and the task were similar to those used in Study 3a. Participants were informed that they would be presented with three different scenarios in which they would be asked to consider two civilizations and judge which one they would prefer to come into existence. They were informed that these civilizations will have no issues with resource depletion, environmental degradation, crowdedness, or overpopulation and that they will have multiple Earth-like planets available to them. Similar to Study 1b, participants were presented with a happiness scale, ranging from −100 to +100.

Participants were presented with three dilemmas in randomized order in which two possible civilizations were pitted against each other. Depending on the condition, the inhabitants of both civilizations were either happy or unhappy. The civilizations differed in size and average happiness level but were equated by design in terms of their total happiness levels. For example: “Civilization A contains 4,000 people at +60 happiness [-60 unhappiness]. Civilization B contains 6,000 people at +40 happiness [-40 unhappiness]”. In the other two dilemmas the population sizes were 4 million/6 million or 4 billion/6 billion respectively. Participants responded on a 7-point scale (1 = *Strongly prefer civilization A*, 4 = *No preference*, 7 = *Strongly prefer civilization B*).

After the main task, we checked whether participants understood how to evaluate populations based on the average and total principles. We presented them once again with one dilemma from the main task (4000 vs 6000 people) and asked them in which civilization, (1) the

average level, and (2) the total amount of happiness (or suffering) is greater, smaller or the same. On the next page, we asked them to explicitly calculate the average and total amounts. For example, we asked: "What is the average level of happiness [unhappiness] in civilization A?" or "What is the total amount of happiness [unhappiness] in civilization A (i.e. the sum of happiness [unhappiness] across all people)?". In the happy population condition, the correct average level of happiness of civilization A (4,000 people at level +60) would be (4000 * 60) / 4000 = 60, which is higher than for civilization B (6,000 people at level +40). The correct total amount of happiness of civilization A would be 4000 * 60 = 240,000, which is the same as for civilization B (6,000 * 40 = 240,000). We expected that a substantial proportion of participants would fail to calculate these numbers correctly and therefore did not pre-register exclusions based on these responses. Finally, participants responded to demographic questions.

13.2. Results

The reported analyses include only participants (N = 825) who correctly responded to the check questions. 41 were excluded either for failing the attention check or for failing to accept the provided information at the beginning of the study.

One sample t-tests against the mid-point revealed that, in the happy population condition, participants preferred (smaller) populations with greater average happiness levels over (larger) populations with lower average happiness levels, notwithstanding that their total happiness levels were equivalent ($M = 2.28$, $SD = 1.42$, aggregated across the three different population size dilemmas in the happy condition, $\alpha = 0.92$), $t(426) = -25.10$, $p < .001$, $d = 1.21$. Similarly, in the unhappy population condition, participants also preferred (larger) populations with lower average suffering levels over (smaller) populations with greater average suffering levels but the same total suffering ($M = 2.34$, $SD = 1.10$, reverse scored and aggregated across the three different population size dilemmas in the unhappy condition, $\alpha = 0.88$), $t(397) = -30.05$, $p < .001$, $d = 2.34$. There was no difference in the strength of this preference between the happy and unhappy conditions, $t(798) = -0.65$, $p = .52$, $d = -0.05$ (Fig. 7). There were also no noteworthy differences between the three types of dilemmas (which varied by population size) within each condition (Table 2). (See Table 3.)

Next, we looked at the follow-up questions. When asked directly, 80% of participants correctly indicated which civilization had a greater average level of (un)happiness (principle check question). 86% (happiness condition) and 88% (unhappiness condition) of participants correctly calculated the average level for civilization A (±60), $\chi^2(1) = 0.71$, $p = .40$. And 85% (happiness condition) and 88% (unhappiness condition) of participants correctly calculated the average level for civilization B (±40), $\chi^2(1) = 0.85$, $p = .36$. When asked directly, only 47% of participants correctly indicated that both civilizations had the same total level of (un)happiness (principle check question). Only 35% (happiness condition) and 26% (unhappiness condition) of participants correctly calculated the total average level for civilization A (±240,000), $\chi^2(1) = 7.86$, $p = .005$. And only 35% (happiness condition) and 25% (unhappiness condition) of participants correctly calculated the total average level for civilization B (±240,000), $\chi^2(1) = 9.29$, $p = .002$. In both conditions, 35–37% of participants incorrectly calculated that the total levels were ±60 (civilization A) and ±40 (civilization B) respectively, suggesting that they misunderstood the question and did not multiply the stated individual happiness level by the number of people. When all participants who responded incorrectly to at least one of the two principle check questions or one of the four calculation check questions were excluded, the pattern of the results remained the same (see Supplementary Materials). There were no noteworthy correlations between the dependent variables and demographic variables.

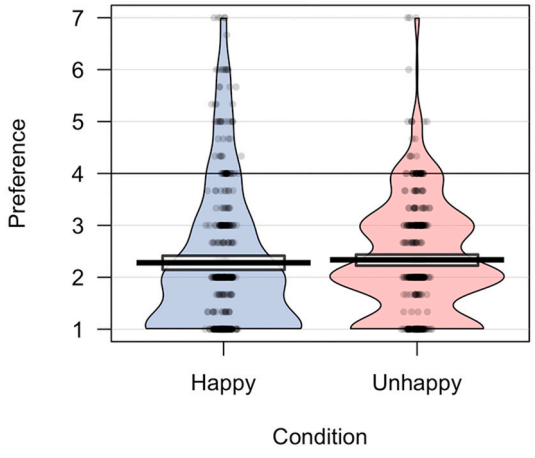

Fig. 7. Participants in Study 3b had a preference for populations with larger average happiness levels (left plot) and smaller average suffering levels (right plot) even though the total amounts of happiness or suffering between the two populations were the same. 1 indicates a preference for the smaller population with higher happiness level (left plot) or the larger population with smaller suffering level (right plot), 4 indicates no preference, 7 indicates the opposite preference.

Table 2
Mean ratings in Study 3b for comparisons between populations of different sizes but equal total amounts of happiness or suffering, i.e., the smaller populations have an average happiness (unhappiness) level of +60 (−60) and the larger populations have an average happiness (unhappiness) level of +40 (−40). 1 indicates a preference for the smaller population with higher happiness level (Happy condition) or the larger population with smaller suffering level (Unhappy condition), 4 indicates no preference, 7 indicates the opposite preference.

	4,000 vs 6,000	4 million vs 6 million	4 billion vs 6 billion
Happy	2.35 (1.55)	2.23 (1.48)	2.25 (1.55)
Unhappy	2.32 (1.24)	2.32 (1.19)	2.37 (1.24)

Table 3
Mean ratings in Study 3c for comparisons between populations of different sizes and different average levels of happiness or suffering. 1 stands for a preference in line with averagism (populations with better average but worse total levels), 4 stands for no preference, 7 stands for a preference in line with totalism (populations with better total but worse average levels). Responses were reverse scored in the unhappiness condition.

	Happiness		Unhappiness	
	Intuition	Reflection	Intuition	Reflection
100,000@90	4.79 (2.13)	5.04 (2.08)	4.34 (2.08)	4.76 (1.99)
100,000@70	3.51 (2.08)	4.21 (2.17)	3.55 (1.96)	4.06 (2.11)
100,000@50	2.71 (1.76)	3.28 (2.08)	3.12 (1.92)	3.19 (2.08)
million@50	2.87 (1.82)	3.49 (2.11)	3.44 (2.05)	3.77 (2.23)
billion@50	2.91 (1.93)	3.55 (2.15)	3.80 (2.18)	4.22 (2.31)

13.3. Discussion

The results of Study 3b show that participants had averagist preferences when the total amount of happiness or suffering was held constant across populations. In Study 3a, we found that participants had broadly totalist preferences when the average amount of happiness or suffering was held constant across populations. Thus, together these two studies suggest that people seem to have both averagist and totalist preferences to some degree.

It was surprising to see that so many participants failed to correctly infer that both populations had the same total amount of happiness. One possibility is that people did not understand how to answer the question correctly. For example, they might not have understood that one can calculate the total level of happiness simply by multiplying the level of a

single individual by the number of people. This calculation is highly abstract and unfamiliar to most people. Furthermore, the units themselves are very abstract and possibly meaningless to many people (i.e., what does it really mean to say that a population has a happiness "level" of 240,000?). Another possibility is that at least some participants rejected the idea that one could answer the question via an algebraic calculation. That is, they might not think that happiness can be quantified and multiplied in this manner. While this sort of philosophical objection could exist, we suspect that it is not very common among lay people. Nevertheless, when we conducted the same analysis including only those participants who answered these questions correctly we found the same pattern of results. This suggests that people have averagist preferences in cases where the averages differ and the totals are constant regardless of whether they correctly calculated the total amounts or not.

14. Study 3c: populations with both varying total levels and average levels of happiness

In Study 3c, we looked at cases in which populations differ both in their total and average levels. The central question was whether people would prefer smaller populations with higher average but lower total happiness levels (averagism) over larger populations with lower average but higher total happiness levels (totalism). (And vice versa for cases involving suffering instead of happiness.) The study had two main objectives. The first objective was to investigate whether people have a mix of both averagist and totalist tendencies in cases where populations differ in both their average and total levels. Our hypothesis was that in such cases, people would have intuitions in line with both principles which could counteract each other. We, therefore, assumed that, across various dilemmas, people would make choices that are neither strongly in line with averagism nor with totalism, but rather that their choices would lie in between the recommendations of these two principles.

As a further subquestion we aimed to determine the point at which (roughly) people cross the "threshold" between favoring the higher average happiness or favoring the higher total happiness. In other words, how much higher must the total happiness of one population be in order to outweigh its reduced average happiness, and vice versa? We did not have a specific hypothesis regarding this question.

Our second objective was to investigate whether people's population ethical intuitions are affected by their thinking style—that is, whether they think intuitively or reflectively. One possibility is that focusing on the average level is intuitively predominant because in certain contexts it is cognitively less taxing. It only requires the single cognitive step of comparing two readily available numbers (at least in our study) and requires no attention to the respective population sizes. Focusing on the total level, by contrast, requires the additional cognitive step of multiplying each population's average level by its size, and then comparing these two numbers with each other. Therefore, we hypothesized that when people think intuitively, their preferences become more in line with averagism, whereas when they think more reflectively, their preferences become more in line with totalism.

Finally, as a confirmation of the effects we found in Studies 3a and 3c, we hypothesized that people's tendency to prefer one population over another would get weaker if both the average and total levels of that population worsen.

Our study was pre-registered at https://aspredicted.org/zw6ek.pdf and had a 2 (valence: happiness vs unhappiness) x 2 (thinking style: intuition vs reflection) between-subjects study design.

14.1. Method

14.1.1. Participants

We recruited 622 US American participants online via MTurk ($0.45 payment per participant). 161 were excluded, leaving a final sample of 461 people (220 female, M_{age} = 39.35, SD_{age} = 11.79). We aimed to recruit 600 participants. The sample size was set in advance based on rough approximations of what would be needed to comfortably detect the smallest effect sizes of interest; but they were not based on precise power analyses.

14.1.2. Procedure and materials

The instructions and the task were similar to those in Study 3b. Participants were informed that they would be presented with five different scenarios in which they are asked to consider two civilizations and judge which one they would find better. Again, they were informed that these civilizations will have no issues with resource depletion, environmental degradation or overpopulation and that they will have multiple Earth-like planets available to them. Participants were again asked whether they accept these assumptions. Similar to Study 1b, participants were presented with a happiness scale, ranging from −100 to +100.

On the next page, participants in the intuition conditions read: "*When you answer the next three questions, please try to respond quickly. Don't think too much about the answer. Just follow your first gut reaction and intuition.*" Participants in the reflection conditions read: "*When you answer the next three questions, please try to think long and hard about the answer. Make sure not to follow your first gut reaction blindly. Instead, try to reflect more deliberately about your answer. With further reflection, you might agree with your initial gut reaction, but you might also disagree with it. Try to rely only on your considered reflection in answering the following questions.*"

Next, participants were presented with five dilemmas in which two possible civilizations (Civilization A and B) were pitted against each other. Depending on the condition, the inhabitants of both civilizations were all either happy or unhappy. In all five dilemmas, Civilization A contained 1000 people on level ±100 (referred to as 1000@100), whereas the size and average level of Civilization B varied.

In the first three dilemmas, the three comparison civilizations differed from Civilization A in both their total and average happiness levels, but they did so to differing extents. Civilization B contained either 1) 100,000 people on level ±90 (referred to as 100,000@90), 2) 100,000 people on level ±70 (referred to as 100,000@70), 3) 100,000 people on level ±50 (referred to as 100,000@50).

In the last two dilemmas, the two civilizations compared against Civilization A had the same average happiness levels as the civilization in the third dilemma (100,000@50), while differing in their total happiness levels. Civilization B contained either 4) 1 million people on level ±50 (referred to as million@50), 5) 1 billion people on level ±50 (referred to as billion@50).

For example, the 100,000@90 dilemma in the happiness condition read as follows: "Civilization A contains 1,000 happy people on level +100. Civilization B contains 100,000 happy people on level +90. If only one civilization could come into existence, which one would be better?" Participants responded on a 7-point scale (1 = *Definitely civilization A*, 4 = *Equally good*, 7 = *Definitely civilization B*).

After the main task, participants had to complete a comprehension check question that asked how someone on level 0 would feel according to the happiness scale. The correct answer was "neither good nor bad", which is the exact wording used when the happiness scale was presented at the beginning of the study. Finally, participants responded to demographic questions.

14.2. Results

The reported analyses include only participants (N = 461) who correctly responded to the check questions. 159 were excluded because they failed the comprehension check that asked how someone on level 0 on the previously described happiness scale would feel (correct answer: neither good nor bad). Of those who answered incorrectly, 96 believed that someone on level 0 would feel extremely unhappy, 46 believed that someone on level 0 would feel extremely happy, and 17

believed that someone on level 0 would feel mildly happy. And 2 additional participants were excluded because they did not accept the assumptions stated at the beginning of the study (regarding resource depletion, etc.). We report the same analyses with the full sample without any exclusions in the Supplementary Materials. The pattern of the results is the same.

In line with our first hypothesis, participants' responses aggregated across all five dilemmas and across thinking style were neither strongly in line with averagism nor with totalism but instead they lay in between the recommendations of these two principles (Fig. 8, Table 4). Note that responses in the unhappiness conditions were reverse coded for all reports and analyses of this study. A one-sample *t*-test against the midpoint (4) revealed that in the happiness conditions, participants had a weak aggregated preference (aggregated, $\alpha = 0.92$) for the smaller populations with higher average happiness but lower total happiness over the larger populations with higher total happiness but lower average happiness ($M = 3.62$, $SD = 1.78$), $t(212) = -3.10$, $p = .002$, $d = 0.21$. In the unhappiness condition, by contrast, participants did not have an aggregated preference (aggregated, $\alpha = 0.90$) for either of the two populations ($M = 3.82$, $SD = 1.76$), $t(247) = -1.62$, $p = .12$, $d = 0.10$.

So far, we have only looked at the aggregated responses across dilemmas. Next, we tested for each individual dilemma (across both thinking style conditions) whether responses were significantly above or below the midpoint (4). This allows us to infer, for each dilemma, whether participants' responses were more in line with averagism, totalism, or whether they lay in between the recommendations of the two principles. The results in Table 4 — for both happy and unhappy populations – show that participants' responses tended to be more in line with totalism in the 100,000@90 dilemma, roughly at the midpoint in the 100,000@70 dilemma, and more in line with averagism in the 100,000@50 dilemma. In the two last dilemmas, million@50 and billion@50, participants' responses tended to be more in line with averagism. The only exception were responses in the billion@50 dilemma in the unhappiness condition, which were roughly at the midpoint.

In the first three dilemmas, predictably, participants were more inclined to favor the population with the better average level (as opposed to the population with the better total level) as the average (and therefore also total) level of the larger population declined (100,000@90 and 100,000@70: $t(460) = 12.04$, $p < .001$, $d = 0.43$; 100,000@70 and 100,000@50: $t(460) = 9.89$, $p < .001$, $d = 0.37$). This was the case in both the happiness and unhappiness conditions. In the last three dilemmas, participants largely continued to favor the population with the better average level (as opposed to the population with the better total level), even when the size (and therefore total level) of the larger population became much bigger. However, on average, as the size of the larger population increased, participants' tendency to favor the population with the better total level slightly increased (100,000@50 and million@50: $t(460) = -4.62$, $p < .001$, $d = 0.16$; million@50 and billion@50: $t(460) = -4.54$, $p < .001$, $d = 0.11$). These effects, however, were only robustly significant in the unhappiness condition (100,000@50 and million@50: $t(247) = -4.19$, $p < .001$, $d = 0.21$; million@50 vs billion@50: $t(247) = -4.76$, $p < .001$, $d = 0.19$) but less so in the happiness condition (100,000@50 and million@50: $t(212) = -2.09$, $p = .04$, $d = 0.09$; million@50 vs billion@50: $t(212) = -0.96$, $p = .34$, $d = 0.03$). This asymmetrical sensitivity for the size of populations is in line with the findings of Study 3a. In that study, we had similarly observed that people's preference for larger over smaller happy populations diminished as the respective population sizes increased (and reversed for very large population sizes); in contrast, smaller unhappy populations were consistently preferred, regardless of population size.

Next, we turned to the question of whether thinking style affects people's population ethical judgments. A two-way ANOVA revealed a main effect for thinking style on the aggregated responses (across all five dilemmas), $F(1, 457) = 7.42$, $p = .007$, $\eta_p^2 = 0.02$. In line with our hypothesis, participants had stronger preferences for the populations

Table 4
One-sample t-tests against the midpoint (4) in Study 3c. Degrees of freedom were 212 in the happiness condition and 247 in the unhappiness condition. A positive t-value indicates that mean responses tended to be more in line with totalism. A negative t-value indicates that mean responses tended to be more in line with averagism.

	Happiness		Unhappiness	
	t-value	Cohen's *d*	t-value	Cohen's *d*
100,000@90	6.27***	0.43	4.19***	0.27
100,000@70	−1.09	0.07	−1.58	0.10
100,000@50	−7.67***	0.53	−6.69***	0.43
million@50	−6.20***	0.42	−3.00**	0.19
billion@50	−5.56***	0.38	0.03	0.002

** $p < .01$.
*** $p < .001$.

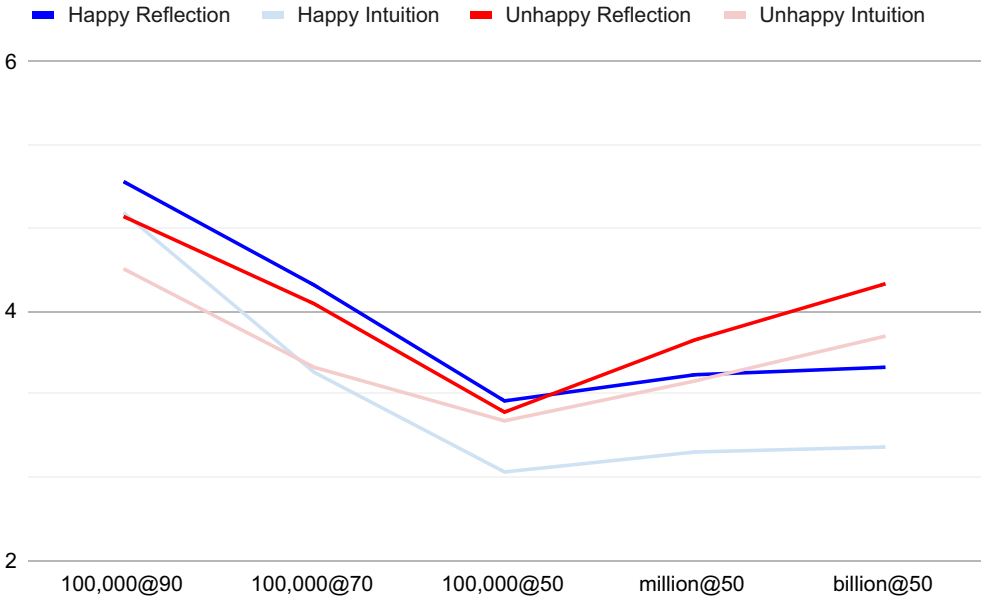

Fig. 8. Participants' preferences in each of the five dilemmas, as function of happiness (happy vs. unhappy) and thinking style (reflective vs. intuitive) in Study 3c. Responses were reverse scored in the unhappiness condition. 1 stands for a preference in line with averagism (populations with better average but worse total levels), 4 stands for no preference, 7 stands for a preference in line with totalism (populations with better total but worse average levels). Participants had to choose between a civilization of 1000 people on level ±100 on the one hand and a civilization of either 100,000 people on level ±90, ±70, ±50, or a civilization of one million people on level ±50, or a civilization of one billion people on level ±50 on the other hand.

with the better average level (and worse total level) when they were thinking intuitively compared to when they were thinking reflectively (Fig. 7). There was no main effect for the valence factor, $F(1, 457) = 1.38$, $p = .24$, $\eta_p^2 = 0.003$, and neither was there an interaction between valence and thinking style, $F(1, 457) = 0.38$, $p = .54$, $\eta_p^2 < 0.001$.

Finally, we looked at associations with demographic measures. Aggregating across all dilemmas and conditions (valence and thinking style), women ($M = 3.74$, $SD = 1.71$) tended to give responses that were more in line with averagism than men ($M = 4.08$, $SD = 1.85$), $t(459) = 2.05$, $p = .04$, $d = 0.19$. This gender effect was moderated by the thinking style manipulation. There was no gender difference in the reflection condition (women: $M = 3.99$, $SD = 1.74$; men: $M = 3.93$, $SD = 1.94$), $t(219) = 0.23$, $p = .82$, $d = 0.03$. But there was a gender difference in the intuition condition, such that women were more inclined to show averagist preferences than men (women: $M = 3.52$, $SD = 1.66$; men: $M = 4.23$, $SD = 1.76$), $t(238) = 3.18$, $p < .001$, $d = 0.41$. There were no other noteworthy associations with demographic measures.

14.3. Discussion

In this study, we investigated participants' preferences in cases where one population has a better average but worse total level of happiness and the other population has a better total but worse average level. We found that in such cases participants' responses were neither strongly aligned with averagism nor with totalism. Instead, they lay between the recommendations of these two principles, suggesting that people apply both principles simultaneously, even when they conflict with each other. Similar to Study 3b, there were no noteworthy differences between the happiness and unhappiness conditions.

Another research question we asked was how people trade-off differences in average or total happiness against each other. We found that, on average, participants were roughly indifferent between a population of 1000 people on level 100 and a population of 100,000 people on level 70. One interpretation is that, in this dilemma, participants considered a drop of 30 average level points roughly to be outweighed by the increase of almost 100,000 people who were still very happy overall. Or, more precisely, they considered the 0.70-fold decrease in average level roughly to be outweighed by a 70-fold increase in total level.

As mentioned above, aggregated across all dilemmas, average responses were neither strongly aligned with averagism nor with totalism but lay between the recommendations of these two principles. However, there were some dilemmas in which participants had clear-cut preferences — for instance, they favored a population of 100,000 people at level 90 over one of 1000 people at level 100 (in line with totalism), and conversely, they quite consistently favored a population of 1000 people at level 100 over populations of 100,000 people, 1 million people, and even 1 billion people at level 50 (in line with averagism). Thus, in accordance with the findings of Study 3a and 3b, we found that responses were influenced by changes in the relative average and total levels of the two populations. The fact that participants' responses stayed roughly in line with averagism even when the size of the larger population with a 50-point average level became extremely large is noteworthy. It suggests that people appear to have a threshold for a minimally acceptable average level that must be met, such that even a much greater total level cannot outweigh the disvalue of a population with an average level that is below this threshold (which seems to be greater than 50).

We also found that, in line with our hypothesis, thinking style affected participants' population ethical intuitions. When prompted to think intuitively, participants' responses were more inclined towards averagism. And when prompted to think reflectively, participants' responses were more inclined towards totalism. In Study 3d, we explore this finding further.

An open research question that our study cannot answer is the relative weight people intuitively place on the averagism and totalism *principles* when these two conflict. Such an analysis is difficult due to an asymmetry in the range of possible values between averagism and totalism. For totalism, there are no lower and upper limits because population sizes can get infinitely big. For averagism, by contrast, clear upper and lower limits were defined as −100 and +100. Because of this, total utility differences can get much bigger (at least in numeric terms) than average utility differences. Further, the availability of such upper and lower limits offers a reference point that makes it easier to evaluate how good or bad a certain average level is (see the evaluability heuristic; Hsee & Zhang, 2010). For example, it is easy to tell that an average level of +90 is very good because it is close to the maximum, whereas it is difficult to tell how good a total level of +9,000,000 is. This difference in evaluability between average and total levels may push people to focus more on average levels. An additional complication is that there could be diminishing marginal sensitivity for total levels. That is, comparing 1 person vs. 4 million people may seem very different from comparing 100,001 people vs. 5 million people. More research is required to investigate more precisely how people weigh averagism against totalism.

The dilemmas used in this study resemble the Repugnant Conclusion, which was discussed by Derek Parfit (1984) as an argument against totalism. Parfit's argument, however, considers an extremely large population with an average happiness level just barely above zero. Our dilemmas did not feature such an extreme population. However, despite this difference, we believe our findings capture the same intuitions people have in response to the philosophical argument. Based on our results, it is likely—as Parfit hypothesized—that people would prefer a smaller population on average level + 100 than an extremely large population with average level + 1, even if the second population has a higher total level.

15. Study 3d: mere additions

In Study 3d, we investigated whether people show averagist tendencies even in cases where averagism favors adding unhappy people or disfavors adding happy people to a population — that is, cases in which most philosophers agree that averagism is wrong (Greaves, 2017).

To test this, we asked participants which out of two populations they consider better: a population consisting of 1000 very happy (unhappy) people or a population of 2000 people consisting of 1000 very happy (unhappy) people and an additional 1000 people who are also happy (unhappy) but to a lesser extent than the first 1000. The only difference between the two populations is, therefore, that the larger population contains an additional 1000 slightly less happy (or unhappy) people. Thus, from a totalist perspective the larger happy population is better because the additional people increase the total happiness. Similarly, from a totalist perspective the larger unhappy population is worse because the additional people increase the total suffering. However, from an averagist perspective, the smaller happy population is better because the additional happy people worsen the average happiness level. Similarly, from an averagist perspective, the larger unhappy population is better because the additional unhappy people improve the average unhappiness level. Based on our previous findings, we expected participants' judgments to be influenced by both averagist and totalist tendencies.

Furthermore, we were interested in whether people's responses are affected by whether they think intuitively or reflectively. In line with our findings of Study 3c, we hypothesized that when people think reflectively as opposed to intuitively this would reduce their averagist tendencies.

Our study was pre-registered at https://aspredicted.org/c56dx.pdf and had a 2 (valence: happiness vs unhappiness) x 2 (thinking style: intuition vs reflection) between-subjects study design.

15.1. Methods

15.1.1. Participants

We recruited 613 US American participants online via MTurk ($0.40 payment per participant). 131 were excluded, leaving a final sample of 482 people (199 female, M_{age} = 37.68, SD_{age} = 10.82). We aimed to recruit at least 600 participants. The sample size was set in advance based on rough approximations of what would be needed to comfortably detect the smallest effect sizes of interest; but they were not based on precise power analyses. Due to the large number of excluded participants, we conducted the same analyses without any exclusions. The pattern of the results remained the same.

15.1.2. Procedure and materials

Participants were randomly assigned to one of the four conditions that resulted from crossing valence with thinking style. First, participants were presented with the happiness scale similar to the previous studies. To ensure that they understood the scale they were also told that "A life above the neutral point (0) means it is a life worth living. A life below the neutral point (0) means it is a life not worth living." Next, they had to indicate whether they accepted these assumptions: "I accept that lives on negative levels, such as lives on level -100, -90, -50, or -10, are not worth living.", "I accept that lives on positive levels, such as lives on level +100, +90, +50, or +10, are worth living.", "I do not accept the information provided.". Participants who did not accept these assumptions were excluded.

Next, participants were again presented with the same prompts of Study 3c to rely either on intuition or on deliberation, depending on condition assignment. Finally, participants were presented with the main task that consisted of three separate dilemmas presented in randomized order. All three dilemmas comprised a contrast between two populations. Civilization A always consisted of 1000 people at level +100 (happiness condition), or level −100 (unhappiness condition), whereas Civilization B consisted of the same 1000 people plus an additional 1000 people at more moderated happiness or unhappiness levels. The key difference between the three vignettes was that the happiness level of the additional 1000 people in civilization B was varied (±10, ±50, ±90). For example, the 'Ten' dilemma, read as follows: "*Consider the following two possible human civilizations: Civilization A contains 1,000 people of which all 1,000 people are on level +100 (extreme happiness). Civilization B contains 2,000 people of which 1,000 people are on level +100 and another 1,000 people are on level +10. (Assume that this civilization has no issues of societal inequality because the two groups live so far apart that they will never meet.). If only one civilization could come into existence, which one would be better?*" Participants responded on a 7-point scale (1 = *Civilization A is much better*, 4 = *Both are equally good*, 7 = *Civilization B is much better*).

Next, participants responded to an attention check question that asked how many people there were in civilization A. Participants were then asked a follow-up question: "How good or bad do you think it is for new happy [unhappy] people (who lead lives that are [not] worth living) to come into existence, if their lives are less intensely happy [unhappy] than the current average happiness [unhappiness] level of the population?" (1 = *Very bad*, 4 = *Neither good nor bad*, 7 = *Very good*). Finally, participants responded to demographic questions.

15.2. Results

The reported analyses include only participants (N = 482) who correctly responded to the check questions. 67 participants did not accept the premise that lives on negative levels are not worth living, as stated at the beginning of the study. 16 participants did not accept that lives on positive levels are worth living. 68 participants failed the attention check (asking how many people civilization A had). We report the same analyses with the full sample without any exclusions in the Supplementary Materials. The pattern of the results is the same.

The responses in the unhappiness conditions were reverse scored, such that low scores indicate averagist tendencies and high scores indicate totalist tendencies. In general, responses tended to be more in line with averagism than totalism — that is, participants tended to prefer not adding new moderately happy people to an already maximally happy population, but they did tend to prefer adding new moderately unhappy people to a maximally unhappy population (see Fig. 8, Table 5). Thus, mean responses in all conditions were significantly below the midpoint of 4. There were only three exceptions: reflection +50 (p = .48), intuition +90 (p = .39), reflection +90 (p = .01). That is, the only task for which responses were totalist was the +90 task in the reflection condition.

For the main analysis, we aggregated responses to each of the three vignettes together to form a single score per participant (α = 0.86). A two-way ANOVA revealed two main effects but no significant interaction effect (Fig. 9). Participants had stronger averagist tendencies when asked about unhappy populations than happy populations, $F(1, 478) = 21.89$, $p < .001$, $\eta_p^2 = 0.05$. And they had stronger averagist tendencies when asked to rely on intuition than on reflection, $F(1, 478) = 7.99$, $p = .005$, $\eta_p^2 = 0.02$. The interaction between valence and thinking style was not significant, $F(1, 478) = 0.82$, $p = .37$, $\eta_p^2 = 0.002$. Tukey HSD post-hoc tests revealed that participants were more totalist under reflection compared to intuition when asked about happy populations, $p = .04$, $d = 0.30$. By contrast, when asked about unhappy populations, participants were not significantly more totalist under reflection compared to intuition, $p = .51$, $d = 0.21$.

Only a minority of participants had strict averagist views (selecting 1) or strict totalist views (selecting 7). Across conditions, the proportion of participants who had strict averagist views was: 40% (±10 task), 30% (±50 task), 17% (±90 task). The proportion of participants who had strict totalist views was: 7% (±10 task), 10% (±50 task), 16% (±90 task). This supports our hypothesis that most people's views are a mixture of these two principles.

When asked explicitly in the abstract, participants said that they would find it good for new happy people (who lead lives that are worth living) to come into existence even if their lives are less intensely happy than the current average happiness level of the population (M = 4.85, SD = 1.51), as revealed by a one-sample t-test against the mid-point of 4, $t(236) = 8.68$, $p < .001$. By contrast, participants said that they would find it bad for new unhappy people (who lead lives that are not worth living) to come into existence even if their lives are less intensely unhappy than the current average unhappiness level of the population (M = 3.31, SD = 1.49), as revealed by a one-sample t-test against the mid-point of 4, $t(244) = -7.21$, $p < .001$. Thus, in the abstract task, the dominant response tendency (totalism) conflicted with the dominant response tendency in the concrete task (averagism). Moreover, 27% of participants in the happiness condition and 35% of participants in the unhappiness condition gave a totalist response to the abstract question but an averagist response in the main task (when aggregating across the three dilemmas). However, there was still some relation between the

Table 5
Mean ratings in Study 3d for comparisons between a population of 1000 people on level + 100 (−100) and a population of 2000 people, consisting of 1000 on level ± 100 and an additional 1000 on level + 10 (±50, ±90 respectively). The responses in the unhappiness conditions were reverse scored. 1 indicates the averagist view (the population with the better average level), 4 indicates the view that both populations are equally good, 7 indicates the totalist view (the population with the better total level).

	Happy		Unhappy	
	Intuition	Reflection	Intuition	Reflection
Aggregated	3.33 (1.92)	3.91 (1.90)	2.73 (1.41)	3.03 (1.51)
±10	2.71 (2.01)	3.39 (2.11)	2.27 (1.56)	2.46 (1.75)
±50	3.11 (2.17)	3.86 (2.15)	2.61 (1.67)	2.98 (1.79)
±90	4.17 (2.20)	4.48 (2.06)	3.3 (1.69)	3.65 (1.83)

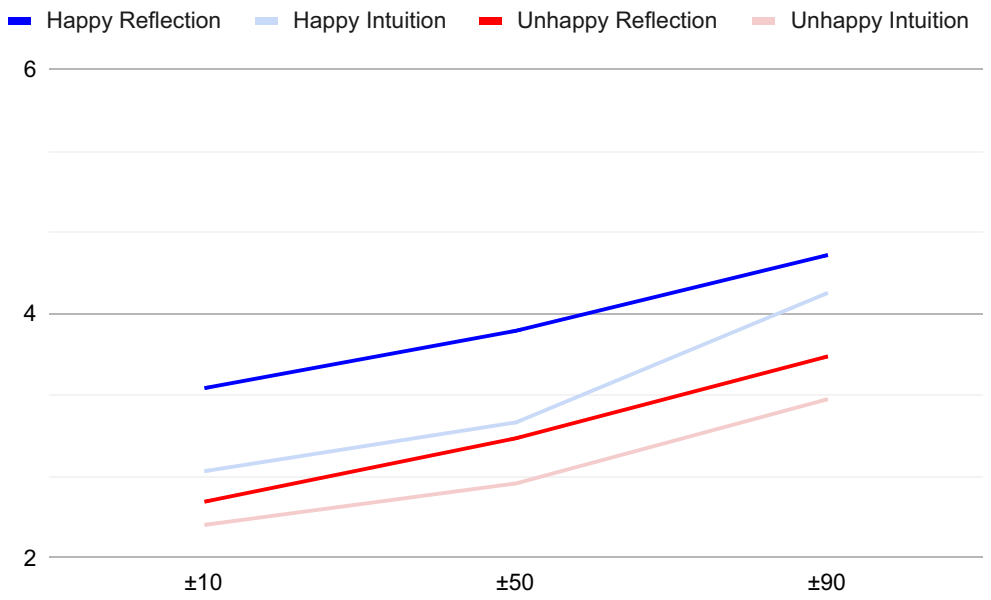

Fig. 9. Participants' responses to the three separate dilemmas in Study 3d, as a function of happiness (happy vs. unhappy) and thinking style (reflective vs. intuitive). The responses in the unhappiness conditions were reverse scored. 1 indicates the averagist view (favoring the population with the better average level), 4 indicates the view that both populations are equally good, 7 indicates the totalist view (favoring the population with the better total level).

two tasks. In both conditions, the more totalist responses were in the main task, the more totalist responses were to the abstract question (happiness: $r(235) = 0.31$, $p < .001$; unhappiness: $r(243) = 0.18$, $p = .004$).

There were no noteworthy associations with demographic measures. In contrast to Study 3c, there were no significant differences between women ($M = 3.41$, $SD = 1.65$) and men ($M = 3.22$, $SD = 1.76$) in their responses, $t(571) = 1.41$, $p = .16$, $d = 0.11$. There were also no gender differences when controlling for condition assignment.

15.3. Discussion

The results of Study 3d demonstrate that participants showed averagist tendencies for both happy and unhappy populations. That is, they preferred a smaller population of very happy people over a larger population consisting of the same people and additional people that are happy to a weaker extent. Similarly, participants preferred a larger population consisting of very unhappy people and people that are unhappy to a weaker extent over a smaller population consisting of only the very unhappy people. Thus, even in an extreme case in which the choice was to add an additional 1000 people whose lives come close to the "absolute worst form of suffering imaginable" (e.g., −90) in order to improve the average well-being of the population (e.g., −100), participants tended to make this choice.

In line with the findings of Study 3c, we found that when participants are asked to rely on deliberate reflection instead of intuition, their averagist tendencies reduce. This suggests that when thinking more deliberately about the question, people tend to realize that focusing just on the average level has undesirable implications. Indeed, most philosophers consider averagism wrong, at least in the case our study covered that involved unhappiness (e.g., Greaves, 2017). In contrast to our hypothesis there was no interaction between thinking style (intuition vs reflection) and valence (happiness vs. unhappiness).

The present findings are in line with the results of Study 3b, which demonstrated a tendency to apply averagism with total well-being held constant. The results also tend to suggest that whereas people are sensitive to total well-being when average levels are held constant (Study 3a), they tend to apply averagism rather than totalism when these two principles come into conflict. This is especially the case for unhappy populations.

The results also arguably conflict with the findings of Studies 2a-b, which showed that people find it good to add a happy person to an empty world and bad to add an unhappy person. If people think that it is bad to add an unhappy person to a world (or vice versa for a happy person), why do they favor larger populations with *more* unhappy people (but lower average unhappiness levels)? One difference between the present study and Studies 2a-b is that the level of happiness or unhappiness of the additional person(s) was maximal (−100 or +100) in Studies 2a-b, whereas it was sub-maximal in the present study. Perhaps more pertinently, the nature of the questions was different. The present study posed a dilemma that called for participants to choose between averagism and totalism, whereas Studies 2a-b posed no such dilemma — participants were asked simply about the goodness or badness of adding a single person to an empty world. In that context, averagism and totalism both favor the same response. A final contrasting factor is that Studies 2a-b's question was abstract and relatively context free, whereas in the present study, participants compared two populations side-by-side, which were described in terms of their population sizes and their well-being distributions. It may be that with this more concrete presentation, people feel a greater pull towards averagism.

Supporting this last speculation, in the present study, when asked explicitly in the abstract, participants indicated that they consider it good to add new happy people to a population even if this worsens the average happiness level, and similarly, that it is bad to add new unhappy people to a population even if this improves the average unhappiness level. That abstract response conflicts with participants' responses in the main task, in which they tended to favor the addition of such new people, thereby showing an averagist preference. What might cause this difference — that is, why would people answer the abstract question differently than the more concrete one? (For an investigation into an abstract/general vs concrete effect in moral judgment, see Caviola, Schubert, & Mogensen, 2021). One possible explanation is that people do not perceive the questions asked in the main tasks as questions about whether it is good or bad to merely add more people on top of an existing population (that is identical to the other population). Note that, strictly speaking, the questions asked in the main tasks were not about adding people to a pre-existing population but instead about comparing two separate populations. In this case, people may perceive and evaluate the two presented populations holistically by applying their averagist intuition to the whole population of 1000 or 2000 people respectively.

That is, their response incorporates the entire surrounding context. By contrast, when they are asked explicitly whether adding people to an existing population is good or bad, particularly when they are not presented with any additional context about the overall population, people may evaluate this question by focusing solely on whether adding these new people is good or bad. In a similar way, it may be that when people are instructed to reflect rather than rely on intuition, they dissect the main task in this analytical way, focusing more on the specific difference between the two populations, rather than coming to a more holistic judgment of the entire population that results.

16. General discussion

Population ethics is an active field of research in philosophy but has so far been neglected by psychologists. In this paper we provided what to our knowledge is one of the first systematic psychological investigations of lay people's population ethical intuitions. Our key insights across nine studies are the following. When people evaluate populations, they 1) weigh suffering more than happiness, 2) do not view creating new people as morally neutral, 3) focus on improving both the average and total happiness level, but seem especially sensitive to average levels, 4) tend to be more sensitive to average levels when relying on intuition than when relying on reason. We will next discuss these insights in more detail.

16.1. Suffering is more bad than happiness is good

We found that people weigh suffering more than happiness when they evaluate the goodness of populations consisting of both happy and unhappy people. Thus, people appear to follow neither strict negative utilitarianism (minimizing suffering, giving no weight to maximizing happiness at all) nor strict classical utilitarianism (minimizing suffering and maximizing happiness, weighing both equally). Instead, the average person's intuitions seem to track a mixture of these two theories. In Studies 1a-c, participants on average believed that approximately 1.5–3 times more happy people are required to outweigh a given amount of unhappy people. The precise trade ratio between happiness and suffering depended on the intensity levels of happiness and suffering, such that a greater proportion of happiness was required as intensity levels increased (Study 1b). (In additional preliminary studies, we found that the trade ratio can also heavily depend on the framing of the question.) Study 1c clarified that, on average, participants continued to believe that more happiness than suffering was required even when the happiness and suffering units were exactly equally intense. This suggests that people generally weigh suffering more than happiness in their moral assessments, above and beyond perceiving suffering to be more intense than happiness. However, our studies also made clear that there are individual differences and that a substantial proportion of participants weighed happiness and suffering equally strongly, in line with classical utilitarianism.

One explanation for why people weigh suffering more than happiness could be a general negativity effect—the phenomenon that negative events have a greater perceived impact than positive events have (Baumeister et al., 2001; Rozin & Royzman, 2001). While a negativity effect has been observed in various domains, to our knowledge, so far it has not been demonstrated before in moral judgments about the goodness and badness of happiness and suffering. It is possible that a negativity effect drives the asymmetric perception of the intensity of suffering and happiness. There could be neurobiological and evolutionary reasons why people demonstrate this negativity effect and, furthermore, why some people's normative evaluation of happiness and suffering is asymmetric even when the intensity levels of happiness and suffering are equalized (Study 1c). A possible ultimate explanation is that evolution has selected for mechanisms that give greater priority to avoiding suffering than pursuing happiness due to recurring asymmetrical costs in the environment of evolutionary adaptation (error management theory; Haselton & Buss, 2000). That is, negative events tend to have greater objective impact on the health, safety, or survival of humans, than equally intense positive events: failing to avoid harmful actions could lead to death, whereas failing to avoid beneficial actions doesn't have similarly bad consequences. This could explain why for most people the worst suffering—either experienced or imagined—is more intense than the best happiness and perhaps also why some people evaluate happiness and suffering asymmetrically even when they are equally intense.

We found no correlation between participants' current or general mood level and their tendency to weigh suffering more than happiness. However, we did find that participants who were more willing to relive their worst day in order to experience their best day tended to weigh happiness and suffering more similarly. This suggests that the way in which people weigh happiness and suffering for themselves personally affects their judgments of how many happy and unhappy people a population should contain. Similarly, we also found that judgments about the acceptable proportion of happy and unhappy people in a population matched judgments about the acceptable proportion of happiness and unhappiness within a single individual's lifetime. This again suggests that the same mechanisms drive intuitions about evaluating the goodness of individual lives and whole populations (cf. Starmans & Bloom, 2015).

16.2. Creating new people is morally relevant

We found that people do not endorse the so-called *intuition of neutrality* according to which creating new people with lives worth living is morally neutral. In Studies 2a-b, participants considered a world containing an additional happy person better and a world containing an additional unhappy person worse.

Moreover, we also found that people's judgments about the positive value of adding a new happy person and the negative value of adding a new unhappy person were symmetrical. That is, their judgments did not reflect the so-called *asymmetry*—according to which adding a new unhappy person is bad but adding a new happy person is neutral. It is surprising that people's judgments about adding a new happy or unhappy person were symmetrical given that they weighed suffering more than happiness when asked about the appropriate proportion of happy vs. unhappy people in a population in Studies 1a-c.

One possible explanation is that people reach different conclusions when they globally assess the final state of a population compared to when they locally assess a change to a population. In Studies 1a-c, participants were asked to evaluate the acceptable happiness-to-suffering ratio within a given population. By contrast, in Studies 2a-b, participants were asked to assess whether a population improves or worsens if a new happy or unhappy person is added to it. That is, participants were asked to evaluate a change to a population, not its final state. Future research could explore this hypothesis further. Future research could also explore to what extent the evaluation of adding one new person depends on the average or total happiness level of the pre-existing population.

Whereas participants considered it good to add a new happy person in Studies 2a-b, in Study 3a we found that they became less sensitive to the value of adding more happy people, the larger the number of people became. For example, participants preferred 10,000 happy people over 1000 happy people but they did not prefer 10 billion happy people over 1 billion happy people, notwithstanding the identical ratio. In contrast, participants remained sensitive to the disvalue of adding more unhappy people, even with larger numbers of people — in fact, they became more sensitive as the numbers increased. For example, they preferred 1 billion unhappy people over 10 billion unhappy people to a greater extent than they preferred 1000 unhappy people over 10,000 unhappy people.

One possible explanation for this pattern could be a general preference for populations that are not too large. For example, people may be worried that excessively large populations could lead to negative

consequences, such as crowdedness or resource depletion. Even though we tried to rigorously control these factors, it is possible, as suggested by several of our participants' comments, that many people find it difficult to completely decontextualize the scenarios and consider the question in its pure abstract form. Thus, such a preference for populations that are not too big could play a significant role in shaping people's population ethical intuitions when the population sizes are large.

16.3. Both the average and the total should be improved

We found that people have intuitions in line with both averagism and totalism. In Study 3a, we found that participants preferred populations with better total levels when the average levels were held constant. In Study 3b, we found that participants preferred populations with better average levels when the total levels were held constant. In Study 3c, we found that most participants' preferences lay in between the recommendations of these two principles when they conflict, suggesting that participants applied both preferences simultaneously in such cases. Thus, we found that, at least in some cases, participants were willing to trade reduced average levels to achieve improved total levels. But when the average levels became too low (e.g., level 50), they believed that even a much greater population size and total level could not outweigh the loss in the reduced average level. Similarly, in Study 3d, we found that in certain cases where averagism and totalism conflict, participants' averagist preferences tend to dominate, suggesting that participants were particularly sensitive to differences in average levels.

Many philosophers have rejected averagism because it leads to disturbing implications that most consider unacceptable under careful reflection. One example is the so-called *Sadistic Conclusion* according to which adding unhappy people can be better than adding happy people. In Study 3d, we found that people showed averagist tendencies akin to the Sadistic Conclusion. For instance, people preferred not to add new happy people to a population (i.e., they preferred the smaller of two populations) when this would result in a lowering of its average happiness level (despite these new people being moderately happy). Even more striking is the fact that people preferred to add new unhappy people to a population (i.e., they preferred the larger of two populations) when this would result in raising its average happiness level (even when these new people were almost maximally unhappy). However, this does not mean that people explicitly endorse averagism as a moral principle. To the contrary, in the dilemmas used in Studies 3c and 3d, we found that people's preferences became less averagist and more totalist when they were prompted to think reflectively instead of intuitively. And when asked explicitly in the abstract whether populations can be improved by adding more unhappy people in order to improve the population's average level, participants rejected that view. It therefore seems plausible that—at least in the specific dilemmas we used in studies 3c and 3d—people find it intuitive to focus on the average level without being fully aware of its implications. Note that we are not claiming that these findings mean that totalism is necessarily the more rational principle than averagism. Our investigation is purely descriptive.

What explains why people intuitively focus on the average level? First, focusing on the average level may be intuitive in contexts where it is easier to infer the average level than the total level. For example, when populations are described by specifying the number of people, all with homogenous individual happiness levels, focusing on the average level only requires the comparison of two readily available numbers (e.g., Study 3c). By contrast, inferring the total levels requires multiplying the number of people by the stated individual happiness level. However, when the happiness level within a population is heterogeneous, inferring the average level is more difficult and may not necessarily be easier than inferring the total level (e.g., Study 3d). However, we did not systematically test nor control for ease-of-inference in our studies. We therefore would like to emphasize that our findings do not suggest that intuition always favors averagism and reflection always favors totalism. We believe it is very well possible that in certain contexts, e.g., when the total level is easier to infer than the average level, people intuitively would focus more on the total level rather than the average level. More research is required to systematically investigate whether and to what extent ease-of-inference of the average and total levels of a population impacts which criterion people focus on more when responding intuitively (or under deliberation).

Second, it is possible that the averagist intuition we have explored is related to the proportion dominance effect. The proportion dominance effect is the tendency to focus on relative savings rather than the absolute savings (Baron, 1997). For example, people prefer saving 10 out of 10 people instead of 10 out of 100 people. This is in line with our observed averagist tendency. It has been found that the proportion dominance effect is driven by unreflective thinking (Bartels, 2006; Mata, 2016). This also seems in line with our finding that averagist tendencies are enhanced when people are prompted to think intuitively instead of reflectively (Studies 3c-d). Thus, it is possible that averagist tendencies and the proportion dominance effect are driven by similar, or even the same, psychological processes.

Third, people may substitute (cf. attribute substitution, Kahneman & Frederick, 2002) the difficult question of how good the distribution of happiness of a population is with the easier question of how good the same distribution of happiness would be for a single individual, such as themselves. That is, rather than conceiving of the welfare distribution as representing distinct people, they instead envisage it as capturing the distribution of hedonic experiences within a single individual's life. Such a heuristic would only take into account the average levels of happiness (or the ratio between happiness and suffering) but not the population size. The fact that people have the same happiness-to-suffering trade-ratios for populations as for single individuals (Studies 1a and 1c) provides some support for this hypothesis.

16.4. The study of axiological judgments

Most previous research in moral psychology has focused on moral judgments about decision-making or acts (deontics) but not about outcomes per se (axiology). To our knowledge, little previous research has explicitly focused on understanding people's axiological judgments (though for one exception, see Goodwin & Landy, 2014). Our findings suggest that the study of axiological judgments can reveal important insights for moral psychology about the way in which people apply and integrate moral values about outcomes and moral values about decision-making (cf. Cushman, 2013).

People's judgments in our studies suggest that they may be more willing to make utilitarian (happiness-suffering) trade-offs when considering just the *outcomes* than when considering concrete *actions* that need to be taken to achieve these outcomes. We found, for example, that people consider world states that contain substantial amounts of (extreme) suffering morally net positive as long as this suffering is outweighed by sufficient amounts of happiness (Studies 1a-b). By contrast, in Footbridge-like trolley problems people have the strong non-utilitarian intuition that it is wrong to take the action that harms one person to help five others (Greene, 2013; Kahane et al., 2018). And more generally, people are often averse to applying cost-benefit analysis to human lives and consider making such trade-offs taboo (Fiske & Tetlock, 1997; Tetlock, 2003). This suggests that in their axiological judgments about outcomes per se, people do apply a broadly utilitarian cost-benefit analysis. It is only (or mostly) in their judgments about decision-making that they deviate from utilitarianism, e.g., due to deontological constraints that prohibit them from bringing about the outcome that they consider better.

It is possible that axiological judgments and deontic judgments can diverge on population ethical issues. As a first example, in our Studies 1a-c we probed participants' axiological judgments about the appropriate trade-ratio of happiness vs. suffering in a population. That is, participants were asked about the value of the population but not about

any specific actions. In preliminary studies, we found that participants gave different responses when they were asked about actions to bring about such populations. That is, participants were asked what the appropriate trade-ratio of happiness vs. suffering in a population would need to be for them to be willing to push a button that would create such a population. When asked in this (deontic) way, participants stated higher trade-ratios (with more happiness being required) than when asked about the population value only (axiology). We note, however, that there could be other factors, such as framing effects, that could partly account for these differences.

As a second example, in Studies 2a-b we found that people had symmetrical axiological judgments about the (positive vs. negative) value of a world that included an additional happy or unhappy person. But this does not rule out the possibility that they could have asymmetrical deontic judgments about the action of creating a new happy or unhappy person. It is possible that people may consider it morally forbidden to create a new unhappy person but only morally supererogatory—good but not required—to create a new happy person. Future research could explore this possibility.

Similarly, to discuss a case outside population ethics, most people likely agree that a world in which one person suffers is better than a world in which five people suffer. But if the better outcome can only be brought about by actively harming one person (e.g., by pushing them off a footbridge), they might consider that action wrong despite their preference for the outcome in which fewer people suffer. More research is required to investigate the relation of deontic and axiological judgments in general and people's deontic judgments about population ethical questions in particular.

16.5. Limitations

In our studies, we focused only on manipulating hedonic welfare, i.e., happiness and suffering, between populations. While people clearly take into account hedonic welfare when making population value judgments, it is likely that they also believe other aspects of people's lives to have intrinsic value. For example, they might think that Van Gogh's life was worth living, notwithstanding that he arguably experienced a net negative balance of suffering to happiness. That is, they would not wish, for his sake, that he had never been born, because he also achieved artistic excellence and accomplished extraordinary things. Thus, it could be that people assume that lives have other features that make life worth living, such as achievement of perfectionist value of the kind that lives like Van Gogh's have in very high degrees (Bradford, 2015). In such cases, the balance of suffering to happiness would have to be strongly asymmetric for the individual life not to be worth living. Our studies did not take such non-welfare values into account.

In our studies, participants were presented with the explicit or implicit assumption that there is a happiness scale, ranging from −100 (extreme suffering), to 0 (neutral), to +100 (extreme happiness). However, we acknowledge that it is not obvious whether such an objective scale exists. Instead, such a scale may rather reflect subjective estimates that can differ between people. A happiness scale might thus not be analogous to objective scales, such as those that assess mass, but rather, analogous to subjective scales, such as those that assess beauty or taste. Due to the subjectivity of perception regarding the evaluation of happiness and suffering it is unclear how exactly participants interpreted the happiness scale they were presented with. It is, for example, unclear whether people accept and converge on the view that there is a neutral happiness point at which life is neither worth living nor not worth living. Despite these uncertainties, our results, including participants' comments in the open text fields, do not suggest that they objected to the presented happiness scale.

As explained in the Introduction section, our studies relied on different questions and response scales to assess participants' population ethical intuitions. In Studies 1a-c and Studies 2a-b, participants were asked about overall population value. In Studies 1a-c, a positive overall population value was specified as reflecting the view that "it would be better for the world to exist rather than not exist". In Studies 2a-b, participants were asked whether one population compared to another would be "better or worse (…) in terms of its overall value". In Studies 3a and 3b, participants were asked which population they would prefer to come into existence. And in Studies 3c and 3d, participants were asked which out of two populations they think would be better to bring into existence. It cannot be ruled out that these different types of questions elicit different responses, and we hope future research tests this.

In most of our studies, participants were forced to decide which population they consider better. However, some philosophers have argued that two populations could also be incommensurable (for a discussion, see Broome, 2005), i.e., that neither is better than the other and yet nor are they exactly equally good. While in these studies we allowed participants to indicate that they are indifferent between the two options by choosing the midpoint, it is possible that some people would have instead indicated that the answer cannot be determined. In Study 2b, we tried to address this question by allowing participants to select an option stating that they consider the question nonsensical or that they don't know how to answer it. We found that less than 2% chose that option, weakening this concern. However, since we did not offer this option in all studies, it is still possible that some participants consider it too difficult or impossible to respond to tricky population ethical questions. Another possibility is that the availability of an option stating that they consider the question nonsensical or that they don't know how to answer it did not allow subjects an appropriate means by which to express the judgment that the outcomes are incommensurable, as this need not entail dismissing the question as meaningless. For example, according to Chang (2002), two items may be evaluatively comparable without it being the case that one must be better than the other or else they must be equally good. Chang claims that there exists a fourth possible relation of value comparability, which she calls 'being on a par'. Future research could explore this further.

16.6. Future research

Our studies primarily focused on demonstrating people's general tendencies to prefer one population over another. Future research could investigate in greater detail what psychological mechanisms drive people's population ethical tendencies. For example, future research could test our hypothesis that the averagist tendency is driven by a heuristic that substitutes the population ethical question with the question of how good an individual life within the given happiness distribution would be. Future research could also investigate the mechanisms that drive people to have asymmetrical judgments when evaluating populations with fixed sizes consisting of both happy and unhappy people but symmetrical judgments when evaluating whether a population improves or worsens when happy or unhappy people are added.

We found that people have both averagist and totalist preferences and that they apply them simultaneously when both the average and total levels vary. This implies that people somehow must make intertheoretic comparisons between averagism and totalism. That is, they must intuitively trade the relative utility gains of one principle against the relative utility loss of the other principle. How precisely this is done is not clear. Philosophers have discussed various approaches to intertheoretic comparisons, such as the normalization of theories based on their variance (Greaves, 2017; MacAskill, 2014; MacAskill, Cotton-Barratt, & Ord, 2020). Future research could investigate how people make such intuitive intertheoretic comparisons.

How do people aggregate vast amounts of small units together? A famous example is the so-called Repugnant Conclusion (Parfit, 1984). Do people think that there is a number of slightly happy people—who are all on level +1—that is large enough for this population to be overall better than a smaller population of people who are all on level +100? Even though we did not directly test this, our results of Study 3c suggest

that people would likely not think that such a number exists.

16.7. Implications

We found that people's population ethical judgments about concrete cases are influenced by different intuitions, which do not always square with judgments made under careful reflection. For example, we found that most participants believed that more happy than unhappy people are required for a population to be net positive (Studies 1a-c), yet most participants made symmetrical judgments regarding the goodness of adding a new happy person or the badness of adding a new unhappy person (Studies 2a-b). And while participants found it good to add a new happy person, the perceived marginal value of adding a new happy (or unhappy) person declined when populations got very large (Study 3a). And even more disturbingly, participants preferred populations with better average levels even if this meant that those populations contained additional people experiencing torture-like suffering on level −90. These are examples in which people's judgments appear unreasonable and at odds with most philosophers' views (and possibly even their own more carefully considered views). They demonstrate that caution is warranted when drawing normative conclusions directly from lay people's population ethical intuitions.

However, this does not mean that it is not valuable to examine lay people's population ethical intuitions. Population ethics has important implications for policy making and global priority setting. Philosophers often rely on their own intuitions when discussing population ethics. An understanding of the psychology of these population ethical intuitions can therefore be informative. For example, greater awareness of the specific psychological mechanisms and biases driving these intuitions could elucidate which ones should be endorsed under reflection and which ones not. The apparent inconsistencies between some of these intuitions demonstrate that it may be impossible to formulate a population ethical theory that is both consistent and intuitive (cf. impossibility theorems; Arrhenius, 2000). One possible solution could be a debunking approach: attempting to understand the psychological underpinnings of different philosophical positions, with an eye to identifying those that result from unreliable or biased cognitive processes. This in turn allows the resolution of inconsistency by discounting certain intuitions as untrustworthy (cf. Greene, 2013). Another possible resolution is to accept the fact that we are internally conflicted and, as a consequence, uncertain which moral theory is right (MacAskill, Bykvist, & Ord, 2020).

Despite the fact that people's population ethical intuitions can be biased and inconsistent, we found some response patterns that appear to be relatively robust. Participants were generally sensitive to the total amount of happiness and unhappiness the world contains, and in most cases they considered creating new happy people morally good and new unhappy people morally bad. We also found that, at least in our dilemmas, people's responses became more in line with totalism when they thought more reflectively about their answers. While it is too early to draw any practical ethical implications from these findings, they suggest that many people, especially after more careful reflection, may be supportive of a broadly totalist position, according to which vast numbers of (un)happy people would be extremely (un)desirable. Given that the number of people who could come into existence in the future vastly outnumbers the number of currently existing generations (Hauser, Rand, Peysakhovich, & Nowak, 2014), this position implies that future generations should be weighed much more heavily in our moral decision making than they currently are (Beckstead, 2013; Bostrom, 2013; Greaves & MacAskill, 2019; Mogensen, 2021; Ord, 2020; Parfit, 2011, section 36). Future research could investigate to what extent lay people endorse these practical implications that may follow from their population ethical values.

Credit author statement

LC conceptualized the project. LC, DA, ALM, and GPG planned the studies. LC and DA collected and analyzed the data. LC, DA, ALM, and GPG interpreted the data. LC wrote the first draft. LC, DA, ALM, and GPG wrote the manuscript.

Author note

Reports of all measures, manipulations, and exclusions, as well as all data, analysis code, and experimental materials are available for download at https://osf.io/qt65w/.

Acknowledgments

We thank Carl Shulman, Joshua Lewis, Pablo Stafforini, William MacAskill and the attendees of the Global Priorities Research Workshop for their valuable inputs. This research was funded by Effective Altruism Funds and the Center on Long-Term Risk.

References

Arrhenius, G. (2000). An impossibility theorem for welfarist axiologies. *Economics and Philosophy, 16*(2), 247–266.
Baron, J. (1997). Confusion of relative and absolute risk in valuation. *Journal of Risk and Uncertainty, 14*(3), 301–309.
Bartels, D. M. (2006). Proportion dominance: The generality and variability of favoring relative savings over absolute savings. *Organizational Behavior and Human Decision Processes, 100*(1), 76–95.
Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370.
Bealer, G. (1998). Intuition and the autonomy of philosophy. *Rethinking Intuition* (pp. 201–240). Lanham, MD: Rowman and Littlefield.
Beckstead, N. (2013). *On the overwhelming importance of shaping the far future*. Doctoral dissertation. Rutgers University-Graduate School-New Brunswick.
Birnbaum, M. H. (1972). Morality judgments: Tests of an averaging model. *Journal of Experimental Psychology, 93*(1), 35.
Birnbaum, M. H. (1973). Morality judgment: Test of an averaging model with differential weights. *Journal of Experimental Psychology, 99*(3), 395.
Biswas-Diener, R., & Diener, E. (2009). Making the best of a bad situation: Satisfaction in the slums of Calcutta. In *Culture and well-being* (pp. 261–278). Dordrecht: Springer.
Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy, 4*(1), 15–31.
Bradford, G. (2015). Perfectionism. In *The Routledge handbook of philosophy of well-being* (pp. 140–150). London, UK: Routledge.
Broome, J. (2004). *Weighing lives*. Oxford, UK: Oxford University Press.
Broome, J (2005). Should we value population? *The Journal of Political Philosophy, 13*(4), 399–413.
Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space_ A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115*(3), 401.
Cappelen, H. (2012). *Philosophy without intuitions*. Oxford, UK: Oxford University Press.
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.
Chang, R. (2002). The possibility of parity. *Ethics, 112*(4), 659–688.
Chudnoff, E. (2013). *Intuition*. Oxford, UK: Oxford University Press.
Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.
Desvousges, William, Johnson Reed, F., Dunford, Richard, Boyle, Kevin, Hudson, Sara, Wilson, Nicole, et al. (1992). *Measuring Nonuse Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy*. NC: Research Triangle Institute.
Diener, E., & Diener, C. (1996). Most people are happy. *Psychological Science, 7*(3), 181–185.
Diener, E., Oishi, S., & Tay, L. (2018). Advances in subjective well-being research. *Nature Human Behaviour, 2*(4), 253–260.
Fetherstonhaugh, D., Slovic, P., Johnson, S., & Friedrich, J. (1997). Insensitivity to the value of human life: A study of psychophysical numbing. *Journal of Risk and Uncertainty, 14*(3), 283–300.
Fiore, J., Becker, J., & Coppel, D. B. (1983). Social network interactions: A buffer or a stress. *American Journal of Community Psychology, 11*(4), 423–439.
Fiske, A. P., & Tetlock, P. E. (1997). Taboo trade-offs: Reactions to transactions that transgress the spheres of justice. *Political Psychology, 18*(2), 255–297.
Frederick, S., & Fischhoff, B. (1998). Scope (in) sensitivity in elicited valuations. *Risk Decision and Policy, 3*(2), 109–123.
Frick, J. D. (2014). *"Making people happy, not making happy People": A Defense of the asymmetry intuition in population ethics (doctoral dissertation)*. Cambridge, MA, USA: Harvard University.

Goodwin, G. P., & Landy, J. F. (2014). Valuing different human lives. *Journal of Experimental Psychology: General, 143*(2), 778.
Greaves, H. (2017). Population axiology. *Philosophy Compass, 12*(11), Article e12442.
Greaves, H., & MacAskill, W. (2019). *The case for strong longtermism*. Global Priorities Institute Working Paper.
Greaves, H., MacAskill, W., O'Keeffe-O'Donovan, R., Trammell, P., Tereick, B., Mogensen, A., … Herrmann, S. (2020). *A research agenda for the Global Priorities Institute*. Global Priorities Institute. https://globalprioritiesinstitute.org/research-agenda/.
Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York, USA: Penguin.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.
Hardin, G. (1968). The tragedy of the commons: the population problem has no technical solution; it requires a fundamental extension in morality. *Science, 162*(3859), 1243–1248.
Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology, 78*(1), 81.
Hauser, O. P., Rand, D. G., Peysakhovich, A., & Nowak, M. A. (2014). Cooperating with the future. *Nature, 511*(7508), 220–223.
Holtug, N. (2004). Person-affecting moralities. In *The repugnant conclusion* (pp. 129–161). Dordrecht, NL: Springer.
Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science, 5*(4), 343–355.
Huemer, M. (2005). In *Ethical intuitionism*. London: Palgrave Macmillan.
Hurka, T. (1983). Value and population size. *Ethics, 93*(3), 496–507.
Hurka, T. (2010). Asymmetries in value. *Nous, 44*(2), 199–223.
Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology, 75*(4), 887–900.
Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*(2), 131.
Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment, 49*, 81.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica, 47*, 263–291.
Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*(4), 341–350.
Lewis, D. (1983). *Philosophical papers volume I*. New York, USA: Oxford University Press.
MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral uncertainty*. Oxford, UK: Oxford University Press.
MacAskill, W., Cotton-Barratt, O., & Ord, T. (2020). Statistical normalization methods in interpersonal and Intertheoretic comparisons. *The Journal of Philosophy, 117*(2), 61–95.
Manne, S. L., Taylor, K. L., Dougherty, J., & Kemeny, N. (1997). Supportive and negative responses in the partner relationship: Their association with psychological adjustment among individuals with cancer. *Journal of Behavioral Medicine, 20*(2), 101–125.
Mata, A. (2016). Proportion dominance in valuing lives: The role of deliberative thinking. *Judgment and Decision making, 11*(5), 441–448.
McMahan, J. (1981). Problems of population choice. *Ethics, 92*(1), 96–127.
Mogensen, A. L. (2021). *Moral demands and the far future*. Philosophy and Phenomenological Research, Article forthcoming.
MacAskill, W. (2014). *Normative uncertainty*. University of Oxford: Doctoral dissertation.
Narveson, J. (1973). Moral problems of population. *The Monist, 57*(1), 62–86.
Nebel, J. M. (2019). Asymmetries in the value of existence. *Philosophical Perspectives, 33*(1), 126–145.
Ng, Yew-Kwang (1989). What should we do about future generations? Impossibility of Parfit's Theory X. *Economics and Philosophy, 5*(2), 235–253. https://doi.org/10.1017/S0266267100002406
Nozick, Robert (1974). *Anarchy, State, and Utopia* (pp. 42–45). New York: Basic Books.
Ord, T. (2020). *The precipice: existential risk and the future of humanity*. New York, NY: Hachette Books.
Parfit, D. (1984). *Reasons and persons*. Oxford, UK: Oxford University Press.
Parfit, D. (2011). *On what matters: Volume two*. Oxford, UK: Oxford University Press.
Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology, 1*(1), 33–60.
Riskey, D. R., & Birnbaum, M. H. (1974). Compensatory effects in moral judgment: Two rights don't make up for a wrong. *Journal of Experimental Psychology, 103*(1), 171.
Roser, Max, & Ortiz-Ospina, Esteban (2020). Global Extreme Poverty. *OurWorldInData.org*. https://ourworldindata.org/extreme-poverty.
Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320.
Schubert, S., Caviola, L., & Faber, N. S. (2019). The psychology of existential risk: Moral judgments about human extinction. *Scientific reports, 9*(1), 1–8.
Sidgwick, H. (1981/1907). *Methods of ethics* (7th Revised ed.). Indianapolis, Indiana: Hackett Publishing Co, Inc.
Spears, D. (2017). Making people happy or making happy people? Questionnaire-experimental studies of population ethics and policy. *Social Choice and Welfare, 49*(1), 145–169.
Spears, D. (2019). The asymmetry of population ethics: Experimental social choice and dual-process moral reasoning. *Economics and Philosophy*, 1–20.
Starmans, C., & Bloom, P. (2015). *A series of separate selves: Happiness distributions across a lifespan mirror happiness distributions across a group*. Yale University.
Tappin, B. M., & Capraro, V. (2018). Doing good vs. avoiding bad in prosocial choice: A refined test and extension of the morality preference hypothesis. *Journal of Experimental Social Psychology, 79*, 64–70.
Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences, 7*(7), 320–324.