Synthese (2024) 203:140
https://doi.org/10.1007/s11229-024-04544-9

ORIGINAL RESEARCH

Reflective equilibrium in practice and model selection: a methodological proposal from a survey experiment on the theories of distributive justice

Akira Inoue[1] · Kazumi Shimizu[2] · Daisuke Udagawa[3] · Yoshiki Wakamatsu[4]

Received: 28 February 2023 / Accepted: 20 February 2024

Abstract

In political philosophy, reflective equilibrium is a standard method used to systematically reconcile intuitive judgments with theoretical principles. In this paper, we propose that survey experiments and a model selection method—i.e., the Akaike Information Criterion (AIC)-based model selection method—can be viewed together as a methodological means of satisfying the epistemic desiderata implicit in reflective equilibrium. To show this, we conduct a survey experiment on two theories of distributive justice, prioritarianism and sufficientarianism. Our experimental test case and AIC-based model selection method demonstrate that the refined sufficientarian principle, a widely accepted principle of distributive justice, is no more plausible than the prioritarian principle. This tells us that some changes of certain intuitions revolving around sufficientarianism should be examined (separately) based on the findings of the survey experiment and AIC model selection. This shows the potential of our approach—both practically and methodologically—as a novel way of applying reflective equilibrium in political philosophy.

✉ Akira Inoue
inoueakichan@g.ecc.u-tokyo.ac.jp

Kazumi Shimizu
skazumi1961@gmail.com

Daisuke Udagawa
udagawa.daisuke@gmail.com

Yoshiki Wakamatsu
nappa10292001@gmail.com

[1] Department of Advanced Social and International Studies, Graduate School of Arts and Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

[2] Waseda University, 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, Japan

[3] Hannan University, 5-4-33 Amamihigashi, Matsubara, Osaka 580-8502, Japan

[4] Gakushuin University Law School, 1-5-1 Mejiro, Toshima-ku, Tokyo 171-0031, Japan

Published online: 24 April 2024

Keywords AIC · Model selection · Prioritarianism · Reflective equilibrium in practice · Sufficientarianism · Survey experiment

1 Introduction

Reflective equilibrium is widely used as a method to bring our intuitions into accord with theoretical principles. While reflective equilibrium is used in philosophy in general (Baumberger & Brun, 2021; Cath, 2016; Elgin, 1996), it is particularly seen as a reasonable method (among others) for justifying normative principles in political philosophy (Knight, 2017; Scanlon, 2003; Sinnott-Armstrong et al., 2010; Varner, 2012).

However, difficulties arise when we try to collect a wide range of people's intuitions as a (provisional) basis for justification in terms of reflective equilibrium. The difficulties arise from the fact that we have to find intuitions latent in our moral practice. Thus, political philosophers do not simply try to gather intuitions: To evoke our intuitions in a controllable manner, they often use thought experiments that consider hypothetical scenarios, but purport to reflect real dilemmas we confront, such that the available options do not allow for desirable outcomes in practical contexts. However, there are also challenges to the use of thought experiments, such that the aroused intuitions are not reliable because the experimental settings are far removed from reality (Anscombe, 1957; Goodin, 1982; O'Neil, 1989). Political philosophers must ward off these challenges.

This paper shows that we can defuse the challenges to the use of thought experiments and thus overcome the difficulties of gathering and inducing intuitions to justify normative principles. We propose a twofold approach. *(i) Use survey experiments (if available):* Survey experiments are conducted by using a poll-style survey to investigate people's intuitive views. We show that this allows for the use of thought experiments—more generally, the use of possible cases—to avoid the aforementioned challenges with available samples. *(ii) Use a model selection method for coping with our intuitive reactions to thought experiments in a systematic manner*. We propose the use of the Akaike Information Criterion (AIC) as a feasible way to select normative principles that would systematically account for the intuitions. This illustrates that a survey experiment and a model selection method can be viewed together as a methodological means of satisfying the epistemic desiderata implicit in reflective equilibrium. To show this, we conduct a survey experiment on two theories of distributive justice, prioritarianism and sufficientarianism. We then analyze the results using a model selection method. Through the combination of the survey experiment and AIC model selection, we are able to show that the refined sufficientarian principle, a widely supported principle of distributive justice, cannot be considered more plausible than the prioritarian principle. This tells us that some changes of certain intuitions revolving around sufficientarianism should be examined (separately), which is an important stage of reflective equilibrium. Thus, our proposed approach can contribute to the development of reflective equilibrium as a method of political philosophy.

2 Reflective equilibrium in practice

2.1 What is reflective equilibrium?

Reflective equilibrium serves as a standard method in philosophy for justifying judgements about what the world is, what the world should be, or what people (should) do. Roughly, reflective equilibrium is the end state of the deliberative process by which we revise our initial beliefs—i.e., initial judgments, or more simply intuitions—about a targeted subject such as justice or knowledge. While the characteristics of these different judgments differ in some important ways, they commonly possess non-inferential warrants for the claims of the targeted subjects.[1] In the context of this study, we can consider all these opinions as simply "intuitions", where intuitions are understood to contribute to the coherent system of judgments in regard to a targeted subject.

Intuitions can thus be viewed as initial input to the end state of the deliberative process in reflective equilibrium. This is illustrated by Scanlon (2003, pp. 140–141), who describes reflective equilibrium as a three-stage process. In the first stage, a relevant set of intuitive judgments is identified. The second stage is to formulate or select theoretical principles that would systematically account for these judgments. In the third stage, any mismatches between the principles and the intuitive judgments are resolved by working back and forth between them.

There is an additional point worth noting: In the three-stage process of reflective equilibrium, the theoretical principles account for more than simply psychological facts or mechanisms. Rather, they are meant to cover the *content* of the intuitions, and the method is supposed to justify principles with the same kind of content. This element of reflective equilibrium requires credible adjustments in the equilibration process in a way that does justice to epistemic desiderata such as parsimony and generality (Brun, 2014, pp. 241–242; Baumberger & Brun, 2021, pp. 7933–7935). This treatment of intuitions suggests that the three-stage process goes beyond the mere achievement of coherence among intuitions as initial input: the adjustment process is required to meet a relevant set of the epistemic desiderata.[2] After all, the point of reflective equilibrium is the reconstructive process of justifying theoretical principles by adjusting intuitive judgments (Baumberger & Brun, 2021, pp. 7935–7938).

[1] Here we treat intuitions as warranted data. Strictly speaking, warrants and data can be distinguished. According to Toulmin (2003, chap. 3), warrants (often implicitly) guarantee the underlying data in one of the claims leading to a conclusion. Because of this, we can treat intuitions (as initial input) as a premise for the relevant inference. For simplicity, we presume the relevance of treating intuitions in the proposed manner.

[2] Some philosophers, who are skeptical of the philosophical practice that relies on intuitions as initial input for justifying theoretical principles, challenge this treatment because intuitions are mere mental states arising non-inferentially (Cappelen, 2012; Deutsch, 2015). While we do not directly defend the proposed treatment of intuitions against this challenge, we provide an (indirect) argument for that treatment by demonstrating the relevance of treating intuitions as such based on the AIC, a goodness-of-fit and parsimonious measure of data, later in this text. For a more direct argument for the treatment of intuitions as evidence in political philosophy, see Conte (2022).

2.2 Two concerns about the use of intuitions in political philosophy

As is well-known, reflective equilibrium is influential in political philosophy. We can partly attribute this to the impact of John Rawls, who was the first to apply this method for the purpose of justifying his two principles of justice. Perhaps even more importantly, as a main subject of political philosophy, justice is distinctly normative—that is, justice has a guiding force that directs people in some collective and compulsory manner. This echoes Rawls's (1971, pp. 5, 48–49) statement about the notion of reflective equilibrium: "it is a notion characteristic of the study of principles which govern actions shaped by self-examination" which denotes the rules that "determine a proper balance between competing claims to the advantages of social life".[3] As such, political philosophers tend to regard our intuitions concerning justice (and morality) as being reflected in the formation of justified judgments about what we ought to do (Copp, 2012; Knight, 2017; Rawls, 1971; Scanlon, 2003).

In political philosophy, two main concerns have been raised with regard to the method of reflective equilibrium. First, some philosophers have expressed concern about manipulating intuitions as initial input to support principles of justice. The intuitions must be *folk* intuitions rather than those of political philosophers. Otherwise, intuitions cannot play a non-circular role in justifying the principles. In the practice of reflective equilibrium, philosophers examine theoretical principles to see whether they are systematically consistent with our intuitive judgments. In political philosophy, however, appeals to intuitions are based on *anticipated* folk intuitions, so that political philosophers take *their own* intuitions to reflect the intuitions of ordinary people. This is often seen in theories of distributive justice; distributive theorists presume that their anticipated intuitions can be equated with people's reactions to the states of affairs that the theories evaluate in terms of the goodness and/or badness of the states of affairs.[4] This raises a question as to whether the theories are tested reflectively in light of folk intuitions: The anticipated intuitions may be merely those of the distributive theorists themselves. This may well motivate the skeptics of reflective equilibrium to suspect that the anticipated intuitions are "rigged up" to countenance their proposed theories in a circular manner with respect to full justification (Brandt, 1979, 1990; Hare, 1973; Singer, 1974, 2005).[5]

[3] This claim may be interpreted such that, based on Rawls's emphasis on self-examination, reflective equilibrium is pursued by *each* individual through the process of deliberating between intuitions and theoretical principles. This may be a reasonable interpretation. Rawls (1971, pp. 48–51) put stress on the role of intuitions made by a particular person who is willing to make a correct decision. However, as argued above, political philosophy requires *collective* and *normative* deliberation. We thus should attempt to merge each person's judgments into a converged state that would fully support the principle of justice and morality. For this point, see Baderin (2017), Daniels (1996), and Scanlon (2003).

[4] There are remarkable exceptions (on this point see Hassoun 2016). In the context of the present analysis, several important experimental studies are relevant (e.g., Bruner, 2018; Bruner and Lindauer, 2020; Inoue et al., 2021; Pölzler and Hannikainen, 2022). Later, we will discuss our own experimental study that advances these earlier lines of research into the manifestation of folk intuitions.

[5] The circularity objection should not be overestimated for two reasons. First, we can interpret reflective equilibrium in some (weakly) foundationalist way so that even when initial intuitive judgments are not infallible, they can enjoy some degree of initial credibility for the justification of theoretical principles (Ebertz, 1993; Holmgren, 1987; McMahan, 2000; Pust, 2000; for the compatibility between reflective

Second, there have been doubts about the (legitimate) use of thought experiments that purport to elicit intuitions as initial input in ethics and political philosophy. Anscombe's (1957) criticism of the use of thought experiments is the classic example: Thought experiments treat morally serious issues in such a flippant way as to dismiss the richness of philosophical discussions about normatively significant and sensitive issues involving people's life and death. The richness in question may have bearings on a key feature of moral principles: They apply to the practical context in which real people confront dilemmatic issues and problems involving various factors concerning just institutions (Goodin, 1982) and moral dilemmas and vicissitudes in real life (O'Neil, 1989). In other words, there may be non-negligible gaps between abstract or purely hypothetical—often modally bizarre—cases (such as Nozick's utility monster), and actual practical cases. Our intuitions prompted by the former, but not those prompted by the latter, are unreliable as reflective warrants for or against theoretical principles in political philosophy. Although this is not a direct challenge against any appeal to intuitions in reflective equilibrium, it does pose a fundamental problem with the use of the method in question; political philosophers often employ thought experiments including those of a purely hypothetical kind as possible cases. Since their arguments rely on folk intuitions about how to respond to such cases, and since intuitions as initial input are a starting point for reflective equilibrium, proponents of reflective equilibrium must respond to this challenge.

We can defuse the two concerns, however. In response to the first concern, the key point is that the intuitions collected must be folk intuitions, not the intuitions of the political philosophers themselves, in order to avoid the charge of manipulating the intuitions to uphold the principles of justice. In response to the second concern, philosophers have to provide a convincing method for the use of thought experiments in political philosophy. Let us explain this by examining a proposal for the proper use of thought experiments in political and moral philosophy. According to Walsh (2011, pp. 478–480), we can conduct thought experiments properly in light of the distinction between their legitimate and illegitimate uses. The illegitimate use of thought experiments is problematic because it ignores the richness of contexts in which the issues and problems arise, such that thought experiments are naïvely used to show the plausibility of theoretical principles in *all logically possible worlds*. Many (if not all) bizarre and purely hypothetical cases are meant to accommodate possible worlds far removed from reality (even if nomologically relevant), and it is this accommodation that skeptics question. However, this does not lead us to repudiate any appeal to possible cases. A use of thought experiments is legitimate if it caters to "the *contingency* of the problems with which applied ethicists characteristically deal" and does not try to "draw conclusions that attempt to accommodate a wide range of merely possible cases rather than the actual case before us" (Walsh, 2011, p. 478; emphasis original). If thought experiments are legitimately used, we may respond to the context-based challenge against the use of thought experiments.

Footnote 5 continued
equilibrium and weak foundationalism, see BonJour, 1985, pp. 26–30; Cath, 2016, pp. 218–220). Second, even without endorsing such (weak) foundationalism, we can reasonably claim that reflective equilibrium is not circular, because the practice of reflective equilibrium involves other considerations, e.g., systematicity.

2.3 Survey experiments and model selection

Up to this point, we agree with Walsh's argument. However, it is not clear how we can legitimately draw on thought experiments in practice. In what follows, we suggest a way to ensure the relevance of appealing to possible cases: Possible cases can be treated in such a way as to satisfy the context-sensitivity of the issues and problems if we conduct *survey experiments* in which we analyze the results *with a proper model selection.*

Clearly, this proposal draws on folk intuitions, because the subjects of survey experiments are ordinary people. The survey experiments aim to ensure the sample size required for quantitative analysis and to facilitate the acquisition of a sample size representative of the population. Hence, the use of survey experiments can respond to the first challenge against the method of reflective equilibrium. As a first approximation, this proposal seems promising too because people's intuitions obtained as the data through survey experiments may well reflect the contextual interactions of relevant factors and vicissitudes of life. This should help to establish a good start at the first stage of reflective equilibrium, and may well allay the concerns of skeptics about the use of thought experiments.

Nevertheless, the mere use of survey experiments is not sufficient for the legitimate use of possible cases. As a second approximation, we suggest the use of a proper model selection by means of the Akaike Information Criterion (AIC). Before exploring this point, let us see how difficult it is to single out particular cases relevant to an issue (such as abortion) in an *ex ante* manner. There are two problems at this point. First, apparently relevant cases often have disanalogies to the issue under consideration that are difficult to discern in advance. This renders the (apparently) intuitive fit with theoretical principles worthless. Second, apparently irrelevant cases could be of the type that steer our intuitions in certain directions. We may doubt that the cases at issue are legitimately excluded and thus reach an unconvincing verdict about the proper (un)fit between our intuitions and the proposed theoretical principles. As long as we cling to the method of cases, we must have a criterion for sorting out possible cases in an *ex ante* manner.

Can we establish such a criterion in an *ex ante* manner? We doubt it, because we can easily point out illegitimate inclusions and exclusions of possible cases if we carefully look through each particular case. We can raise famous examples of philosophical arguments relating to illegitimate inclusions and exclusions. Thomson's (1971) violinist may be seen as an example of illegitimate inclusions: There might be disanalogies between unplugging an individual from the famous unconscious violinist and allowing the abortion of pregnant women who were raped (Davis, 1983). Foot's (original) trolley problem has been questioned as an example of legitimate exclusions in order to support the killing-and-letting-die principle: The other cases as variants of the trolley problem cannot be covered by the killing-and-letting-die principle, such as the case where the trolley driver has just died and a passenger must decide whether to turn the trolley around (Thomson, 1976, 1985). To avoid misunderstandings, we do not underestimate the significance of the philosophical discussion over case-based explorations such as Thomson's violinist and Foot's trolly problem. Nor do we deny the possibility of establishing an *ex ante* criterion for sorting out possible cases in a relevant manner. We only claim that there are difficulties in establishing the proper criterion in an *ex ante* manner, given these famous examples and arguments, and that it may be feasible to have a different manner of dealing with possible cases. (Of course, ours is not the only pertinent way to handle possible cases.)

Our suggestion is as follows: We should use a model selection method for coping with possible cases in thought experiments in an *ex post* manner. That is, we propose to use AIC-based model selection as a practical method for reconciling intuitions with theoretical principles in a systematic manner, ensuring that the epistemic desiderata, particularly parsimony (simplicity), are honored in the practice of reflective equilibrium. While this method does not directly search for the relevant similarities of possible cases, it leads to the justification of a targeted principle by virtue of the systematic adjustments of intuitions that possible cases evoke; satisfying the epistemic desiderata of generality and (especially) parsimony would guarantee the legitimate use of possible cases, even if they may include irrelevant cases. Obviously, this method reflects the virtue of reflective equilibrium. Let us elaborate on this point in more detail below.

To begin with, let us explain why we suggest the use of AIC. While there are criteria that differ quantitatively from AIC (such as the Bayesian Information Criterion (BIC)), AIC is simply defined and can be seen as generalizable in a perspicuous manner (Forster & Sober, 1994, p. 2). Indeed, AIC is a widely used method for evaluating how well a model befits the obtained data. Roughly, AIC is calculated by the number of independent variables for constructing the model and by the maximal likelihood estimate of the model (i.e., the higher the likelihood of a model with few independent variables yielding the data, the better the model). According to AIC, the best model has the greatest predictive ability measured by estimated likelihood (P (data | model)). AIC aims to achieve the maximum degree of data fit by incorporating a minimal number of independent variables, in keeping with the condition of parsimony as a theoretical virtue for reflective equilibrium practice.

We can now state the philosophy underlying model selection as follows: Although multiple models are always maintained, they can be compared and ranked according to specific criteria and based on data. Notably, this is different from the Neyman–Pearson philosophy based on frequentism, which is a theory about which hypothesis should be accepted as true or rejected as false based on existing data. But why is AIC better than the other criteria in our argument?[6] To see this, let us focus on the comparison of AIC with BIC. While empirical studies often recommend competing models based on both criteria, our reason for choosing AIC over BIC is that the former measures the predictive accuracy of a model based on existing data, without the specific information that certain empirical observations carry (Sober, 2002, 2008; see Otsuka, 2021, p. 55). BIC measures the likelihood (posterior probability) of a model relative to existing data. Importantly, BIC does not necessarily follow the principle of Occam's razor: the simpler the model, the better. By contrast, AIC recommends a model with higher predictive accuracy for future data, which plausibly favors simplicity. Thus, AIC-based model selection can be used as an *ex post* way of dealing with possible cases, which incorporates the epistemic virtue of parsimony in practicing reflective equilibrium.

[6] We owe this important question to an anonymous reviewer.

Let us explain this point in more detail. According to AIC, we can comparatively evaluate how well each theoretical principle fits with the data obtained from survey experiments. There are two advantages of using AIC in this way. First, this method takes into account the limited availability of relevant cases that persist in survey experiments. AIC is used for the estimation of a model's predictive performance within the confinement of the available data. Second, the statistical model selection can be viewed as a reasonable estimate of the maximally relevant set of independent variables that determine the predictive performance of a targeted theoretical principle. Importantly, from a reflective equilibrium perspective, the estimated independent variables coupled with the principle single out the significant features of the principle that pertain to people's intuitions prompted by possible cases, whether relevant or irrelevant. More concretely, due to the emphasis on parsimony, the principles and parameters will not fit to every intuition regarding every case. We can thus hope that intuitions that are misled due to problematic cases are effectively not taken into account. Rather, the final model concentrates on a relatively small set of principles and parameters that capture the intuitive reactions of people overall well. In this way, the AIC-based model selection allows us to sidestep significant challenges with an *ex ante* case selection, dispensing with illegitimate inclusions and exclusions of relevant possible cases. The AIC-based model selection can be seen as a kind of *ex post* case selection.[7]

We can now say that our proposal serves as the three-stage process of reflective equilibrium in which principles (i.e., models in this context) are adjusted based on intuitions that respond to possible cases in thought experiments. Intuitions as a starting point are input commitments for building or selecting a relevant principle. This is the first stage of reflective equilibrium, and it is carried out using survey experiments. The process of model selection can be seen as achieving the second stage of reflective equilibrium, in which we check whether the principle can systematically account for the intuitions. This is because its epistemic goal is to obtain the best and most parsimonious fit between a model and the data obtained. Since we can grasp the intuitions as the relevant data from survey experiments, a theoretical principle that would pass the AIC test can reasonably be regarded as the best—or at very least a better—model. More moderately, we can view a model that shows a bad (worse) AIC score as a less plausible model (compared to one which has a better score). For this reason, we can consider this use of AIC as a formal and feasible method to simplify the factors required in the second stage of reflective equilibrium.

Note that AIC model selection does not itself cover the third stage of reflective equilibrium: that any possible systematic disparity between our intuitions and the principle is resolved in such a way as to work back and forth between them.[8] In AIC model selection, a model is selected simply based on its high predictive performance

[7] Note that AIC-based model selection does not serve as a standard of discarding counterintuitive cases that go against certain theoretical principles. AIC is used to assess how well theoretical principles (as statistical models) parsimoniously fit intuitions (as data) by checking whether the coefficients of the variables in the data analysis are significant. This approach allows us to avoid the enormously difficult task of discarding irrelevant cases *ex ante*. We thus refer to this as "a kind of" *ex post* case selection, since, as argued above, the use of AIC does not involve the selection of relevant cases. We appreciate the editor(s) for recognizing the importance of this point.

[8] This remark is owed to the editor(s).

for future data in a parsimonious manner. As shown above, this can be seen as the second stage of reflective equilibrium, in which the principle of justice is selected that would systematically account for the intuitions that are invoked by possible cases in thought experiments. However, this does not itself involve any change of some existing intuitive judgments that would be an expected result of going back and forth between principles and intuitions.[9] In our argument, what would be involved in this third stage of reflective equilibrium? Our answer is that the third stage is outside the statistical analysis in our argument: Any modification of certain intuitions should be done *separately* in light of the results of the survey experiment and the AIC model selection. This separate process can be better illustrated through the use of a test case, which is one of the tasks in the upcoming sections. The results of our test-case analysis will be presented in Sect. 4.3.

3 Testing the theories of distributive justice

In this and the next sections, we highlight the practicality and significance of the proposed practice of reflective equilibrium by referring to the debates over theories of distributive justice, in particular the debate between prioritarianism and sufficientarianism. Using a survey experiment and a model selection method, we show that the sufficienciantarian principle cannot be evaluated as a better theoretical principle than the prioritarian principle. This will serve to illustrate how the proposed method can be exercised as reflective equilibrium in practice. First, our method allows us to examine whether folk intuitions indeed fit well with sufficientarianism, such that many political philosophers would intuitively support the indisputability of a minimal threshold. Second, we may then consider the modification of some intuitions in light of the results of the statistical analysis, which is an important part of working back and forth between principles and intuitions (i.e., the third stage of reflective equilibrium).

3.1 Egalitarianism, prioritarianism, and sufficientarianism

Let us first introduce popular theories of distributive justice. Egalitarianism is certainly the best-known of these theories. Although egalitarianism has variants in terms of people getting the same, being treated the same, or being treated as equals (Arneson, 2013), egalitarianism as defined here is simply concerned with people being equally well-off. According to egalitarianism, it is intrinsically bad if some people are worse off than others.[10] Many (if not all) political philosophers argue that endorsing the

[9] Note that any change of certain intuitions is part of the reflective equilibrating process. The discrepancies may well lead us to modify principles in light of the intuitions. This is because the intuitions may be an important source of new principles that provide more reasonable coverage of all relevant intuitions in a parsimonious way.

[10] Egalitarianism here is taken as a telic—or more precisely, axiological—form in that it is bad *in itself* if some people are worse off than others. This formulation of egalitarianism is based on the state of affairs and serves as a sufficient basis for our experimental study to comparatively evaluate the two theories of distributive justice. For discussion of this point, see Hirose (2015, chap. 3), Lippert-Rasmussen (2007), Parfit (2000, pp. 84–88), and Segall (2016, pp. 10–15).

badness of distributive inequalities *simpliciter* is unreasonable because it is objectionable to claim the intrinsic value of eliminating distributive inequalities by radically reducing the overall welfare of all people (Holtug, 1998; Parfit, 2000; Temkin, 2000). The so-called "leveling down objection" encourages many political philosophers to suggest two different theories: *prioritarianism* and *sufficientarianism*. Prioritarians assert that gains in well-being are more valuable, the worse off the person would otherwise be (Arneson, 2022; Hirose, 2015, chap. 4; Holtug, 2007, 2010, chap. 8; Parfit, 2000, pp. 101–106). According to sufficientarianism, whether a person has enough of some goods matters rather than being concerned with inequalities as such (Frankfurt, 1987; Gosseries, 2011; Hirose, 2015, chap. 5; Shields, 2020). These two theories of distributive justice have been seen as attractive alternatives to egalitarianism.

The appeal of the theories has been strengthened by respective refinements. In particular, sufficientarianism has been elaborated in an alluring manner. A refined version of sufficientarianism incorporates two "enough" thresholds, the *minimal* and *maximal* thresholds. The minimal threshold is the point where basic needs are met, whereas people above the maximal threshold have good (content) lives in terms of healthy and cultured living (Huseby, 2010, 2017). According to refined sufficientarianism, welfare shortfalls below the minimal threshold are simply (non-gradually) morally bad; welfare shortfalls between the two thresholds become (gradually) worse as their number and sizes increase (Huseby, 2017, p. 74). Refined sufficientarianism powerfully embraces the intuitive aspects of egalitarianism and prioritarianism. It endorses the complex evaluations of inequalities, in that it is not concerned with the badness of distributive inequalities *simpliciter*, but rather with people's worse-off positions below the threshold(s). As such, the refined sufficientarian approach to distributive justice has gained popularity in political philosophy.[11]

3.2 The method of cases and reflective equilibrium in practice: the example of the theories of distributive justice

Our interest lies in how sufficientarians attempt to compete with the prioritarian principle. As many political philosophers do, sufficientarians have appealed to people's intuitions, but exactly how? There are two ways of appealing to intuitions. First, sufficientarians can point to the popularity of the maximizing principle with an income floor rather than Rawls's difference principle—a prioritarian principle in the not-strict sense[12]—among ordinary people. The popularity of moral principles restricted with a sufficientarian threshold was shown first by Frohlich and Oppenheimer's (1992, pp. 58–60) laboratory experiments and later replicated in other studies (Bruner, 2018;

[11] This does not mean that the refined sufficientarianism has not been subject to criticism. For one such criticism, see Segall (2016, chap. 5).

[12] The difference principle contends that inequalities are justified only when and because they maximize the expectations of the worst off. Although this principle appears to be similar to prioritarianism, it differs from prioritarianism in two important ways. First, the difference principle gives priority to the worst off in an absolute manner (i.e., maximin and leximin) rather than giving additional weight to their interests. Second, it is not concerned with how we should weight the interests of people in the case that such weighting would not affect the worst (worse) off. About these points, see Rabinowicz (2002, p. 13) and Hirose (2015, pp. 95–98).

Bruner & Lindauer, 2020; Lissowski et al., 1991; cf. Inoue et al., 2021). However, this appeal to people's intuitions is not (explicitly) employed by political philosophers, because they have recourse to the method of cases by illustrating the (im)plausibility of competing theoretical principles. This is the second way of appealing to people's intuitions.

To illustrate: Sufficientarians raise the so-called "Beverly Hills case" in order to show the plausibility of their sufficientarian proposal (and the untenability of prioritarianism). The Beverly Hills case is as follows: Suppose we must choose between benefiting the rich and benefiting the super-rich. While many ordinary people would intuitively not prioritize the rich in this case, the prioritarian position defies that intuition, espousing a policy of always benefitting the rich rather than the super-rich simply because the rich are worse off than the super-rich (Benbaji, 2006; Crisp, 2003). By contrast, as mentioned above, refined sufficientarianism appeals to people's intuition in that the different thresholds are germane to the differential degree of moral urgency assigned to the states of affairs involving the thresholds. On this basis, Huseby (2010, p. 183) contends that sufficientarianism (with its use of a maximal threshold above which people should have content lives) can respond to Holtug's (2007, pp. 149–150) case—which can be dubbed "the Left-Behind case"—against simple sufficientarianism, i.e., sufficientarianism with only one threshold: Only one individual at the threshold level is left behind in the boom of the world economy where everyone else enjoys much better-off positions than hers. Huseby (2010, p. 183) believes that while "[t]he relative deprivation of the person left behind in Holtug's scenario, makes it very hard for her to be content in an environment in which she is considerably worse off than all others", the maximal threshold of sufficientarianism would license her claim for "a level of welfare at which she would be content". As such, sufficientarians use the method of cases against the prioritarian principle by appealing to people's intuitive responses to the states that the theoretical principles (do not) endorse.

However, as argued in the previous section, it seems reasonable to ask whether ordinary people would truly find no plausibility in the proposed theoretical principles in possible cases. Nor can we ensure that the cases in question involve neither illegitimate inclusions nor illegitimate exclusions; there might be some disanalogies or a result of snubbing relevant cases. These concerns can reasonably be defused if we adopt the method of reflective equilibrium in practice. More specifically, we can conduct a survey experiment using apparently relevant cases (the first stage of reflective equilibrium) and then analyze people's intuitive responses to the possible cases using AIC (the second stage of reflective equilibrium). We can then compare theoretical principles—here the prioritarian principle and the refined sufficientarian principle—to see which principle better fits the data obtained from a survey experiment; AIC *ex post* indicates which of the two principles better fits with people's intuitions. In other words, we do not need to select possible cases before investigating the intuitive judgments. The third stage of reflective equilibrium involves attempts to resolve any inconsistencies between the selected principle and certain existing intuitions. In this context, the modification of some intuitions supporting, for example, sufficientarianism may be considered in light of the results of the survey experiment and the AIC model selection. While, here again, any such modifications must be conducted separately from the statistical analysis, they nonetheless play an important role in our approach, and distinguish it from the method of cases.

Let us further note the relevance of reflective equilibrium in practice to the debate over the two theories of distributive justice. In light of people's intuitions about cases of an apparently relevant sort, we will require a sophisticated analysis of the distributive theories. As a test case to clarify the significance of our proposal of reflective equilibrium in practice, we will attempt to compare the refined sufficientarian principle that incorporates the two thresholds with the prioritarian principle. Specifically, we will investigate how sensitive ordinary people are to distributive inequalities in the presence of minimal and maximal thresholds. We can thereby evaluate the states of affairs involving the different types of inequalities and worse-off positions below or above the two threshold(s). From the viewpoint of reflective equilibrium in practice, it is important to examine (i) whether the state of equality is more supportable than unequal states, (ii) whether ordinary people tend to prioritize the worse off in apparently relevant cases (including the Beverly Hills case and the Left-Behind case), and (iii) whether the multiple thresholds concern the ordinary judgments in such cases. We can then compare prioritarianism with refined sufficientarianism in terms of whether they each systematically befit people's intuitive judgments. Finally, we can consider modifying some intuitions related to the principles when one of the principles (models) is selected on the AIC.

4 Experiment

The aim of our experiment is fourfold. First, we want to find out whether ordinary people are generally egalitarian or not. Second, we investigate whether ordinary people react significantly to distributive inequalities in a variety of apparently relevant cases (including the Beverly Hills case and the Left-Behind case). Third, we examine which of the prioritarian principle and the refined sufficientarian principle fits better with systematically captured intuitions, presented with the two thresholds, based on the model selection method.[13] Fourth, we will consider modifying or eliminating certain intuitions in light of the selected principle. For this purpose, we conduct a survey experiment that focuses on how ordinary individuals react to distributive cases of an apparently relevant kind.

[13] One might claim that this experimental setting cannot be useful for comparing prioritarianism with sufficientarianism, because it explicitly presents the two thresholds, the minimal and maximal thresholds. Admittedly, folk intuitions may be disturbed by the presence of the two thresholds, which might distort people's preferences. However, in our argument, this experiment plays a role only as a *test case* for the exercise of reflective equilibrium: Whether the prioritarian principle and/or the refined sufficientarian principle truly match prioritarianism and sufficientarianism, respectively, does not strictly matter to our argument. They are nothing more or less than a test case for illustrating the significance of the model selection method as the method of reflective equilibrium in practice.

4.1 Method

4.1.1 Participants

A private research company (Rakuten Insight, Inc.) was asked to recruit respondents for our online experiment. These respondents had voluntarily applied to the research company to participate in experiments from their homes by answering questions via the Internet. The instructions were presented on their computer. After the experiment, the company randomly chose some of the respondents and paid them a fee of 500 yen (approximately US$5). The experiment took place from March 23rd to 29th 2022, with 2,707 subjects (1,352 females, 1,344 males, and NA 11). The age distribution was 12 respondents in their teens, 397 in their 20s, 398 in their 30s, 520 in their 40s, 471 in their 50s, 450 in their 60s, 455 in their 70s, and 4 in their 80s. Our sample roughly corresponds to the age and gender distribution of the actual population in Japan.

4.1.2 Design and materials

We constructed ten cases based on a between-subject design in which respondents were randomly assigned to each of the ten cases. Each case was described as a figure showing two distributive states of affairs that the respondents were requested to evaluate comparatively.[14] Four features were common to all cases. The first of these features was that each state had two persons, x and y. Second, the bar heights indicated the levels of each person's income. Third, the first distributive state had an unequal distribution of income (person x was better off than person y), whereas the two persons enjoyed equal income in the second state. In both states, the sum of income is the same. Fourth, dashed lines were drawn to represent the two thresholds (maximal threshold: 4 million yen per year; minimal: 2.5 million yen per year) in every case. An income of 4 million yen was chosen as the maximal threshold because this is the average annual income in Japan. This can reasonably be seen as a threshold above which people can lead healthy and cultured lives. An income of 2.5 million yen was considered the minimal threshold because this is the approximate income qualifying for public assistance in Japan. This can reasonably be regarded as a threshold where the basic needs of people are met. The difference among the ten cases thus boiled down to whether each income was above or below the minimal and/or maximal thresholds.

The ten cases cover all potentially relevant differences. In Case 3, for example, the first (unequal) state (Society A) has person x, whose income is between the minimal and maximal thresholds, and person y, whose income is below the minimal threshold, while persons x and y have incomes between the minimal and maximal thresholds in the second (equal) state (Society B). The following figure was used in Case 3 (Fig. 1).

[14] For all ten cases with figures, see the Appendix.

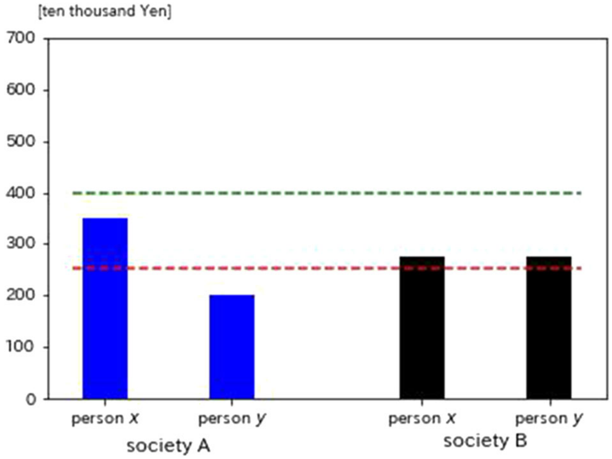


Fig. 1 Unequal society A vs equal society B in Case 3

The ten cases can be described in terms of the two thresholds, such that:

Case 1: Society A (Lx > Ly), Society B (Lx = Ly).
Case 2: Society A (Mx > Ly), Society B (Lx = Ly).
Case 3: Society A (Mx > Ly), Society B (Mx = My).
Case 4: Society A (Mx > My), Society B (Mx = My).
Case 5: Society A (Ux > Ly), Society B (Lx = Ly).
Case 6: Society A (Ux > Ly), Society B (Mx = My).
Case 7: Society A (Ux > My), Society B (Mx = My).
Case 8 (the Left-Behind case): Society A (Ux > Ly), Society B (Ux = Uy).
Case 9: Society A (Ux > My), Society B (Ux = Uy).
Case 10 (the Beverly Hills case): Society A (Ux > Uy), Society B (Ux = Uy).

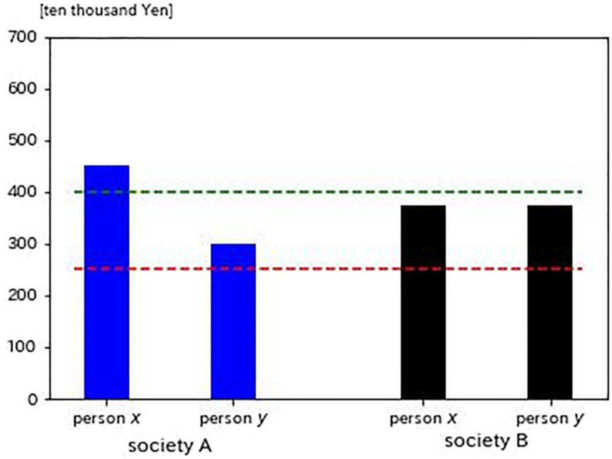
Note: L means a level of income below the minimal threshold. M means a level of income between the minimal and maximal thresholds. U means a level of income above the maximal threshold.

4.1.3 Procedure

Respondents completed the questions online, in their own time. Before beginning, they read a consent form and were assured of the anonymity of their data. After granting consent, they were presented with a written scenario and a figure (Case 7 is shown below as an example) and were asked to respond to a question:

> *The following figure shows two societies where two persons, x and y, can gain different levels of income. The blue bars indicate the levels of income (unit: yen) that x and y will get when society A is realized, whereas the black bars indicate the levels of income that x and y will get when society B is realized.*
> *Moreover, the green dashed line represents enough income for one individual to lead a healthy and cultured life. The red dashed line represents enough income for one individual to lead a barely healthy and cultured life.*

In this case, which set of incomes do you prefer, the blue bars or the black bars, and how strong is your preference? Please choose the option most close to your view.



(1) *Blue is strongly preferable.*
(2) *Blue is preferable.*
(3) *Blue is slightly preferable.*
(4) *Both are on par.*
(5) *Black is slightly preferable.*
(6) *Black is preferable.*
(7) *Black is strongly preferable.*

This question is intended to capture folk intuitive reactions to distributive inequalities in the presence of the two thresholds. Their reactions to the presence of persons below and above each threshold will also be revealed through their answers to this question. We can reasonably expect that the results will elucidate how people's intuitions are manifested in the face of different states.

4.2 Results

As Fig. 2 shows, respondents showed a general tendency to prefer Society B (an equal society) to Society A (an unequal society). However, Society A was preferred in some cases. Interestingly, there was a difference in people's preferences between the Left-Behind case and the Beverly Hills case: In the former, more people preferred Society B over Society A, whereas Society A was more often preferred to Society B in the latter. That is, people preferred a state in which no one was left behind, but found inequalities above the threshold tolerable. This seems to illustrate that (1) simple egalitarianism may not suit people's intuitions in the apparently relevant cases, and that (2) we cannot claim that people are either prioritarian or refined sufficientarian in light of the descriptive statistics based on the two cases; neither the preferences of

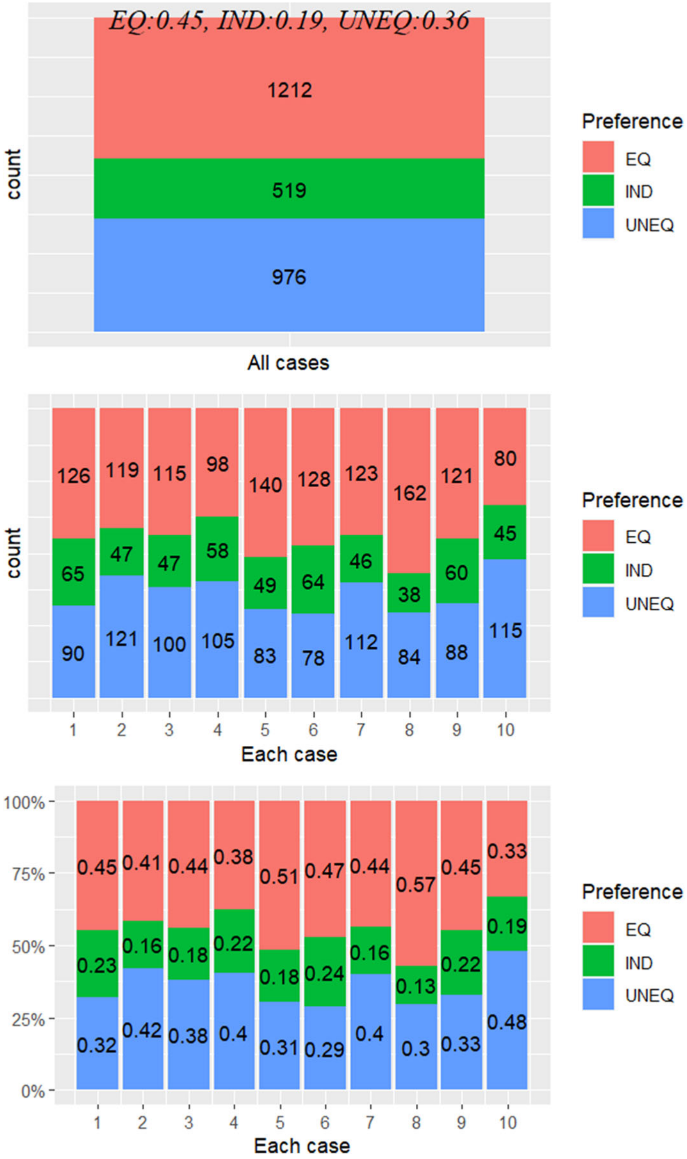


Fig. 2 People's preferences in regard to the two distributive states in ten cases **Note*: EQ is the number and ratio of respondents who chose answers (5), (6) or (7). IND is the number and ratio of respondents who chose answer (4). UNEQ is the number and ratio of respondents who chose answers (1), (2) or (3)

prioritarians nor those of refined sufficientarians consistently matched the preferences of the respondents in these two cases. This, we believe, supports the use of a model selection method to examine which of these principles befits intuitions systematically captured in the apparently relevant cases.

Let us apply the model selection method for the comparative evaluation of prioritarian and sufficientarian principles (statistical models) in terms of the extent to which they fit well with people's intuitions in the apparently relevant cases. First, we suggest two models, **P1** and **S1**, both of which contain all independent variables and control variables.

P1 consists of the three important independent variables (for details, see the note in Table 1): *poorLevel* (the variable reflecting the income level of the poor person *y* in Society A), *richLevel* (the variable reflecting the income level of the rich person *x* in Society A) and *middleLevel* (the variable reflecting the income level of persons *x* and *y* in Society B). We also control for participants' demographic characteristics in the model to ensure accurate and unbiased estimation of the important independent variables (see the note in Table 1). The negative value of the coefficient *poorLevel* ($\text{Coef}_{\text{poor}} = –0.33$; $p < 0.001$) indicates that if the income level of a poor person in

Table 1 Coefficients of prioritarian model 1 (**P1**)

The dependent variable is a preference for the equal society B	**P1**		
Predictors	*Log-Odds*	*Std. Error*	*P value*
poorLevel	−0.33	0.06	**< 0.001**
richLevel	0.08	0.06	0.189
middleLevel	0.13	0.06	**0.031**
Groupsize	−0.02	0.01	0.058
Numeracy	0.06	0.03	0.063
Female	0.5	0.07	**< 0.001**
Age	0.12	0.02	**< 0.001**
Income	−0.01	0.01	0.422
Observations	2664		
Adjusted R^2	0.115		
AIC	9595.639		

Note: poorLevel is an ordered variable that was set at 3 if the income level of the poor person *y* in Society A was above the maximal threshold, set at 2 if their income level was between the minimal and maximal thresholds, and set at 1 if their income level was below the minimal threshold. *richLevel* is an ordered variable that was set at 3 if the income level of the rich person *y* in Society A was above the maximal threshold, set at 2 if his/her income level was between the minimal and maximal thresholds, and set at 1 if his/her income level was below the minimal threshold. *middleLevel* is an ordered variable that was set at 3 if the income levels of persons *x* and *y* in Society B were above the maximal threshold, 2 if their income levels were between the minimal and maximal thresholds, and 1 if their income levels were below the minimal threshold. Coefficients are estimated by the ordered logit model. We control for participants' demographic characteristics (gender, age, income, numeracy, and size of groups where participants work or study in their daily lives). The bold numbers in the table highlight those with *p*-values less than 5%, in accordance with the conventions of quantitative analysis

Society A increased from below the minimal threshold to above the maximal threshold, it would very likely cause respondents to change their preference from Society B to Society A. The positive value of the coefficient of *middleLevel* ($\text{Coef}_{\text{middle}} = 0.13; p = 0.031$) means that if the income level of average persons in Society B increased from below the minimal threshold to above the maximal threshold, it would very likely cause respondents to change their preference from Society A to Society B. Thus, comparing the values of these coefficients, the poor person has more impact than the person with average income. As we see it, **P1** approximately represents the prioritarian principle such that **P1** echoes intuitions changed in an egalitarian direction especially if we attended to the level of the poor, and also those changed (slightly weakly) in an egalitarian direction if we attended to the level of people in equal Society B.

As Table 2 shows, **S1** is composed of four important variables. The first two are *DifMiserableLine* (the dummy variable that was coded 1 if a distributive inequality between the rich person x and the poor person y existed across the minimal threshold in Society A, and otherwise 0) and *DifSufficientLine* (the dummy variable that was coded 1 if a distributive inequality between the rich person$x$ and the poor person$y$ existed across the maximal threshold in Society A, and otherwise 0). The second two are *numDemiserablized* (the variable representing the net number of persons who would move across the minimal threshold if unequal Society A were changed to equal Society B) and *numSufficiencialized* (the variable representing the net number of persons who would move across the maximal threshold if unequal Society A were changed into equal Society B). As can be seen, the two thresholds affected the intuitive judgments of

Table 2 Coefficients of sufficientarian model 1 (**S1**)

The dependent variable is a preference for the equal society B	**S1**		
Predictors	*Log-Odds*	*Std. Error*	*P value*
DifMiserableLine	0.12	0.07	0.096
DifSufficientLine	0.29	0.07	**< 0.001**
numDemiserablized	0.06	0.05	0.267
numSufficiencialized	0.08	0.05	0.132
Groupsize	−0.02	0.01	**0.047**
Numeracy	0.06	0.03	0.058
Female	0.5	0.07	**< 0.001**
Age	0.11	0.02	**< 0.001**
Income	−0.01	0.01	0.454
Observations	2664		
Adjusted R^2	0.113		
AIC	9599.122		

Note: We control for participants' demographic characteristics (gender, age, income, numeracy, and size of groups where participants work or study in their daily lives). The bold numbers in the table highlight those with *p*-values less than 5%, in accordance with the conventions of quantitative analysis respondents. We also control for participants' demographic characteristics in the model to ensure accurate and unbiased estimation of the important independent variables (see the note in Table 2). Regarding *DifMiserableLine*, respondents tended to marginally prefer equal Society B to unequal Society A if the distributive inequality existed across the minimal threshold (coefficient odds ratio: $\text{Coef}_{\text{min}} = 0.12$; $p = 0.096$). The threshold sensitivity was confirmed distinctly regarding *DifSufficientLine* (coefficient odds ratio; $\text{Coef}_{\text{max}} = 0.29$; $p < 0.001$). These results show: First, ordinary people attend to the two thresholds, in that they would prefer egalitarian societies when distributive inequalities hold across the two thresholds; second and more importantly, ordinary people are more sensitive to the maximal threshold than the minimal one ($\text{Coef}_{\text{max}} = 0.29 > \text{Coef}_{\text{min}} = 0.12$, $p = 0.096$ and $p < 0.001$, respectively).[15] Thus, we can tentatively say that the statistical results shown in Table 2 barely support the refined sufficientarian principle.

Next, following the standard procedure of model selection, we reconstruct **P2** and **S2** by eliminating insignificant variables. Here, **P1** and **P2** represent the prioritarian principle and **S1** and **S2** the sufficientarian principle. Let us first examine the two prioritarian models.

In tandem with the usual model selection process, we build **P2** because the coefficient of *richLevel* is not significant in **P1**, which would, very likely, indicate the irrelevance of that variable to a prioritarian statistical model. This selection also seems reasonable in the prioritarian theory because, according to prioritarianism, the worse off people are, the more morally important it is to benefit them. With these models in hand, although the AIC of **P2** is almost the same as that of **P1** shown in Table 3, we can regard **P2** as a relevant prioritarian model accommodating the relevant independent variables.

Now let us turn to the two sufficientarian models, **S1** and **S2**.

Under the usual model selection process, **S2** is built based only on the respondents' reaction to the presence or absence of the distributive inequality across the two thresholds. This is because *numDemiserablized* and *numSufficiencialized* are not at all significant in **S1**. Under the sufficientarian theory, any transitional change from Society A to Society B is axiologically irrelevant: We ought to evaluate each state independently and compare them. As Table 4 shows, since the AIC of **S2** is slightly smaller than that of **S1**, we can regard **S2** as a more relevant sufficientarian model in terms of accommodating the relevant independent variables.

We are now in a position to evaluate **P2** (the prioritarian statistical model) and **S2** (the sufficientarian statistical model) in terms of their fit with the observed data that echo intuitions systematically captured in the apparently relevant cases. While the AIC of **P2** is 9595.364, that of **S2** is 9598.407. The smaller the value of AIC, the better fit

[15] There may be two ways of interpreting this result. First, we can interpret the result in the following manner: Ordinary people are more sensitive to whether people enjoy their lives than whether the basic needs of people are met. Particularly in developed countries such as Japan, distributive inequalities across the line of good (content) lives may attract more attention than those across the line of basic needs being met, because the important threshold for ordinary people in developed countries may be the maximal one, since the basic needs of most individuals in developed countries tend to be met. Second, we can interpret the result in ways that ordinary people take the "maximal" threshold in our experiment as the minimal threshold and regard the "minimal" threshold in our experiment as far below the minimal one. To determine which interpretation holds true, however, we need to conduct more experiments

Table 3 Coefficients of prioritarian models 1 and 2 (**P1** and **P2**)

The dependent variable is a preference for the equal society B	**P1**			**P2**		
Predictors	*Log-Odds*	*Std. Error*	*P value*	*Log-Odds*	*Std. Error*	*P value*
poorLevel	−0.33	0.06	**< 0.001**	−0.33	0.06	**< 0.001**
richLevel	0.08	0.06	0.189			
middleLevel	0.13	0.06	**0.031**	0.18	0.05	**0.001**
Groupsize	- 0.02	0.01	0.058	−0.02	0.01	0.058
Numeracy	0.06	0.03	0.063	0.06	0.03	0.062
Female	0.5	0.07	**< 0.001**	0.5	0.07	**< 0.001**
Age	0.12	0.02	**< 0.001**	0.12	0.02	**< 0.001**
Income	−0.01	0.01	0.422	−0.01	0.01	0.431
Observations	2664			2664		
Adjusted R^2	0.115			0.115		
AIC	9595.639			9595.364		

Note: As for the explanation of independent variables (*poorLevel*, *rich Level* and *middleLevel*), see the note in Table 1. Coefficients are estimated by the ordered logit model. We control for participants' demographic characteristics (gender, age, income, numeracy, and size of groups where participants work or study in their daily lives). The bold numbers in the table highlight those with *p*-values less than 5%, in accordance with the conventions of quantitative analysis

is the model, and the gap is 3.043. This implies that, at the very least, we cannot claim that **S2** is better-fit than **P2**.

5 Summary

The results of our experiment suggest that (1) while ordinary people tend to prefer equal societies rather than unequal societies, we cannot dismiss the tendency to prefer unequal societies in some cases; (2) ordinary people are more sensitive to the maximal threshold than the minimal one; and (3), most importantly, according to the AIC-model selection method, we can in no way claim that **S2** is better-fit than **P2**. In light of these results, we can also state: (4) Some changes, or more specifically the elimination of people's intuitive judgments revolving around refined sufficientarianism may be considered part of the process of going back and forth between principles and intuitions. No doubt these results are important not only because sufficientarianism has enjoyed a wide range of support from philosophers, but also because we cannot evaluate whether the refined sufficientarian principle is more plausible than the

Table 4 Coefficients of sufficientarian models 1 and 2 (**S1** and **S2**)

The dependent variable is a preference for the equal society B	**S1**			**S2**		
Predictors	*Log-Odds*	*Std. Error*	*P value*	*Log-Odds*	*Std. Error*	*P value*
DifMiserableLine	0.12	0.07	0.096	0.11	0.07	0.104
DifSufficientLine	0.29	0.07	**< 0.001**	0.29	0.07	**< 0.001**
numDemiserablized	0.06	0.05	0.267			
numSufficiencialized	0.08	0.05	0.132			
Groupsize	−0.02	0.01	**0.047**	−0.02	0.01	0.05
Numeracy	0.06	0.03	0.058	0.06	0.03	0.054
Female	0.5	0.07	**< 0.001**	0.5	0.07	**< 0.001**
Age	0.11	0.02	**< 0.001**	0.12	0.02	**< 0.001**
Income	−0.01	0.01	0.454	−0.01	0.01	0.461
Observations	2664			2664		
Adjusted R^2	0.113			0.115		
AIC	9599.122			9598.407		

Note: Coefficients are estimated by the ordered logit model. We control for participants' demographic characteristics (gender, age, income, numeracy, and size of groups where participants work or study in their daily lives). The bold numbers in the table highlight those with *p*-values less than 5%, in accordance with the conventions of quantitative analysis

prioritarian principle by the method of cases, i.e., by appealing only to the Beverly Hills case and the Left-Behind case.[16]

6 Discussion

In this paper, we have shown that survey experiments can be used to demonstrate whether theoretical principles are systematically consistent with people's intuitions prompted by possible cases. In a case study, we have conducted an experiment on competing principles of distributive justice, refined sufficientarianism and prioritarianism. What is unique about this experiment is that its results differ from—indeed, contradict—what refined sufficientarians try to show using the two single cases, the Beverly Hills case and the Left-Behind case: We find that **S2** (the sufficientarian statistical model) cannot be said to fit better than **P2** (the prioritarian statistical model). In other words, the experiment shows that the systematically captured folk intuitions

[16] It should be noted that our project does not aim to work on and defend (one of) the two theoretical positions. Rather, we use the statistical models that take into account variables that somehow reflect the ideas behind the positions. In other words, we try to refine the positions so that they are more systematically consistent with intuitions. Thus, our argument does not deny that there is room for further refinement of prioritarianism and even sufficientarianism. This is, we believe, in the spirit of reflective equilibrium. We thank the editor(s) for this point.

did not support what some philosophers and (perhaps) ordinary people find plausible in the context of distributive justice. We can thus say that the particular intuitions evoked by the two particular cases should not simply be taken to speak for or against the principles of distributive justice.

This illustrates the importance of going beyond the method of cases and practicing the method of reflective equilibrium through AIC-based model selection in three respects. First, since AIC measures the predictive accuracy of the model based on the existing data, we can use the model selection method to make reasonable judgments about possible cases. By limiting the complexity of the model, we can make the model easier to use to estimate what a society ought to do without much reducing its predictive power. Second, our appeal to AIC can be regarded as the central stage in the process of reflective equilibrium. That is, as the results of our experiment have shown, the model selected by the AIC scores may run counter to some intuitive reactions to individual cases. Since the goal of AIC is to evaluate the plausibility of relevantly stripped-down models, the proposed method systematizes folk intuitions to pave the way for the justification of the theoretical principle. Third, with the result of model selection, we may expect some intuitions (here, those concerning refined sufficientarianism) to be discarded (tentatively). The practice of reflective equilibrium involves moving back and forth in the system of judgments, which often requires changing intuitions as initial input. In our experiment, particular intuitions in the Beverly Hills case and the Left-Behind case are very likely to be revised in light of the better results of **P2**. To be sure, this is not a full endorsement of prioritarianism, but the results of our experiment provide an important challenge to our reliance on what refined sufficientarians and (some) ordinary people find plausible. With these findings in hand, our proposal can reasonably be thought of as a method of reflective equilibrium, in such a way as to distinguish reflective equilibrium essentially from the method of cases. In Rawls's terms, reflective equilibrium is the systemic equilibrium "reached when someone has carefully considered alternative conceptions of justice and the force of various arguments for them," as distinct from a mere state in which "general beliefs, first principles, and particular judgments are in line" (Rawls, 2001, pp. 30–31).

It may be objected that the proponents of reflective equilibrium, of whom Rawls is representative, are not committed to the idea that intuitions as initial input should be elicited through unrealistic thought experiments.[17] This seems to contradict our appeal to the model selection using survey experiments that involve possible cases of a not-always-common kind. We concede that this is true. But there is a camp of reflective equilibrium theorists who do not ignore the intuitions of ordinary people stimulated by unrealistic thought experiments (De Vries & Van Leeuwen, 2010; Savulescu et al., 2021). We can thus say that at least many proponents of reflective equilibrium cannot ignore what our argument shows.

Admittedly, our argument has limitations. First, our experiment is based on the major debate about theories of distributive justice. Therefore, it remains to be shown whether our argument has broader implications for other philosophical debates. Second, while our experiment has a large sample size, the vast majority of experimental works in philosophy use small samples. It is recognized that AIC is justified in a

[17] This objection is raised by the editor(s).

very general framework and, as a result, provides a crude estimator of the expected discrepancy: one that has a potentially high degree of negative bias in small-sample applications (Cavanaugh, 1997). While our proposal tells us that experiments in philosophy should employ a sample size that is as representative of the population as possible, it may not be broadly generalizable to common experiments in philosophy.[18] Third, since AIC is a measure of the relative quality of a statistical model for a given set of data, if all candidate models fit the data poorly, then AIC may be of no use. This is another limitation of generalizing the use of "imprecise" model selection as a method of reflexive equilibrium in practice.[19] In other words, the AIC-based *ex post* model selection cannot be an "all-purpose tool" in philosophy. However, unless we use it in the wrong way, we believe it is an effective tool.

Acknowledgements We are grateful to the three editors, Claus Beisbart, Gregor Betz, Georg Brun, and two anonymous reviewers for their insightful and helpful comments based on their careful readings of this paper. We would also like to thank Thomas Pölzler for his invaluable comments and useful suggestions on an earlier version of this paper.

Author contributions The authors are responsible for all aspects of this research including analysis, discussion, and writing of the manuscript.

Funding Open Access funding provided by The University of Tokyo. This work was partly supported by JSPS KAKENHI Grant Number JP18H00602, JP20H01446, JP22H00598, and JP23H00802.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical approval This article was written in accordance with the ethical standards of the institutional and/or national research committee.

[18] This point is owed to an anonymous reviewer.

[19] This is also a point made by an anonymous reviewer. However, this limitation should not be exaggerated; if all models fit the data poorly, it can be seen that the experimental conditions with independent variables may not properly reflect the philosophical hypotheses. We can then reconsider the experimental conditions (the content of the independent variables). This reconsideration, we believe, is not inconsistent with the practice of reflective equilibrium.

Appendix

Figures of ten cases used in the survey experiment (See Figures 3, 4, 5, 6, 7, 8, 9, 10, 11, 12).

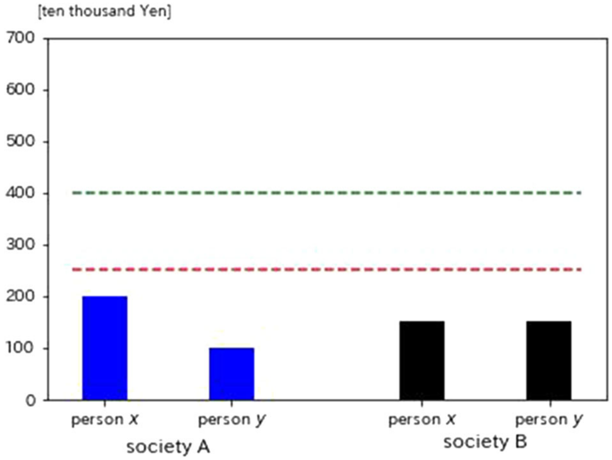


Fig. 3 Case 1: Society A (Lx > Ly), Society B (Lx = Ly)

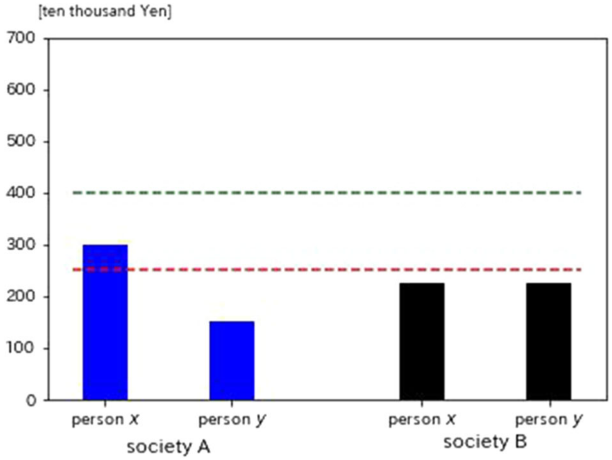


Fig. 4 Case 2: Society A (Mx > Ly), Society B (Lx = Ly)

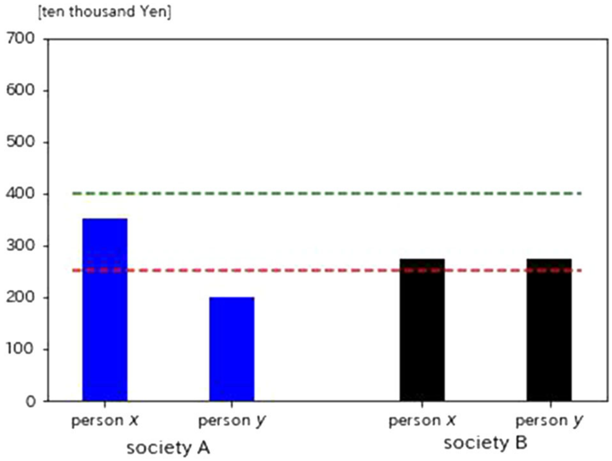


Fig. 5 Case 3: Society A (M*x* > L*y*), Society B (M*x* = M*y*)

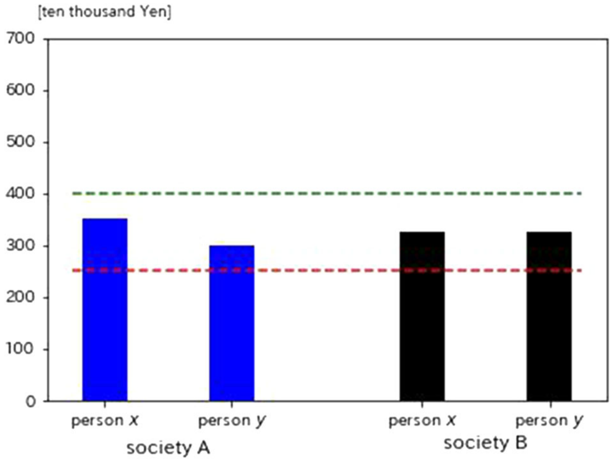


Fig. 6 Case 4: Society A (M*x* > M*y*), Society B (M*x* = M*y*)

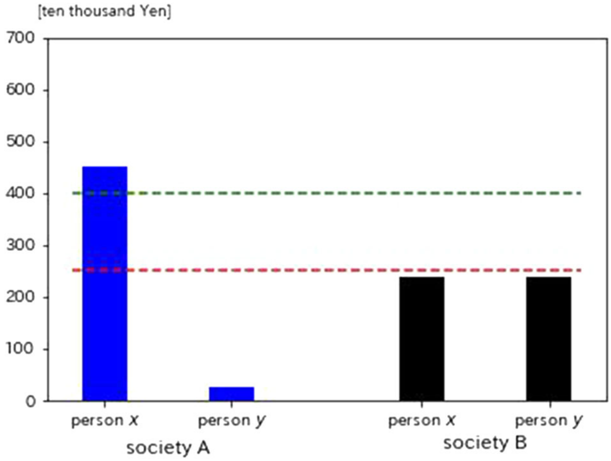


Fig. 7 Case 5: Society A (Ux > Ly), Society B (Lx = Ly)

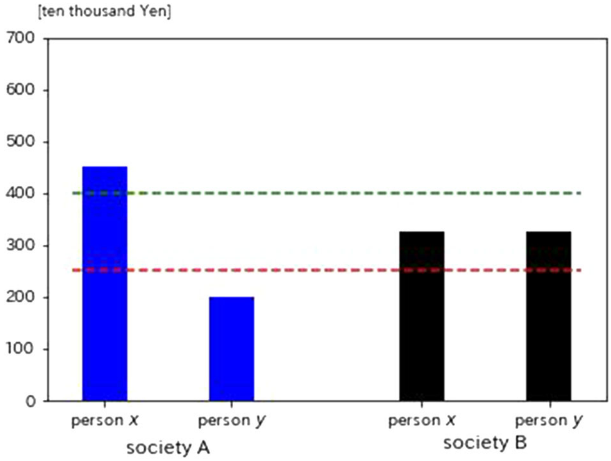


Fig. 8 Case 6: Society A (Ux > Ly), Society B (Mx = My)

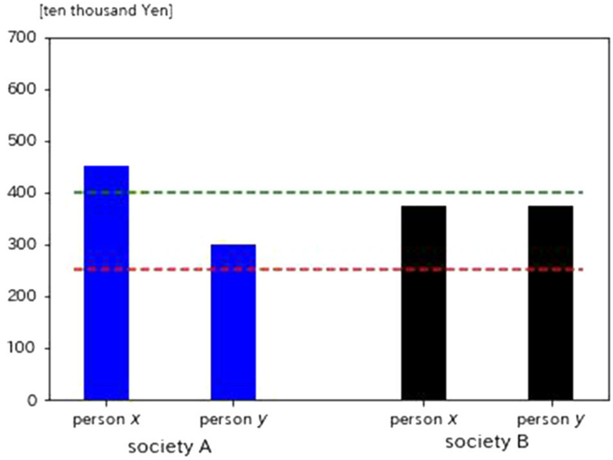


Fig. 9 Case 7: Society A (Ux > My), Society B (Mx = My)

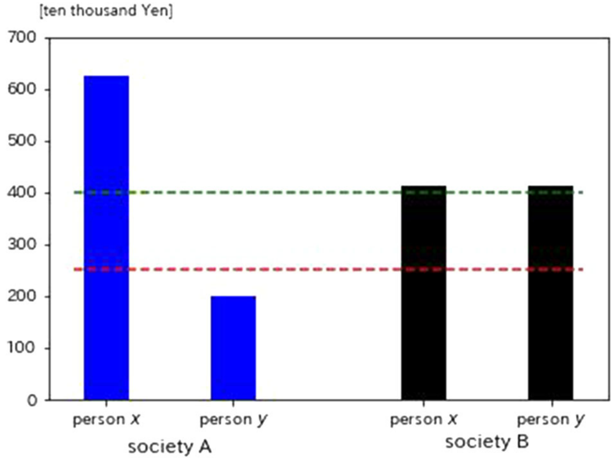


Fig. 10 Case 8 (the Left-Behind case): Society A (Ux > Ly), Society B (Ux = Uy)

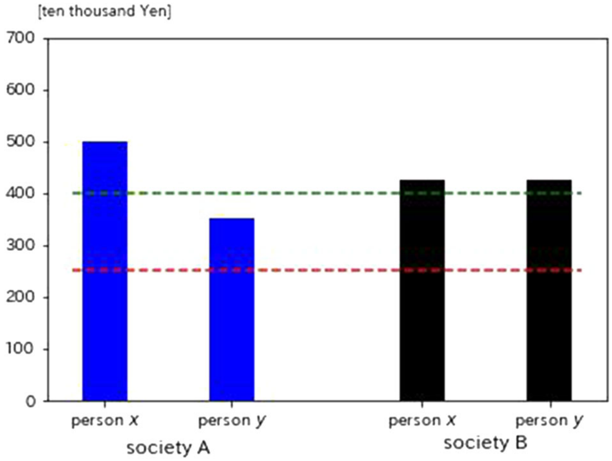


Fig. 11 Case 9: Society A (U*x* > M*y*), Society B (U*x* = U*y*)

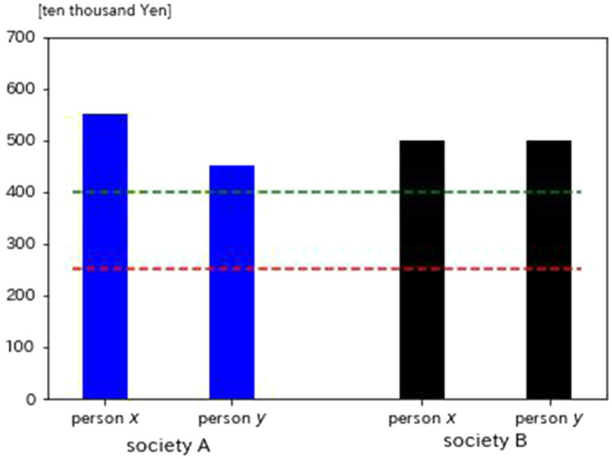


Fig. 12 Case 10 (the Beverly Hills case): Society A (U*x* > U*y*), Society B (U*x* = U*y*)

References

Anscombe, G. E. M. (1957). Does Oxford moral philosophy corrupt the youth? *Listener, 57*(1455), 266–271.

Arneson, R.J. (2013). Egalitarianism. In E.N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/sum2013/entries/egalitarianism/

Arneson, R. J. (2022). *Prioritarianism*. Cambridge University Press.

Baderin, A. (2017). Reflective equilibrium: Individual or public. *Social Theory and Practice, 43*(1), 1–28. https://doi.org/10.5840/soctheorpract20174311

Baumberger, C., & Brun, G. (2021). Reflective equilibrium and understanding. *Synthese, 198*(8), 7923–7947. https://doi.org/10.1007/s11229-020-02556-9

Benbaji, Y. (2006). Sufficiency or priority? *European Journal of Philosophy, 14*(3), 327–348. https://doi.org/10.1111/j.1468-0378.2006.00228.x

BonJour, L. (1985). *The structure of empirical knowledge*. Harvard University Press.

Brandt, R. B. (1979). *A theory of the good and the right*. Clarendon Press.

Brandt, R. B. (1990). The science of man and wide reflective equilibrium. *Ethics, 100*(2), 259–278. https://doi.org/10.1086/293176

Brun, G. (2014). Reflective equilibrium without intuitions? *Ethical Theory and Moral Practice, 17*(2), 237–252. https://doi.org/10.1007/s10677-013-9432-5

Bruner, J. P. (2018). Decisions behind the veil: An experimental approach. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 2, pp. 167–180). Oxford University Press.

Bruner, J. P., & Lindauer, M. (2020). The varieties of impartiality, or, would an egalitarian endorse the veil? *Philosophical Studies, 177*(2), 459–477. https://doi.org/10.1007/s11098-018-1202-8

Cappelen, H. (2012). *Philosophy without intuitions*. Oxford University Press.

Cath, Y. (2016). Reflective equilibrium. In H. Cappelen, T. Gendler, & J. Hawthorne (Eds.), *The Oxford handbook of philosophical methodology* (pp. 213–230). Oxford University Press.

Cavanaugh, J. E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters, 33*(2), 201–208. https://doi.org/10.1016/S0167-7152(96)00128-9

Conte, S. J. (2022). Are intuitions treated as evidence? Cases from political philosophy. *Journal of Political Philosophy, 30*(4), 411–433. https://doi.org/10.1111/jopp.12277

Copp, D. (2012). Experiments, intuitions, and methodology in moral and political theory. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics* (Vol. 7, pp. 1–36). Oxford University Press.

Crisp, R. (2003). Equality, priority, and compassion. *Ethics, 113*(4), 745–763. https://doi.org/10.1086/373954

Daniels, N. (1996). *Justice and justification: Reflective equilibrium in theory and practice*. Cambridge University Press.

Davis, M. (1983). Foetuses, famous violinists, and the right to continued aid. *Philosophical Quarterly, 33*(132), 259–278. https://doi.org/10.2307/2219225

De Vries, M., & Van Leeuwen, E. (2010). Reflective equilibrium and empirical data: Third person moral experiences in empirical medical ethics. *Bioethics, 24*(9), 490–498. https://doi.org/10.1111/j.1467-8519.2009.01721.x

Deutsch, M. (2015). *The myth of the intuitive*. MIT Press.

Ebertz, R. P. (1993). Is reflective equilibrium a coherentist model? *Canadian Journal of Philosophy, 23*(2), 193–214. https://doi.org/10.1080/00455091.1993.10717317

Elgin, C. Z. (1996). *Considered judgment*. Princeton University Press.

Forster, M. R., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45*(1), 1–35. https://doi.org/10.1093/bjps/45.1.1

Frankfurt, H. (1987). Equality as a moral ideal. *Ethics, 98*(1), 21–43. https://doi.org/10.1086/292913

Frohlich, N., & Oppenheimer, J. A. (1992). *Choosing justice: An experimental approach to ethical theory*. University of California Press.

Goodin, R. E. (1982). *Political theory and public policy*. University of Chicago Press.

Gosseries, A. (2011). Sufficientarianism. In *The Routledge encyclopedia of philosophy.* https://www.rep.routledge.com/articles/thematic/sufficientarianism/v-1

Hare, R. M. (1973). Rawls' theory of justice–I. *Philosophical Quarterly, 23*(91), 144–155. https://doi.org/10.2307/2217486

Hassoun, N. (2016). Experimental or empirical political philosophy. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 234–246). Wiley-Blackwell.

Hirose, I. (2015). *Egalitarianism*. Routledge.

Holmgren, M. (1987). Wide reflective equilibrium and objective moral truth. *Metaphilosophy, 18*(2), 108–124. https://doi.org/10.1111/j.1467-9973.1987.tb00192.x

Holtug, N. (1998). Egalitarianism and the levelling down objection. *Analysis, 58*(2), 166–174. https://doi.org/10.1111/1467-8284.00118

Holtug, N. (2007). Prioritarianism. In N. Holtug & K. Lippert-Rasmussen (Eds.), *Egalitarianism: New essays on the nature and value of equality* (pp. 125–156). Clarendon Press.

Holtug, N. (2010). *Persons, interests, and justice*. Oxford University Press.

Huseby, R. (2010). Sufficiency: Restated and defended. *Journal of Political Philosophy, 18*(2), 178–197. https://doi.org/10.1111/j.1467-9760.2009.00338.x

Huseby, R. (2017). Sufficiency, priority, and aggregation. In C. Fourie & A. Rid (Eds.), *What is enough? Sufficiency, justice, and health* (pp. 69–84). Oxford University Press.

Inoue, A., Zenkyo, M., & Sakamoto, H. (2021). Making the veil of ignorance work: Evidence from survey experiments. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 4, pp. 53–80). Oxford University Press.

Knight, C. (2017). Reflective equilibrium. In A. Blau (Ed.), *Methods in analytical political theory* (pp. 46–64). Cambridge University Press.

Lippert-Rasmussen, K. (2007). The insignificance of the distinction between telic and deontic egalitarianism. In N. Holtug & K. Lippert-Rasmussen (Eds.), *Egalitarianism: New essays on the nature and value of equality* (pp. 101–124). Clarendon Press.

Lissowski, G., Tyszka, T., & Okrasa, W. (1991). Principles of distributive justice: Experiments in Poland and America. *Journal of Conflict Resolution, 35*(1), 98–119. https://doi.org/10.1177/0022002791035001006

McMahan, J. (2000). Moral intuition. In H. LaFollette (Ed.), *The Blackwell guide to ethical theory* (pp. 92–110). Blackwell.

O'Neill, O. (1989). *Constructions of reason: Explorations of Kant's practical philosophy*. Cambridge University Press.

Otsuka, J. (2021). Ockham's proportionality: A model selection criterion for levels of explanation. In T. Matsuda, T. J. Wolff, & T. Yanagawa (Eds.), *Risks and regulation of new technologies* (pp. 47–64). Springer.

Parfit, D. (2000). Equality or priority? In M. Clayton & A. Williams (Eds.), *The ideal of equality* (pp. 81–125). Palgrave Macmillan.

Pölzler, T., & Hannikainen, I. R. (2022). The typicality effect in basic needs. *Synthese, 200*(5), 1–26. https://doi.org/10.1007/s11229-022-03859-9

Pust, J. (2000). *Intuitions as evidence*. Routledge.

Rabinowicz, W. (2002). Prioritarianism for prospects. *Utilitas, 14*(1), 2–21. https://doi.org/10.1017/S0953820800003368

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Rawls, J. (2001). In E. I. Kelly (Ed.), *Justice as fairness: A restatement.* Harvard University Press.

Savulescu, J., Gyngell, C., & Kahane, G. (2021). Collective reflective equilibrium in practice (CREP) and controversial novel technologies. *Bioethics, 35*(7), 652–663. https://doi.org/10.1111/bioe.12869

Scanlon, T. M. (2003). Rawls on justification. In S. Freeman (Ed.), *The Cambridge companion to Rawls* (pp. 139–167). Cambridge University Press.

Segall, S. (2016). *Why inequality matters: Luck egalitarianism, its meaning and value*. Cambridge University Press.

Shields, L. (2020). Sufficientarianism. *Philosophy Compass, 15*(11), 1–10. https://doi.org/10.1111/phc3.12704

Singer, P. (1974). Sidgwick and reflective equilibrium. *The Monist, 58*(3), 490–517. https://doi.org/10.5840/monist197458330

Singer, P. (2005). Ethics and intuitions. *Journal of Ethics, 9*(3–4), 331–352. https://doi.org/10.1007/s10892-005-3508-y

Sinnott-Armstrong, W., Young, L., & Cushman, F. (2010). Moral intuitions. In J. M. Doris & Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 246–272). Oxford University Press.

Sober, E. (2002). Instrumentalism, Parsimony, and the Akaike framework. *Philosophy of Science, 69*(S3), S112–S123. https://doi.org/10.1086/341839

Sober, E. (2008). *Evidence and evolution*. Cambridge University Press.

Temkin, L. (2000). Equality, priority, and the levelling-down objection. In M. Clayton & A. Williams (Eds.), *The ideal of equality* (pp. 126–161). Palgrave Macmillan.

Thomson, J. J. (1971). A defense of abortion. *Philosophy and Public Affairs, 1*(1), 47–66.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*(2), 204–217. https://doi.org/10.5840/monist197659224

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395–1415. https://doi.org/10.2307/796133

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Varner, G. E. (2012). *Personhood, ethics, and animal cognition: Situating animals in Hare's two level utilitarianism*. Oxford University Press.

Walsh, A. (2011). A moderate defence of the use of thought experiments in applied ethics. *Ethical Theory and Moral Practice, 14*(4), 467–481. https://doi.org/10.1007/s10677-010-9254-7