COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12841

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12841

How to Be Helpful to Multiple People at Once

Vael Gates,[a] Thomas L. Griffiths,[b] Anca D. Dragan[c]

[a]*Department of Psychology, University of California, Berkeley*
[b]*Department of Psychology, Princeton University*
[c]*Department of Electrical Engineering and Computer Sciences, University of California, Berkeley*

Received 12 September 2019; received in revised form 31 March 2020; accepted 6 April 2020

Abstract

When someone hosts a party, when governments choose an aid program, or when assistive robots decide what meal to serve to a family, decision-makers must determine how to help even when their recipients have very different preferences. Which combination of people's desires should a decision-maker serve? To provide a potential answer, we turned to psychology: What do *people* think is best when multiple people have different utilities over options? We developed a quantitative model of what people consider desirable behavior, characterizing participants' preferences by inferring which combination of "metrics" (*maximax*, *maxsum*, *maximin*, or *inequality aversion* [*IA*]) best explained participants' decisions in a drink-choosing task. We found that participants' behavior was best described by the *maximin* metric, describing the desire to maximize the happiness of the worst-off person, though participant behavior was also consistent with maximizing group utility (the *maxsum* metric) and the *IA* metric to a lesser extent. Participant behavior was consistent across variation in the agents involved and tended to become more *maxsum*-oriented when participants were told they were players in the task (Experiment 1). In later experiments, participants maintained *maximin* behavior across multi-step tasks rather than shortsightedly focusing on the individual steps therein (Experiment 2, Experiment 3). By repeatedly asking participants what choices they would hope for in an optimal, just decision-maker, and carefully disambiguating which quantitative metrics describe these nuanced choices, we help constrain the space of what behavior we desire in leaders, artificial intelligence systems helping decision-makers, and the assistive robots and decision-makers of the future.

Keywords: Fairness; Preferences; Assistive artificial intelligence; Maximin; Modeling

1. Introduction

Consider a schoolteacher trying to plan a field trip. Some students learn best kinetically, others enjoy verbal challenges, and some would be overwhelmed by new locations.

Correspondence should be sent to Vael Gates, Psychology Department, University of California, Berkeley, 2121 Berkeley Way, Berkeley, CA 94704. E-mail: vgates@berkeley.edu

If there is no one action that will maximally satisfy everyone, what is the teacher to do? Now consider a concerned citizen determining how to donate their money. They furrow their brow, staring at the screen: donate to the most needy, the organization where donations will be matched, or the recipient with the tightest time constraints? Next consider a high-level government worker, puzzling over who to prioritize, faced with an array of programs that will benefit everyone to some extent but some more than others. Finally consider an autonomous household robot, trying to determine what to make for the main family meal with one kid who only wants to eat orange foods and one parent who is vegan. What should all of these decision-makers do?

In today’s society, as preference-aggregation problems become more complex, we can turn to tools from computer science to address them. If modern artificial intelligence sys-tems know about the preferences of their recipients, they can use vast computational resources to optimize: airplane scheduling and ride-share services are examples of using artificial intelligence to optimize many people’s preferences. However, these consumer services optimize based on the principles of first-come-first-serve, more resources for more money, and efficiency. Individuals put in a bid and they receive some service. But in other types of situations, people advocate for others rather than themselves. People choose which organizations to donate money towards, and governments strive to serve their entire populations with aid programs. These situations are just as complex and deserving of computational analysis, but they will need a different optimization proce-dure: one that likely incorporates fairness.

To develop artificial intelligence tools to help solve coordination problems, we need to know what the ground truth is. What internal algorithms do people use to make decisions that benefit others? People do not always act in the ways they say they do, but by observ-ing their ground truth behavior, we can gain a quantitative understanding of what behav-iors people think are right. In computer science, there are existing tools for inferring the values, preferences, and utilities that motivate people’s actions and choices. One standard method, inverse reinforcement learning (Ng & Russell, 2000), has been used to infer util-ity functions from behavior in both the robotics community (e.g., Kuderer, Gulati, & Bur-gard, 2015 in which people’s driving styles were inferred from demonstrations) and the computational cognitive science community (e.g., Baker, Saxe, & Tenenbaum, 2009, in which intended goals were inferred from partially completed paths). There are thus estab-lished methods to learn human preferences from actions: Given examples of a person’s driving, we are advancing in our ability to tell an autonomous vehicle how to bring them to the store; given examples of a person tidying up their home, we are becoming able to tell a robot what to clean. Yet these methods only apply to a single person’s preferences at a time. What should be done when there is more than one person, and there are a com-bination of utilities to optimize? Should an artificial intelligence sum up all of its users’ utilities, or would it be more “fair” to minimize the upset of the unhappiest member of the group? When we have assistive robots who must act as decision-makers themselves, how should they be programmed? We think the first step to answering these questions is to quantitatively determine how people—with their intuitions about fairness and effi-ciency and what is good—behave.

In this paper, we set out to map the human sense of compromise, and the challenge of the benign dictator: how to solve the age-old problem of acting in *everyone's* best interest without a vested interest of one's own. It is a challenging problem that has puzzled philosophers, arbitrators, and governors for centuries. The question of what should be done when people disagree is essential, and to build tools that can help solve it, we must understand the ground-truth of what we want done. This problem becomes increasingly important not only to advise decision-makers, but to develop artificial intelligence systems to help decision-makers and eventually create artificial intelligence agents that can make choices that match what we do ourselves. In the face of this unsolved problem, we turn to psychology, developing a quantitative model of what people think should be done when an agent can take only one action that brings different utilities to different people.

Many researchers have thought about these questions. Their work often falls under the heading of *fair allocation*, which investigates how items should be divided among people. The study of fair allocation and fair division (Brams & Taylor, 1996; Konow, 2003) spans many fields, including those of social choice (Gaertner & Schokkaert, 2012; Moulin et al., 2016), neuroscience (Hsu, Anen, & Quartz, 2008), artificial intelligence (Dickerson, Goldman, Karp, Procaccia, & Sandholm, 2014), justice and policy (Fleurbaey, 2008; Gollwitzer & van Prooijen, 2016; Konow, 2003), and behavioral economics and game theory, especially within the dictator, ultimatum, and estate games (Ashlagi, Karagözoğlu, & Klaus, 2012; Chmura, Kube, Pitz, & Puppe, 2005; Dreber, Fudenberg, & Rand, 2014; Fisman, Kariv, & Markovits, 2007; Huck & Oechssler, 1999; Nowak, Page, & Sigmund, 2000; Pálvölgyi, Peters, & Vermeulen, 2010). Empirical work on fairness is similarly wide-ranging, including work on cultural differences (Gaertner, Jungeilges, & Neck, 2001; Jungeilges & Theisen, 2008; Schäfer, Haun, & Tomasello, 2015; Schokkaert & Devooght, 2003), the developmental trajectory of fairness (Wittig, Jensen, & Tomasello, 2013), the impact of other-regarding preferences in games (Austerweil et al., 2015; Bolton, Brandts, Katok, Ockenfels, & Zwick, 2008), and the weighting of distributive allocation versus the procedure by which items are allocated (Cooney, Gilbert, & Wilson, 2016; Dupuis-Roy & Gosselin, 2011). Wanting quantitative measures, researchers have also developed metrics to quantitatively evaluate the fairness of solutions. Several of these metrics have been compared empirically (Dupuis-Roy & Gosselin, 2009; Engelmann & Strobel, 2004; Fehr, Naef, & Schmidt, 2006; Herreiner & Puppe, 2007) and include, for example, envy-free fairness, proportional fairness, and inequality aversion (IA). The tradeoffs between fairness and efficiency (maximizing the joint utility of all agents) are often considered and have been evaluated theoretically (e.g., Bertsimas, Farias, & Trichakis, 2011, 2012).

Much work on fair allocation, however, falls outside our purview, as it considers a self-interested agent deciding how to distribute resources between themself and others. When participants have a personal stake in the situation, biases can appear in their interpretation of what is fair or what behavior they exhibit (Babcock, Loewenstein, Issacharoff, & Camerer, 1995; Beckman, Formby, Smith, & Zheng, 2002; Binmore, 1994; Cappelen, Nielsen, Sørensen, Tungodden, & Tyran, 2013; Croson & Konow, 2009; Ellingsen & Johannesson, 2001; Gächter & Riedl, 2006; Herrero, Moreno-Ternero, &

Ponti, 2010; Konow, 2000, 2009; Traub, Seidl, Schmidt, & Levati, 2005). This paper focuses on the perspective of a third party, the scenario in which an impartial decision-maker is choosing how to best help a group of people.

As such, our experiments are most similar to the subset of fair allocation experiments in which the decision-maker's utility is tied only to the utilities of the agents it is serving (e.g., selfless artificial intelligences built to serve human needs). In one such set of studies, participants acted as dictators to fairly allocate goods when the dictator's own payoffs were fixed (Engelmann & Strobel, 2004; Fehr et al., 2006). Another two related papers investigated the division of multiple goods by an uninvested agent (Herreiner & Puppe, 2007; Yaari & Bar-Hillel, 1984). In these cases, participants chose how to distribute multiple items across agents whose utility functions were represented in payoff matrices. Our experiments have a similar structure in that they consider preferred allocations over payoff matrices.

Our work is distinct from those previous fair allocation studies, however, in the problem it is solving: Here, *the uninvested agent can take one action that has consequences for multiple individuals.* Our problem formulation is more general than the "fair allocation" problem, since the idea of "choosing actions that have implications across multiple utility functions" applies to many situations outside the distribution of items. Specifically, the problem setup we describe, with utilities over any possible action, is more general than having utilities over specific items being distributed. While we are still working with abstract entities in this work, payoff matrices, and in this work only working with positive utilities, this problem formulation admits a range of scenarios. For example, a schoolteacher trying to determine what field trip to bring students on is an example of our problem, but not a fair allocation problem. Necessarily, our formulation also encompasses the class of fair allocation problems, including problems like sorting multiple indivisible items into "bundles" to be distributed to individuals (e.g., Herreiner & Puppe, 2007). In this case, each "action" is to give each agent each sorted bundle, and if necessary each of these agent-specific actions could be characterized as a single larger action that could be repeated over multiple allocation decisions.

It is also worth emphasizing how the question posed in this paper differs from others in the literature, even those that do not involve a self-interested party. For one, this paper does not singularly employ the zero-sum scenarios present in fair allocation studies. In fair allocation studies, participants distribute a set number of items between agents, and if one agent receives an item, another agent loses it. In this work, participants choose to create an item that can bring utility to multiple agents. For these questions, high utility for one agent does not necessarily imply low utility for another.

In addition, this paper asks participants "what *should* be done" rather than what is "fair." Empirically, "fairness" can correspond more to ideas like equality and helping the needy, while questions like "in which society would you like to live?" can evoke responses that take into account the trade-off between equality and maximizing the benefit to the group. Fairness can also correspond to distributional fairness or procedural fairness—examining the fairness of the final allocation solutions or the process by which items are allocated. This work focuses not on fairness but on what decisions participants believe a third party should make, in terms of creating a shared item to be distributed.

Similarly, the question of "what should be done" is incredibly dependent on context (see Konow, 2003; Konow & Schwettmann, 2016 for reviews), and our question of what an uninterested artificial intelligence should do constrains the space of those contexts enormously. In many decision-making tasks, arbitrators are contending with preferences for honesty (Dana, Weber, & Kuang, 2007), altruism (Andreoni, 1989; Pelligra & Stanca, 2013), cooperation (Dreber et al., 2014), previous experiences, friendship, spite (Beckman et al., 2002; Levine, 1998), reciprocity (Berg, Dickhaut, & McCabe, 1995; Bolton & Ockenfels, 2000; Cox, 2004; Dufwenberg & Kirchsteiger, 2004), and any number of highly relevant factors of what is fair and what is preferred. We are interested in what people think should be done in the absence of such considerations: what they would like their leaders or assistive artificial intelligence systems to advise before previously existing social relationships come into play. Artificial intelligences are by default honest and not more altruistic or cooperative than preferred, so they offer a unique opportunity to act as independent advisors free from spite, obligation, or reciprocity.

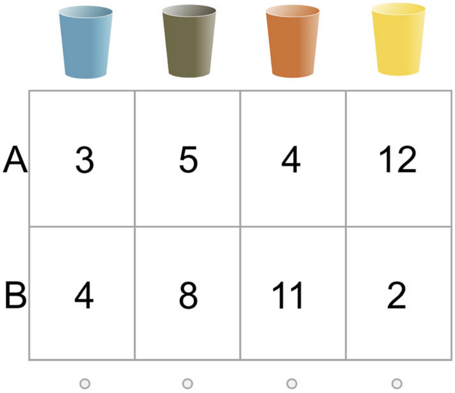
Our question could integrate important factors like need or desert/merit (e.g., Alesina & Angeletos, 2005; Fleurbaey, 2008; Fong, 2001; Hoffman & Spitzer, 1985; Konow & Schwettmann, 2016; Nord, Richardson, Street, Kuhse, & Singer, 1995; Schokkaert & Devooght, 2003; Schokkaert & Lagrou, 1983; Schokkaert & Overlaet, 1989), but the simplified task we choose does not involve agents who are needy or who have worked harder than others, despite this being a major factor in society. Another way to extend our study would be to use a paradigm that uses bundles or collections of objects, either for allocation or as creating these shared resources (both can be construed as an action). Bundles allow investigation of envy-freeness, and also proportionality, like in "claims problems" when agents may each have a claim to a resource that is greater than the resource is worth (Bosmans & Schokkaert, 2009; Thomson, 2003). Additionally, Konow (2003) discusses the differences between subjective values and objective values, which we simplify here by directly presenting agent utilities. Given the difficulty of the question of what should be done, we used a simplified task and save these questions for future work. For our question and paradigm, we consider it reasonable to limit our literature review to those topics that are most aligned with our question, though we provide references to the reader for additional study.

Our task, then, is to determine what an uninvested, autonomous agent should do in a decision-making scenario with multiple recipients. To investigate what defines a "helpful" action when balancing multiple utilities, we created the following problem setting: We asked what decision a manager should make in choosing a drink for two guests. Participants acted as the manager, repeatedly making choices that we used to develop a model of what people think are good decisions. Drawing on the literature in economics, computer science, and philosophy, we considered four metrics as hypotheses to capture participants' behavior: *maximax*, *maxsum*, *maximin*, and *IA*. We determined which combination of metrics best explained participants' behavior by statistical analyses and evaluating the output of a maximum entropy inverse reinforcement learning (MaxEnt) model. Having inferred the preferred metrics, we then used these metrics to compare participants' behavior across conditions.

1.1. Metrics

We now delve into the specifics of our metrics and what stimuli they were applied to. Many different fields have investigated the question of how people make decisions when balancing multiple tradeoffs—for example, people could choose to maximize the group utility or distribute resources evenly. As such, several quantitative models describing "fair" behavior have been suggested. We investigated four metrics that were often used (often in combination) in previous studies to characterize human behavior (e.g., Brams, Edelman, & Fishburn, 2003; Dupuis-Roy & Gosselin, 2009; Engelmann & Strobel, 2004; Herreiner & Puppe, 2007). We do not believe these metrics span the space of reasonable preferences people may hold—people's algorithms for determining what to do in the world can be extremely complex—but we selected these metrics based on what has been widely and empirically observed in the literature. We applied these four metrics to a set of payoff matrices $\mathcal{M}$: in each matrix $\mathcal{M}_m$, two agents' utilities (A and B) were shown for each of four drink options (see Fig. 1). We investigated each of the metrics for these payoff matrices.

Intuitively, one method to select a shared option is to add up the utilities of all agents, and pick the option that maximizes this joint utility value. This is a metric called *maxsum*, which maximizes the happiness of the group rather than individuals. It is a Pareto optimal option. In Fig. 1, the best *maxsum* option would be the third cup, because when utilities of agents A and B are added together, these sums are as follows: 7 (first cup), 13 (second cup), 15 (third cup), and 14 (fourth cup). A different metric enforcing fairness is *maximin*: maximizing the utility of the person who is worst-off. Intuitively, this metric means making sure that no individual agent is very unhappy. In Fig. 1, the best *maximin* option would be the second cup, because when we look to which of the agents are worst-off in each of the pairs, these agents' utilities are as follows: 3 (first cup), 5 (second



A	3	5	4	12
B	4	8	11	2
	○	○	○	○

Fig. 1. Example matrix $\mathcal{M}_m$. Columns are options o^j, while rows show each agent i's utility $\mathcal{U}^i$. Here, the problem is to decide which drink to serve to both agents A and B.

cup), 4 (third cup), and 2 (fourth cup), and the second cup leaves the worst-off agent with the highest possible utility. Another measure of fairness is *IA*: decreasing the difference in utilities between agents. Intuitively, this metric means that agents should be equally happy. In Fig. 1, the best *IA* option would be the first cup, because the agents' utilities will be maximally close to each other. A final possible measure is *maximax*: maximizing the highest utility, searching across both agents. Intuitively, this metric means that any one agent should be made as happy as possible. This option might not initially seem fair, but if presented with the opportunity for repeated choices, pleasing one agent each round may seem like the best solution. In Fig. 1, the best *maximax* option would be the fourth cup, because this cup allows one agent to have its highest possible utility. Note that in Fig. 1, the best example of each metric was a different cup, but in most of our payoff matrices, the best example of multiple metrics was the same cup. Thus, in this paper, we evaluated the four metrics of *maximax*, *maxsum*, *maximin*, and *IA*.

Metrics like *maximin* and *maxsum* have a long history in the literature. The concept of *maximin* was popularized by Rawls (1971, 1974), who advocated for allocating resources to the least well-off individual. Harsanyi (1975) argued in favor of expected utilitarianism, loosely the *maxsum* principle, except in cases where the *maximin* choice was similar to the *maxsum* choice, as in the case in our experiments. There has been a strong focus on *maximin* mathematically and in economics (e.g., Amanatidis, Markakis, Nikzad, & Saberi, 2017; Barman & Krishna Murthy, 2017; Dubois, Fargier, & Prade, 1996; Escoffier, Gourvès, & Monnot, 2013; Kurokawa, Procaccia, & Wang, 2016; Procaccia & Wang, 2014), including applications to networking (e.g., bandwidth-sharing) (Salles & Barria, 2008). Choices aligned with the *maximin*, *maxsum* metric, and *IA* metrics are often compared in studies, and results are often mixed, with participants trading off between different metrics depending on the numbers and contexts involved (Ahlert, Funke, & Schwettmann, 2013; Charness & Rabin, 2002; Engelmann & Strobel, 2004; Faravelli, 2007; Fehr et al., 2006; Gaertner et al., 2001; Gaertner & Schwettmann, 2007; Konow, 2001, 2003; Konow & Schwettmann, 2016; Mitchell, Tetlock, Mellers, & Ordonez, 1993; Ordoñez & Mellers, 1993; Pelligra & Stanca, 2013; Schwettmann, 2009, 2012). These studies differ in their experimental paradigms, and results have been subsequently different, even as the concept of tradeoffs between a few principles of justice has remained similar (Konow, 2003). These experiments reveal a few common difficulties as well. Many experiments attempt to target each metric in isolation, which does not account for the correlations between choices: A choice that maximizes the *maximin* metric also tends to rate highly on the *IA* metric and less well on the *maxsum* or *maximax* metrics. Additionally, experiments often present somewhat extreme choices and models contain variables that are occasionally confounded. We aim to address these difficulties in our study by presenting many nuanced choices to participants, and then using a computational model and statistical analyses to account for correlations and confounding between metrics. Using these tools, we gain the ability to distinguish the relative influence of metrics on sets of participant choices.

In this paper, we aim to address the question of what policies people would like decision-makers, and the assistive technologies assisting decision-makers, to have in the future. To this end, we focus on research studies that are aimed at non-interested third parties reasoning about helpful decision-making, the results of which are not always consistent. In these studies, participants are presented with options that implement various metrics and have to choose how to allocate items across agents (Engelmann & Strobel, 2004; Fehr et al., 2006; Herreiner & Puppe, 2007; Yaari & Bar-Hillel, 1984). Participants are often shown choices that are used to disambiguate between metrics, and because these choices are so distinct, participants may only make a single or small number of choices for any given prompt. We took a different approach with our work: On each prompt we presented many choices to participants, accepting the high amount of overlap among metrics necessary to describe participants' responses. We constructed a computational model to accommodate these correlations and used this model to reveal the relative contributions of different metrics across many probes of participant behavior. We tested the generalizability of our findings by ensuring the participant behavior was similar across different prompts. Additionally, previous work often focuses on short-term, single decisions; we investigated participants' intuitions in the repeated setting where they could make more long-term decisions. In summary, our work compares four common metrics in a setting where participants are making more choices than in previous work and the correlations among metrics describing those choices are accounted for. We use this higher resolution into the relative contributions of each of these metrics to directly compare them, and we probe many questions—including longer-term decision-making—to determine the generalizability of our findings. These improvements take place in a novel setting, in which participants are not being asked about how to allocate many items, but to create a resource to be shared among multiple agents. We thus present a novel test of how correlated metrics interact, in the context of how people feel third-party decision-makers should balance others' utilities: a problem formulation that will occur more and more often in technologies in the future.

In three experiments, we investigated the contributions of the metrics of *maximax*, *maxsum*, *maximin*, and *IA* in describing what people consider good or helpful behavior. In Experiment 1, we examined participants' choices in the drink task and determined whether these choices were consistent when the hypothetical decision-maker was described as a human or a robot, and when the agents receiving the resources were described as friends or strangers. Our results indicate that participants were consistent in using the *maximin* metric to make decisions (maximizing the utility of the worst-off agent) despite variations in the agents involved. In Experiment 2, we asked what decisions people would make when they had the opportunity to offer more than one drink to the same set of agents. We tested whether people would make choices while considering the entire multi-decision expected utility or focus on the individual decisions within. We found that participants reasoned over the entire multi-decision process, and the *maximin* metric could again describe their choices. In Experiment 3, we validated our results from Experiment 2 by presenting participants with the cumulative sum of the choices available in Experiment 2, and observing that when the multi-decision problem was condensed to a single instance again, participants reliably made choices described by the *maximin* metric.

2. Experiment 1: Which metrics describe people's behavior? Are they robust to changes in agent?

Here we tested which combination of metrics described participants' behavior on a drink-choosing task (Fig. 1). We tested a variety of different phrasings and altered scenarios to ensure that the results were consistent across changes in presentation. Specifically, we tested whether participants' choices changed depending on if the decision-maker was an artificial intelligence (a robot) or a human, whether the recipients of the drinks were friends or strangers, and whether the participant was stated to be one of the recipients of a drink. We might expect that a participant would perform differently in the "Robot" condition if they thought that what a robot should do in a manager role was different from what a human manager should do. For example, participants may think that artificial intelligences need to remain completely impartial or compute everything exactly, whereas humans should rely on gut instincts. Similarly, we might expect that participants would have different intuitions about a server giving drinks to strangers rather than friends. Perhaps participants would think that when serving friends, a server should worry more about joint happiness rather than making sure utilities were equitable, since friends could make it up to each other later, while the same would not be true of strangers. We were also interested in whether participants' opinions of what decision should be made would change if they themselves were receiving an item rather than hypothetical recipients. In Experiment 1, we were aiming to test whether the participants' empirical intuitions of good decision-making would extend across these variations in scenarios, which are important for the generalizability of our findings.

2.1. Method

2.1.1. Participants

Participants with U.S. IP addresses were recruited from Amazon Mechanical Turk across five conditions: "Nominal" ($n = 36$, 0 participants excluded), "Robot" ($n = 35$, 1 participant excluded), "Robot Friends" ($n = 35$, 1 participant excluded), "Robot Strangers" ($n = 34$, 2 participants excluded), and "Veil of Ignorance" ($n = 33$, 3 participants excluded). Participants were paid between \$2.50 and \$3.00 for their participation. Participants were excluded if they failed the included attention check or indicated that they did not understand the experiment.

2.1.2. Stimuli

Stimuli were chosen such that participants would reason over a set of positive-utility actions, and the effects of each metric could be quantitatively isolated from these data. To keep the paradigm simple, utilities in the form of numbers in a table were used ("payoff matrices"), with utilities kept small to allow participants to use simple addition. To maximize the generalizability of the findings, several constraints on the matrices were put in place to ensure participants were reasoning over a wide, randomized range of independent matrices.

Participants viewed 20 of these payoff matrices (set of matrices $\mathcal{M}$), one at a time. Each matrix $\mathcal{M}_m$ was 2 × 4, where the rows indicated the two agents, the columns indicated the four colored drinks, and each agent's utility for each drink was shown.

Matrices were generated by rejection sampling. Matrices were subject to the following general constraints. The sum of agents' utilities for each option o^j, $\sum_i \mathcal{U}^i(o^j)$, was constrained to be within 2 and 16, so that there would be a wide range of choices available but participants would only need to use simple addition. Within each matrix, columns (both agents' utilities for an option) were not allowed to repeat, including permutations within columns, so that there would always be four independent choices within a matrix. Moreover, within each matrix, for every column, there could be no other column that strictly dominated that column according to all metrics, because such a column would rarely be picked and so would be uninformative according to the goals of this study. Since we were interested in the question of what desired resource should be shared among recipients, we constrained our actions to positive utilities and did not allow matrices to contain zeros or negative numbers. Finally, in a set of matrices $\mathcal{M}$, any two matrices were not permitted to have more than two of the same columns, where "sameness" included column permutation, in order to create a set of independent matrices.

In addition to the general constraints, "class" constraints were established to ensure that the chosen matrices could isolate the effects of each of the metrics. There were thus four classes: (a) four matrices in which each option was the best choice according to one of the metrics (e.g., the example in Fig. 1), (b) four matrices in which the *maxsum* metric was held constant: all choices had the same joint utility, (c) four matrices in which the *maximin* metric was held constant: for all choices, the worst-off person had the same utility, and (d) eight matrices that were randomly generated. The *maximax* and *IA* metrics were not held constant because matrices constructed in this manner did not meet the general constraints described above. The matrices used in this study are included in the Supplemental Methods.

2.1.3. Procedure

Upon viewing each matrix, participants read the following text: "You're the manager at a hotel and want to serve a drink to the room. Archie and Ben are your guests and have told you how much they enjoy different drinks (higher numbers mean more enjoyment). Which drink would you like to serve?" Participants then had to select one of the four drinks and justify their response. The names "Archie" and "Ben" were substituted with other names beginning with "A" and "B" for each matrix.

We expected that participants would employ a consistent understanding of what made a "good" decision in the drink-choosing task, but wanted to test the robustness of participants' choices by employing several task variations. The variations we tested were the identity of the server (either human or robot), the relationships of the agents being served (either friends or strangers), and whether the participant was described as a recipient of a drink.

In the "Nominal" condition, participants were presented with the default text ("You're the manager at a hotel and want to serve a drink to the room..."). In the "Robot" condition, the prompt was: "There is a robot manager at a hotel which will serve a drink to the room. Archie and Ben are its guests and have told it how much they enjoy different drinks (higher numbers mean more enjoyment). Which drink would you like the robot to serve?" In the "Robot Friends" condition, the prompt was the same as in the "Robot" condition, but the following phrase was added: "Archie and Ben are its guests *(they are friends with each other)....*" In the "Robot Strangers" condition, the following phrase was substituted: "Archie and Ben are its guests *(they are strangers to each other)....*" Italics were not included in the participant text. In the "Veil of Ignorance" condition, the instructions were modified to be: "The manager at a hotel wants to serve a drink to the room, where you and another guest are sitting. The manager has learned how much you both enjoy different drinks (higher numbers mean more enjoyment). Given you do not know which guest (A or B) you are, which drink would you like the manager to serve?" This condition is named the "Veil of Ignorance" condition as a reference to Rawls (1971), who suggested that a method of eliciting moral judgments without self-interest would be to present scenarios in which participants had to make judgments about hypothetical societies they would like to live in, while not knowing anything about their place in the hypothesized social order. Thus, participants would be "behind a veil of ignorance" in that they would have to make decisions without knowing whose place they would fill. Here, in this experimental condition, participants were told they were recipients of a drink but did not know which recipient (A or B) they were, thus enacting a version of the Rawlsian thought experiment.

To analyze participant responses according to our metrics, we calculated the value of each metric ($\mathcal{F}_q$), where q indexes the individual metric (*maximax, maxsum, maximin, IA*), and $\mathcal{F}$ is a vector. Values for the metrics were calculated from the utilities of agent i ($\mathcal{U}^i$) for each option o^j:

$$\mathcal{F}_{maximax}\left(o^j\right) = \max_i \mathcal{U}^i\left(o^j\right)$$

$$\mathcal{F}_{maxsum}\left(o^j\right) = \sum\nolimits_i \mathcal{U}^i\left(o^j\right)$$

$$\mathcal{F}_{maximin}\left(o^j\right) = \min_i \mathcal{U}^i\left(o^j\right)$$

$$\mathcal{F}_{IA}\left(o^j\right) = \prod\nolimits_i \frac{\mathcal{U}^i(o^j)}{\sum_i \mathcal{U}^i(o^j)}$$

These $\mathcal{F}_q(o^j)$ values were then normalized to account for the tradeoffs between each option o^j in the matrix $\mathcal{M}_m$[1]:

$$\mathcal{F}_q\left(o^k\right) = \frac{\mathcal{F}_q\left(o^k\right)}{max_{o^j \in \mathcal{M}_m}\left\{\mathcal{F}_q\left(o^j\right)\right\}}, \forall q \tag{1}$$

2.2. Results and discussion

We were interested in what metrics participants preferred, meaning which metrics could describe participants' demonstrated behavior. We determined preferred metric through an independent statistical analysis, and then via a maximum entropy model. We examined the proportion of participants preferring specific metrics in the "Nominal," "Robot," "Robot Friends," "Robot Strangers," and "Veil of Ignorance" conditions.

2.2.1. Statistical analysis

We sought to determine which metrics individuals used to make their choices. Before asking which metrics *best* described participants' choices, we first asked whether participants were behaving according to any of the metrics, by comparing participants' empirical choices to a simulated baseline participant making random choices. For each metric q, we determined whether the $\mathcal{F}_q$ values from participants' empirical choices were significantly larger than the $\mathcal{F}_q$ values from random choices, summed across participants and matrices. For our baseline comparison, we drew from the null distribution to generate enough choices for a complete experiment (choices for each matrix and for each participant, summed) and repeated that process 10,000 times. We then computed z-scores and associated p-values for whether the empirical $\mathcal{F}_q$ values (H1) were significantly greater than our baseline $\mathcal{F}_q$ values (H0) for each metric q. We found that the *maxsum*, *maximin*, and *IA* metrics significantly matched participant behavior, and that the *maximin* metric best matched participant behavior (had the highest z-scores) for all conditions except the "Veil of Ignorance" condition (Table 1). In the "Veil of Ignorance" condition, the *maxsum*, *maximin*, and *maximax* metrics significantly matched participant behavior, and the highest z-score was for the *maxsum* metric, closely followed by the *maximin* metric. This analysis was chosen because the comparison between the empirical and null distributions accounted for the biases within metrics and choice of specific matrices.

2.2.2. Inferring metrics used with a maximum entropy model

To further test which combination of metrics described participants' choices, we constructed a maximum entropy inverse reinforcement learning (MaxEnt) model (Ziebart, Maas, Bagnell, & Dey, 2008). In this model, we inferred weights, which were associated with each of the metrics. Mathematically, the weight vector θ was indexed by q and composed of $[\theta_{maximax}, \theta_{maxsum}, \theta_{maximin}, \theta_{IA}]$. Participants' preferred metric was evaluated as the metric with the largest associated weight θ_q.

We computed a maximum a posteriori estimate for θ for each participant given their choices by combining a uniform prior over the parameters of θ with the likelihood of each individual's choice o^k over the options o for each matrix. Specifically, within each matrix $\mathcal{M}_m$, for each participant's choice o^k over options o, we maximized the function $\log P(o^k|\theta)$ over θ (a vector of length q), where:

$$P(o^k|\theta) = \frac{e^{\theta^T}\mathcal{F}(o^k)}{\sum_j e^{\theta^T}\mathcal{F}(o^j)} \tag{2}$$

Table 1
Results for Experiment 1, showing relationship between participants' responses and hypothetical null distribution for each metric

	Maximax	*Maxsum*	*Maximin*	*IA*
Nominal	−0.2 (.58)	11.2 (0)	19.7 (0)	10.9 (0)
Robot	0.6 (.27)	10.7 (0)	16.7 (0)	8.6 (0)
Robot Friends	1.6 (.05)	13.6 (0)	19.0 (0)	9.7 (0)
Robot Strangers	0.6 (.27)	9.5 (0)	14.7 (0)	7.3 (0)
VoI	3.1 (.002)	8.7 (0)	7.8 (0)	−0.1 (.53)
Rep2x: Ind.(C1)	4.1 (0)	9.7 (0)	7.8 (0)	1.9 (.03)
Rep2x: Ind.(C2)	1.3 (.10)	5.6 (0)	6.9 (0)	2.8 (.002)
Rep3x: Ind.(C1)	−1.2 (.89)	5.7 (0)	11.9 (0)	6.9 (0)
Rep3x: Ind.(C2)	0.4 (.34)	3.6 (1e-4)	5.4 (0)	3.0 (.001)
Rep3x: Ind.(C3)	−2.1 (.98)	1.8 (.03)	7.3 (0)	5.0 (0)
Rep2x: Summed	−9.4 (1)	9.4 (0)	19.2 (0)	14.4 (0)
Rep3x: Summed	−24.0 (1)	−0.007 (.50)	26.1 (0)	20.8 (0)
Follow-Up (2x)	−11.0 (1)	5.3 (0)	22.1 (0)	14.7 (0)
Follow-Up (3x)	−6.4 (1)	7.8 (0)	18.8 (0)	9.9 (0)

Note. Determining which metrics describe participant choices, as compared to what would be expected given random choices, summed across participants and matrices. *Z*-scores are shown, calculated as the [empirical (H1) summed scores—mean of the null (H0) distribution summed scores] divided by [*SE* for the null (H0) distribution summed scores], along with associated *p*-values in parentheses, for all metrics. (Summation is across all matrices and participants.) High scores for any metric indicate that participants often chose responses in accordance with that metric, above and beyond what they would have done had they been choosing randomly. Results are shown for all conditions in all experiments: "Nominal," "Robot," "Robot Friends," "Robot Strangers," "Veil of Ignorance," "Repeated 2x: Independent (Choice 1)," "Repeated 2x: Independent (Choice 2)," "Repeated 3x: Independent (Choice 1)," "Repeated 3x: Independent (Choice 2)," "Repeated 3x: Independent (Choice 3)," "Repeated 2x: Summed," "Repeated 3x: Summed," "Follow-Up (2x)," and "Follow-Up (3x)." 10,000 samples of summed scores from the null distribution were taken. Condition names are shortened: "VoI" represents the "Veil of Ignorance" condition and "Rep2x: Ind.(C1)" represents the "Repeated 2x: Independent (Choice 1)" condition. "0" in the table refers to <0.0001.

Recall that $\mathcal{F}$ is the vector containing the values of the individual metrics $[\mathcal{F}_{maximax}, \mathcal{F}_{maxsum}, \mathcal{F}_{maximin}, \mathcal{F}_{IA}]$.

We summed across matrices $\mathcal{M}_m$ to create the final cost function $\sum_m \log P\left(o^k(\mathcal{M}_m)|\theta\right)$, and we used the following constraints to compare the relative use of each of the metrics: $\sum_q \theta_q = 1$, $\theta_q \geq 0$. We optimized using sequential least squares programming (SLSQP) in Python's *scipy* package. To calculate the weight vector of an average individual, we calculated the arithmetic mean over individual participants' θ vectors. Thus, overall the θ_q value for any given metric was determined by the closeness between the responses expected under that metric (e.g., *maxsum*) and a participant's responses. In interpreting the results, if any given θ_q value was high, then participants often chose responses in accordance with that metric.

We found that participants' behavior was best explained by the *maximin* metric for all conditions according to the MaxEnt model, as shown in Fig. 2. Our conclusion that participants' behavior was best explained by the *maximin* metric was also supported by using an alternative method of calculating the average weights across participants. Throughout this work, we sought to determine which combination of metrics described an "average" individual; however, there are several ways of calculating an average weight vector. In the main analysis, we set "average" as the arithmetic mean of each individual participant's weight vector θ. An alternative average assumes that all participants' data came from one individual and, given this, calculates a single weight vector θ under the assumptions of the MaxEnt model described above. Using this, $\theta_{maximin} = 1$ and $\theta_{maximax}$, θ_{maxsum}, $\theta_{IA} = 0$ for the "Nominal," "Robot," "Robot Friends," and "Robot Strangers" conditions.

Note that the results from the MaxEnt model support and also further the results from the statistical analysis. In the statistical analysis for Table 1, we compared participants' responses to those expected from a participant choosing randomly. We saw, compared to this null distribution, that participants' choices tended to encompass the *maxsum*, *maximin*, and *IA* metrics across the "Nominal," "Robot," "Robot Friends," and "Robot Strangers" conditions. While the results from that analysis suggested that participants' responses slightly more matched the responses expected from a *maximin* policy (indicated by the higher *z*-scores under the *Maximin* comparison) compared to other metrics' policies, this comparison was indirect (participant and random responses were compared along each of the metrics, and then the metrics were compared, rather than directly comparing the relative influence of each metric in participant responses). To more directly compare which metrics best described participants' responses, we examine the results from the MaxEnt model designed for this purpose. The results from the MaxEnt model for these conditions clearly differentiate the *maximin* metric as more descriptive of participants' responses compared to the other metrics.

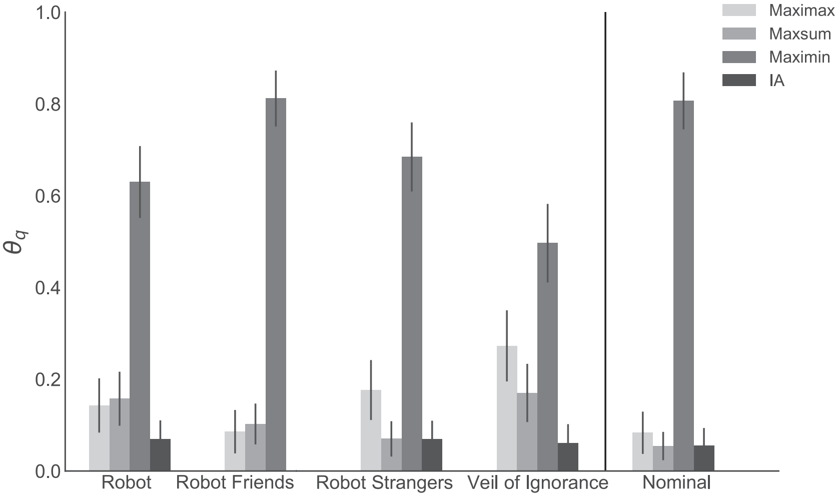


Fig. 2. MaxEnt model results for Experiment 1, showing mean inferred weight θ ± *SE* for each metric across participants in the "Nominal," "Robot," "Robot Friends," "Robot Strangers," and "Veil of Ignorance" conditions. The *maximin* metric best described participant behavior.

For the “Veil of Ignorance” condition, the MaxEnt results were more evenly split between the *maximin* and *maxsum* metrics: $\theta_{maximin} = 0.52$, $\theta_{maxsum} = 0.48$, and $\theta_{maximax}$, $\theta_{IA} = 0$. Of note, while the MaxEnt results emphasized the *maximin* and *maximax* results in Fig. 2, the results from fitting a single θ emphasize the *maximin* and *maxsum* results. This pattern could arise if there were some participants who behaved very consistently according to the *maximax* metric, leading to their being highly represented when θ values were calculated for each participant and averaged, but these participants’ behavior washing out when all participants’ behavior was aggregated. It is noteworthy that the *maximax* and *maxsum* metrics were highly correlated (the same is true for the *maximin* and *IA* metrics) such that if participants were commonly making the best *maximax* choices, these choices were likely to be well-described by the *maxsum* metric as well (Fig. 3, leftmost). This correlation is how there could be a result for the single θ providing more support for the *maxsum* metric with an average mean providing more support for the *maximax* metric. The shared emphasis on the *maximin*, *maxsum*, and *maximax* metrics in the “Veil of Ignorance” condition is also evident from the statistical analysis described above. In sum, across each of these result measures behavior of participants in the “Veil of Ignorance” condition appears to be described by the *maximin*, *maxsum*, and *maximax* metrics.

The MaxEnt model results were designed to isolate which metric best described participants’ behavior by accumulating evidence over all trials. As such, the MaxEnt model seems to clearly indicate participants were always behaving according to, for example, the *maximin* metric. However, in each individual trial, participants could not behave according to an isolated metric, since each choice they made provided some evidence for each of the metrics. This inseparability—caused by the correlations among metrics, see Fig. 3—is what motivated our use of the MaxEnt model. One could argue that participants may be making choices that were most supporting a single metric on each trial, and this is possible but not what was empirically observed. Though we do not include trial-by-trial data, Table 2 provides a histogram of the metrics participants’ behavior most adhered to, according to the MaxEnt model, and the Supplemental Results (Tables S1–S14) show each participant’s adherence to each the metrics according to statistical analyses. Participant behavior generally was in accordance with several of the metrics, and most in accordance with the *maximin* metric in aggregate.

After constructing the MaxEnt model, we wanted to check that our weight vectors θ generalized and that the model was predictive of human behavior. To that end, we estimated participants’ weight vectors and used these to predict held-out choices. Specifically, given participants’ weight vectors θ, we could predict participants’ choices given a new set of options in matrix M_m. The predictions o_m^k were calculated by $\text{argmax}_j P(o_m^j|\theta)$. We calculated weight vectors θ for 50% of participants, each trained on 50% of the matrices. We then used the average weight vector from these participants to predict the remaining 50% of participants’ choices on the remaining 50% of matrices and report this prediction accuracy. As a comparison, we also report the accuracy of prediction when each participant’s weight vector, trained on all matrices, is used to predict that same participant’s choices on each matrix. We report the average of 100 training runs for each prediction.

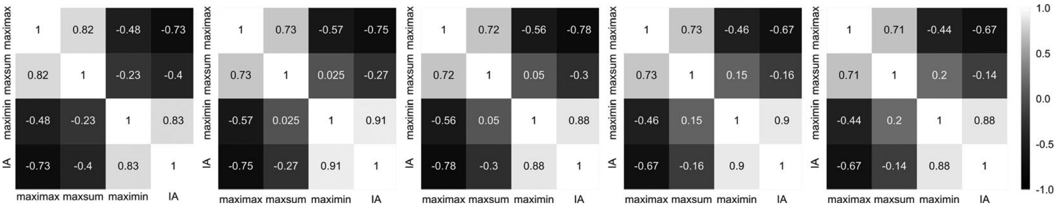


Fig. 3. Correlations across values of the metrics in the “Nominal” (leftmost), “Follow-Up (2x)” (left), “Follow-Up (3x)” (center), “Repeated 2x: Summed” (right), and “Repeated 3x: Summed” (rightmost) conditions. For each condition, we computed the value of each metric for all possible choices across all matrices. We concatenated all of the possible choices across all matrices into one vector for each metric, and we computed the Pearson correlation coefficient across metrics. Choices well-described by the *maximax* metric were also well-described by the *maxsum* metric (high correlation between *maximax* and *maxsum*), while choices well-described by the *maximin* metric were also well-described by the *inequality aversion (IA)* metric (high correlation between *maximin* and *IA*). These clusters (*maximax* and *maxsum*) and (*maximin* and *IA*) tended to be anti-correlated. Note that the “Nominal” condition plot also describes the “Robot,” “Robot Friends,” “Robot Strangers,” “Veil of Ignorance,” “Repeated 2x: Independent,” and “Repeated 3x: Independent” conditions, since participants saw the same stimuli. The summed possible choices used in the “Repeated 2x: Summed” and “Repeated 3x: Summed” correlation matrices were used only for data analysis and not shown to the participants.

Table 2
Percentage of participants whose behavior is best described by each metric, according to the MaxEnt model

	Maximax	*Maxsum*	*Maximin*	*IA*
Nominal	8	3	83	6
Robot	14	17	63	6
Robot Friends	9	9	83	0
Robot Strangers	18	6	71	6
VoI	27	18	48	6
Rep2x: Ind.(C1)	29	12	58	0
Rep2x: Ind.(C2)	25	21	46	8
Rep3x: Ind.(C1)	14	10	62	14
Rep3x: Ind.(C2)	19	29	38	14
Rep3x: Ind.(C3)	29	5	48	19
Rep2x: Summed	0	12	87	0
Rep3x: Summed	10	5	71	14
Follow-Up (2x)	3	6	90	0
Follow-Up (3x)	17	0	83	0

Note. Each participant’s final θ vector according to the MaxEnt metric was computed for each condition; the highest-value θ_q value was taken as the metric that best described the participant’s behavior. Shown is the percentage of participants for which each metric best described their behavior. These more individual-level results closely resemble the aggregated results shown in the Figures. Condition names are shortened: “VoI” represents the “Veil of Ignorance” condition and “Rep2x: Ind.(C1)” represents the “Repeated 2x: Independent (Choice 1)” condition.

Training and testing on 50% participants and matrices, the “Nominal” condition had 72.9% predictive accuracy (non-held-out comparison: 75.4%); the “Robot” condition had 66.4% predictive accuracy (non-held-out comparison: 70.1%); the “Robot Friends” condition had 71.3% predictive accuracy (non-held-out comparison: 70.9%); the "Robot Strangers" condition had 64.6% predictive accuracy (non-held-out comparison: 73.3%); the "Veil of Ignorance" condition had 59.5% predictive accuracy (non-held-out comparison: 78.2%). Chance accuracy was 25%. We observed high predictive accuracy both when testing sets were and were not held out compared to chance accuracy, supporting the validity of our inferred θ vectors.

2.2.3. What metric was preferred across conditions?

We also asked whether participants' judgments of fairness would differ based on what type of agent they were considering, and whether the participant was considered a drink recipient. We found that on the whole, participants were consistent in their behavior across conditions: whether they were thinking about a human manager or a robot manager, whether the robot manager was serving friends or strangers, and whether they were to receive a drink. Specifically, participants' preferred metric did not differ significantly between the "Nominal," "Robot," "Robot Friends," "Robot Strangers," and "Veil of Ignorance" conditions (chi-squared test over $\mathcal{F}_q$ summed over participants and matrices: $\chi^2 = 6.67$, $p = .88$, $df = 12$). We could have generated many more conditions, but our results suggest that participants' fairness judgments are robust to at least some changes in agent identity, and generally that participants' ideas of fairness are similar across varied situations.

The result of similar behavior across conditions is important, because in this work, our goal is to examine humans' views of what decisions decision-makers and assistive artificial intelligences should implement in the future. In this experiment we asked participants what a desirable decision would be and then checked that the results were not a specific consequence of slight differences in how we could have asked the question. There were indeed no significant differences in participants' reports of what they considered a good decision, despite differences in the agent considered. This finding is important with respect to the generalizability of our findings and how we might outsource decisions to technology in the future.

While results were similar across conditions, it is worth noting that the results from the "Veil of Ignorance" condition subtly differed from the rest, for example by encouraging behavior more closely matching the *maxsum* metric. In the "Veil of Ignorance" condition, participants seemed to engage with the question as a new context: Qualitatively, participants seemed to think less about fairness and more in the sense of "gifts" or "gambling." When asked to justify their choices, participants in the "Veil of Ignorance" condition would say things like "one of us might as well be very happy," "I don't know what guest I am so I made the safest choice," or "Neither of us will get something we don't enjoy at all." Whereas in the other conditions, participants tended to more often use words like "fair," "balanced," or "equal." A possible explanation for why behavior in the "Veil of Ignorance" condition was not so closely matched to the *maximin* metric as was true in the "Nominal," "Robot," "Robot Friends," and "Robot Strangers" conditions is this apparent difference in framing. On the one hand (corresponding to the "Veil of Ignorance" condition), the participant could either gamble to win a desired item or offer it generously to an opponent, or alternatively play it safe with a drink everyone would like a little (Konow & Schwettmann, 2016) reviews the evidence on whether "fair" behavior is actually risk-averse behavior). On the other hand (corresponding to the other conditions), the participant would be responsible for the outcomes of two people one does not know, perhaps encouraging more “fair” allocation strategies.

While participants’ behavior was still more *maximin*-aligned than *maxsum*-aligned, the fact that participants’ behavior was relatively more similar to the *maxsum* metric in the “Veil of Ignorance” condition was interesting because the participants could have taken the opposite perspective. In the “Veil of Ignorance” condition, the participants had (unknown) stake in the proceedings and could have made sure not to end up with the unfair end of a bargain by engaging in even more *maximin* behavior relative to the other conditions. Instead, participants behaved relatively more according to the *maxsum* metric. This result is important and should be further explored, since most of us are recipients and not the decision-maker in many real-life situations, perhaps making the “Veil of Ignorance” condition the most descriptive of real situations.

Various other studies have empirically evaluated preferences under the veil of ignorance (Andersson & Lyttkens, 1999; Carlsson, Gupta, & Johansson-Stenman, 2003; Frohlich, Oppenheimer, & Eavey, 1987; Johansson-Stenman, Carlsson, & Daruvala, 2002; Oleson, 2001). Of the studies that evaluated metrics similar to ours, Frohlich et al. (1987) found that participants preferred maximizing average income with a floor constraint, and Oleson (2001) found that participants demonstrated both *maximin* and *IA* behavior—however, this collection of studies was different enough from our paradigm (e.g., studying risk aversion, or looking at welfare over an entire hypothetical society) that the different contexts likely significantly affected outcomes. A study by Bosmans and Schokkaert (2004) was similar to our work in that they asked participants about their preferences directly, and also under a veil of ignorance. They observed different results between the directly elicited preferences and veil of ignorance conditions (but not maximally different results, which occurred in comparing directly elicited preferences to a third self-interested condition). These results were similar to ours, as we found a small difference in response profiles between our “Nominal” and “Veil of Ignorance” conditions. A final aspect of our “Veil of Ignorance” condition is that since our question was about autonomous third-party decision-makers with no stake in the game, we designed the condition such that participants would not receive a payout according to their choices. If participants had a stake in the game (albeit an unknown one), their behavior might have shifted to be more conservative under the assumption of maximizing self-interest—it is known that participants do change their behavior based on whether they are being paid or not (e.g., Gächter & Riedl, 2006; Herrero et al., 2010). Altogether, the “Veil of Ignorance” results were similar to those of the other conditions, but future work will have to continue evaluating the contexts that influence participants’ conceptions of good decision-making.

3. Experiment 2: Repeated choices

Previously, we evaluated what people thought a decision-maker should do when taking a single action. However, in the real world, decision-makers will act many times: the schoolteacher choosing a new lesson plan for their students every day, or the government delivering many aid programs over time. Here we investigated whether people have different intuitions for what decision-makers should do when facing repeated decisions with the same set of agents. Perhaps the decision-maker should give one agent its best choice the first time, and a different agent its best choice the second; or perhaps the conclusion is to compromise each time or to attempt to maximize fairness across the sum of all decisions. We tested whether different metrics would describe participants' actions when participants were given repeated choices.

Repeated decision-making has been studied in the literature, often under the name of sequential or dynamic choices in game theory games. In the repeated version of the Prisoner's Dilemma, participants choose whether to betray their partner or cooperate on each turn, and there has long been research into the optimal strategy for this game (Andreoni & Miller, 1993; Boyd & Lorberbaum, 1987) and other competitive/cooperative games (Kuzmics, Palfrey, & Rogers, 2014), wherein each partner obtains more information about the other on each turn. Behavior in repeated games is sensitive to anticipated repetitions: game outcomes are empirically different when participants are engaging in a finitely repeated game compared to a repeated one-shot game (Andreoni & Miller, 1993). In other sequential games, the sequence is not of both agents making choices simultaneously and then repeating the game, but rather of agents taking turns right after each other, for example by claiming an item via sequential allocation (Bouveret & Lang, 2011; Kalinowski, Narodytska, & Walsh, 2013), or fair queueing (Demers, Keshav, & Shenker, 1989; Moulin & Stong, 2002). There are also "online" games, in which agents may arrive at different times (e.g., Kash, Procaccia, & Shah, 2014; Walsh, 2011) and resources must be divided between them repeatedly. Alternatively, items may arrive over time to the same set of agents who have preferences over them (e.g., Aleksandrov, Aziz, Gaspers, & Walsh, 2015), which is more similar to the structure of our experiment.

There is a subset of research studying the situation of a single item being distributed to a set of agents repeatedly. In this case, researchers are often testing optimum strategies for allocation that maximize total agent utility, while ensuring that their strategy has useful qualities like being efficient and fair, and encouraging truth-telling of preferences from each agent on each round. This work can fall under the heading of "dynamic mechanism design" and has been widely investigated (e.g., Bergemann & Valimaki, 2006; Cavallo, 2008; Guo, Conitzer, & Reeves, 2009). Other researchers have argued that these paradigms do not capture the structure of real-world problems, and so introduced a real-life food distribution problem in which a central decision-maker must repeatedly allocate food to different charities in an online fashion over complex preferences where fairness and efficiency are both considered (Aleksandrov et al., 2015; Walsh, 2015). Our Experiment 2 is similar to this food distribution problem in that we have a repeated task in which a central decision-maker makes resource choices over the preferences of several agents, and similar to the dynamic mechanism design studies in that there is a single item being repeatedly allocated. Our Experiment 2 is different, however, in that the single item is not being allocated to a single recipient, but rather being created and shared across agents on every round.

Freeman, Zahedi, and Conitzer (2017) have perhaps the most similar motivation to our experiment, as they ask if a central decision-maker should make allocations that are optimized for fairness within every round, or if fairness should be optimized across rounds. Freeman et al. (2017) analyzed (non-empirically) fair social choice in dynamic settings with many agents with changing preferences over multiple goods. They chose to optimize for the overall product of all agents' utilities and investigated strategies from there. However, the question of whether real participants would optimize for global utilities across choices is undetermined, including whether they would optimize for additive utilities (the global *maxsum* solution) or the product of utilities (more similar to *IA*). Given previous research, we hypothesized that participants would behave differently when presented with the same choice repeatedly; in other words, we expected that the choices from Experiment 1 would not be repeated three times. We were interested in determining whether participants' choices would be described by the same metric in each round, or across all rounds, and whether these choices would change depending on the number of anticipated rounds. In summary, in Experiment 2 we presented participants with two or three repetitions of the same payoff matrices to determine how they would act, as central decision-makers, in the common real-world situation of providing a resource multiple times to the same group of people.

3.1. Method

3.1.1. Participants

Participants with U.S. IP addresses were recruited from Amazon Mechanical Turk for two additional conditions: "Repeated 2x" (n = 24, 0 participants excluded) and "Repeated 3x" (n = 21, 4 participants excluded). Participants were paid between $2.50 and $3.00 for their participation. Participants were excluded if they failed the included attention check or indicated that they did not understand the experiment.

3.1.2. Stimuli and procedure

In Experiment 2, the procedure and stimuli were similar to Experiment 1, except instead of viewing each matrix once, participants saw each matrix twice (in the "Repeated 2x" condition) or three times (in the "Repeated 3x" condition). Matrices were repeated an even ("Repeated 2x" condition) and odd ("Repeated 3x" condition) number of times to probe how participants would balance their choices. Procedurally, participants read the prompt from the "Nominal" condition, made a choice for the presented matrix, and justified their answer, as in Experiment 1. Then they saw the following prompt: "Your guests have finished the drinks, and you can now put out another. Which drink would you like to serve?" and were shown the same matrix again, and asked to make a choice and justify their answer. The latter prompt and matrix appeared once more in the "Repeated 3x" condition. Participants were able to scroll up and down the page of prompts and responses to determine the total number of drinks they were serving, and could change their choices at any time within the round.

3.1.3. Analysis

In Experiment 2, participants had the opportunity to serve multiple drinks. In the "Repeated 2x" condition they chose an option o^k from matrix $\mathcal{M}_m$ two times (o^k_1, o^k_2), while in the "Repeated 3x" condition they chose an option o^k matrix $\mathcal{M}_m$ three times (o^k_1, o^k_2, o^k_3).

Unlike in Experiment 1, choices from each repeated matrix were not independent given the matrix $\mathcal{M}_m$. We thus calculated two variants on individuals' preferred metrics. For the naïve analysis, we analyzed each choice in the repeated case as an independent decision. Specifically, we assigned all choices after o^k_1 to new "independent" participants. Thus, in the "Repeated 2x" condition, the number of participants artificially doubled, with the first set of participants making choices o^k_1 for each matrix, and the second set of participants making choices o^k_2 for each matrix. $\mathcal{F}^{indep}_q$ was calculated and predictions were made using the same procedure as in Experiment 1, where "*indep*" indicates the artificial creation of independent participants.

For the more sophisticated analysis, we assumed that participants were making one large decision across two or three unordered matrices, rather than making semi-independent decisions o^k_1, o^k_2, (o^k_3). We hypothesized that $\mathcal{U}\left(o^k_1 + o^k_2 + o^k_3\right)$ would not be equivalent to $\mathcal{U}\left(o^k_1\right) + \mathcal{U}\left(o^k_2\right) + \mathcal{U}\left(o^k_3\right)$ (the assumption that choices were independent). We thus mapped the repeated choices to a more sophisticated non-repeated choice: one between all 10 (for the "Repeated 2x" condition) or 20 (for the "Repeated 3x" condition) combinations of individual choices o^k_1, o^k_2, (o^k_3) participants could have made. We then calculated $\mathcal{F}^{summed}_q$ based on the summed agent utilities across all of these potential combination of matrices. Order of choices was not taken into account, but repetition of the same choice o^k for o^k_1, o^k_2, (o^k_3) was allowed. In short, by following this summing procedure, in the "Repeated 2x" condition, participants had 10 hypothetical choices, rather than the 4 they actually faced (two times). In the "Repeated 3x" condition, participants had 20 hypothetical choices, rather than the 4 they actually faced (three times). The analysis proceeded in the same way as for the "Nominal" condition in Experiment 1 except that the matrices $\mathcal{M}_m$ expanded to contain 10 or 20 choices.

3.2. Results and discussion

In Experiment 2, we wanted to test how participant behavior would change across repeated conditions. One hypothesis was that participants would, as in Experiment 1, use the *maximin* metric for each individual choice, not keeping in mind the overall structure of the repeated choices. A second hypothesis was that participants would maintain *maximin* behavior when considering their combined choices, but would engage in different behavior (e.g., suboptimal behavior with respect to the *maximin* metric) within each individual matrix. A third hypothesis was that participants could have chosen to maximize the utility for one agent, then the other, alternating the optimal *maximax* options for each individual choice. This third hypothesis was not borne out in the data and so will not be further discussed.

Results especially supported the second hypothesis, that participants may not have acted independently according to the *maximin* metric on each choice, but rather made sure that the overall outcome after all actions was a *maximin*-supporting outcome.[2] To test this, we evaluated the metrics' values over participants' combined choices ($\mathcal{F}_q^{summed}$). Statistically, we compared participants' empirical $\mathcal{F}_q^{summed}$ values to 10,000 samples of $\mathcal{F}_q^{summed}$ values from randomly generated choices, where the values were summed over participants and matrices. This comparison yielded *z*-scores and *p*-values relating the empirical results to the null distribution for each metric. (The statistics here are the same as explained in Experiment 1.) For both the "Repeated 2x: Summed" and the "Repeated 3x: Summed" conditions, the *z*-scores for the *maximin* metric were higher than for any other metric (Table 1), and they were in the same general range as those of the "Nominal" condition. The statistical analysis thus supported the conclusion that participants' behavior could be explained by applying the *maximin* metric across the set of repeated matrices they were reasoning over. We also analyzed the results through the MaxEnt model. We found that the degree to which the *maximin* metric explained behavior in the "Repeated: Summed" conditions, both for "Repeated 2x: Summed" (two repeated choices) and "Repeated 3x: Summed" (three repeated choices), was high and almost equivalent to in the "Nominal" condition (Fig. 4).

Results also supported the first hypothesis to a lesser extent: that participants' behavior could be described by the *maximin* metric within each individual choice they made throughout the repeated choices. Here, we analyzed the data in terms of individual choices, completing a statistical analysis on the summed empirical and null distributions for the $\mathcal{F}_q^{indep}$ values. *Z*-scores were highest for the *maximin* metric for almost all conditions ("Repeated 2x: Independent (Choice 2)," "Repeated 3x: Independent (Choice 1)," "Repeated 3x: Independent (Choice 2)," and "Repeated 3x: Independent (Choice 3)") (Table 1). The exception was that the *z*-score for the "Repeated 2x: Independent (Choice 1)" condition was highest for the metric *maxsum*. To further analyze our data, we used the MaxEnt model to infer the combination of metrics that best described participant behavior over individual choices. We found that the *maximin* metric best described participant behavior for all of the conditions assuming independence: "Repeated 2x: Independent (Choice 1)," "Repeated 2x: Independent (Choice 2)," "Repeated 3x: Independent (Choice 1)," "Repeated 3x: Independent (Choice 2)," and "Repeated 3x: Independent (Choice 3)" (Fig. 4). The first hypothesis that participants would act according to the *maximin* metric for each of their individual choices was thus supported, though not uniformly across conditions, and contribution from a combination of metrics was visible within the MaxEnt results.

Supporting the importance of the *maximin* metric in describing behavior, estimating a single θ across all participants resulted in values dominated by the *maximin* metric for all but one condition in Experiment 2. Specifically, for the "Repeated 2x: Independent (Choice 1)" condition, $\theta_{maxsum} = 1$ and $\theta_{maximin}$, $\theta_{maximax}$, $\theta_{IA} = 0$, the sole condition where the *maxsum* metric was ranked highest. For the "Repeated 2x: Independent (Choice 2)" condition, $\theta_{maximin} = 1$ and $\theta_{maximax}$, θ_{maxsum}, $\theta_{IA} = 0$. For the "Repeated 2x: Summed" condition, these averages were $\theta_{maximin} = 0.81$, $\theta_{maxsum} = 0.19$, and $\theta_{maximax}$,

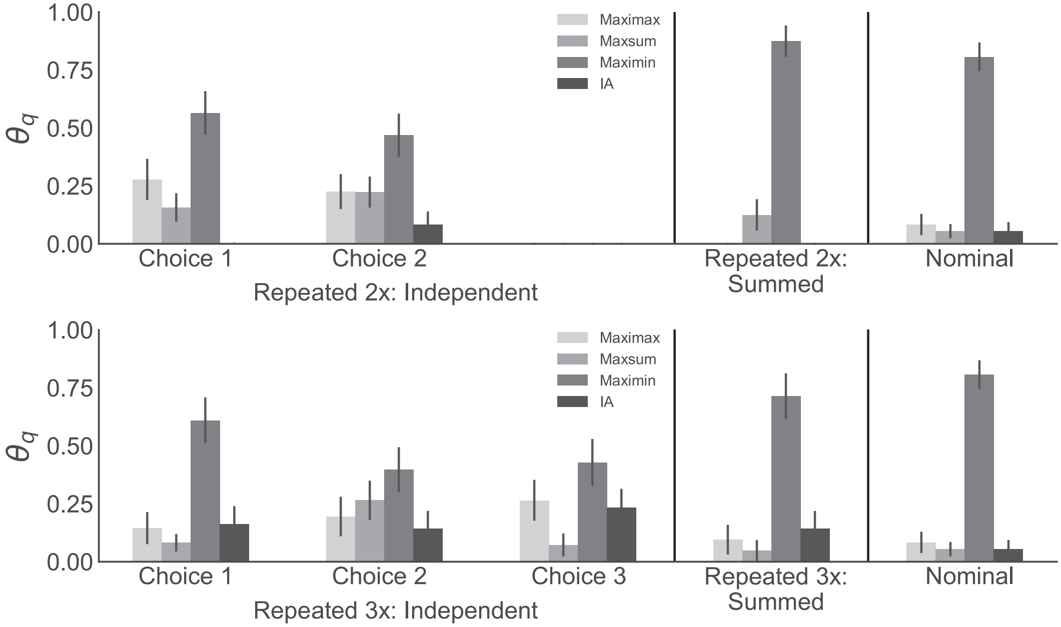


Fig 4. MaxEnt model results for Experiment 2, showing mean inferred weight $\theta \pm SE$ for each metric across participants. (Above) Participants in the "Repeated 2x" condition, and "Nominal" condition. (Below) Participants in the "Repeated 3x" condition, and "Nominal" condition. From left to right: θ_q for each choice in the repeated condition (choices were treated as independent); θ_q for summed choices in the repeated condition, where new matrices were calculated according to $\mathcal{F}_q\left(o_1^k + o_2^k + ...\right)$; and θ_q from the "Nominal" condition.

$\theta_{IA} = 0$. For the "Repeated 3x: Independent (Choice 1)" condition, these averages were $\theta_{maximin} = 0.92$, $\theta_{maxsum} = 0.08$, and $\theta_{maximax}$, $\theta_{IA} = 0$. For the "Repeated 3x: Independent (Choice 2)," "Repeated 3x: Independent (Choice 3)," and the "Repeated 3x: Summed" conditions, these averages were $\theta_{maximin} = 1$ and $\theta_{maximax}$, θ_{maxsum}, $\theta_{IA} = 0$.

This pattern of results suggests that participants consider all of their choices and maintain a running *maximin* metric calculation, though they also seem to apply the *maximin* metric to a lesser degree within each individual choice. Support for this argument draws from the following observation: The dominance of the *maximin* metric in explaining behavior is greater for conditions in which choices were considered cumulatively (the "Repeated: Summed" conditions) than for conditions in which each choice was considered independently (the "Repeated: Independent" conditions). Specifically, the *maximin z*-scores were higher in the statistical analyses and the *maximin* θ values were higher in the MaxEnt model for the "Repeated: Summed" conditions compared to the "Repeated: Independent" conditions, and they were in fact very similar to those in the "Nominal" condition.

Finally, to verify that our weight vectors in the MaxEnt model generalized, we used participants' θ vectors to predict held-out data. Training and testing on 50% participants and matrices, the “Repeated 2x: Independent” condition had 47.0% predictive accuracy (non-held-out comparison: 53.6%, chance accuracy: 25%); the “Repeated 2x: Summed” condition had 39.3% predictive accuracy (non-held-out comparison: 43.3%, chance accuracy: 10%); the “Repeated 3x: Independent” condition had 45.8% predictive accuracy (non-held out-comparison: 57.9%, chance accuracy: 25%); the “Repeated 3x: Summed” condition had 10.9% predictive accuracy (non-held-out comparison: 17.6%, chance accuracy: 5%). Our weight vectors were less accurate in predicting “Repeated: Independent” conditions than they were compared to the non-repeated conditions, but had relatively high predictive accuracy for held-out test sets compared to non-held-out test sets. In the “Repeated: Summed” conditions, as chance accuracy fell, predictive accuracy also fell. With such low predictive accuracy in the “Repeated 3x: Summed” condition especially, we sought to test whether this was an artifact of the many potential choices available to participants in the “Repeated” conditions or a consequence of the utilities of the choices available, motivating Experiment 3.

We observed that in both the “Repeated” conditions and the conditions from Experiment 1, participants’ behavior was best described by the *maximin* metric. The lack of difference in the “Repeated” conditions could be construed as a null result, under the assumption that we did not manipulate the multi-decision conditions strongly enough to cause a change in participants’ responses. To show that participants were producing different behavior in the “Repeated” conditions, but nevertheless overall their behavior could be best described by the *maximin* metric, we show that there is a different pattern of responses for each “Repeated” independent choice condition compared to the “Nominal” condition. Specifically, we computed the percentage of matrices wherein participants had significantly different responses between conditions. We found that that percentage was lower in comparing the “Repeated (Choice 1)” and “Nominal” conditions, and higher in comparing the “Repeated (Choice 2)” and “Nominal” conditions (Table 3). This percentage also changed with the “Repeated 3x: (Choice 3)” and “Nominal” condition comparison (Table 3). These percentage differences indicate that participants were making different choices within each matrix of the “Repeated” conditions, even while the overall results from the MaxEnt model showed continued *maximin* behavior.

4. Experiment 3: Summed repeated choices

When taking repeated actions, participants’ individual choices tended toward being described by the *maximin* metric, but they were described by a combination of other metrics as well. Why were participants’ repeated choices not as clearly described by the *maximin* metric as they were when making a single choice? It could be that participants were attempting to view all of the repeated trial as a single decision, and were trying to maximize the *maximin* solution across all choices, but that the mental overhead for this computation led them to less clearly *maximin* solutions. Alternatively, perhaps computational overhead was not very influential in participants making different choices than they had in Experiment 1, and there was something about the way that the repeated stimuli were presented that led to dissimilar results in Experiment 1 and 2—for example, perhaps participants felt pressure to vary their choices in Experiment 2 when faced with the same questions.

To isolate whether computational overhead was an effect, we presented participants with the summed versions of the repeated choices they saw in Experiment 2. Participants could then see the summed version of choices in Experiment 2, so the computation was easier if they were attempting to sum utilities across decisions. This manipulation also allowed us to present participants with a slightly different one-shot matrix (with six options rather than four) to see if the results from Experiment 1 generalized. We investigated whether participant choices would be similar to those in Experiment 1, individual decisions in Experiment 2, and the artificially summed decisions in Experiment 2, with an emphasis on whether participants' behavior would be more or less clearly described by the *maximin* metric, as we expected from Experiments 1 and 2. In summary, we simplified the multi-decision problem presented before, testing what metrics described participants' solutions when Experiment 2's repeated choices were condensed into a single decision again.

4.1. Method

4.1.1. Participants

Participants with U.S. IP addresses were recruited from Amazon Mechanical Turk across two additional conditions: "Follow-Up (2x)" (n = 31, 0 participants excluded) and "Follow-Up (3x)" (n = 29, 3 participants excluded). Participants were paid between \$2.50 and \$3.00 for their participation. Participants were excluded if they failed the included attention check or indicated that they did not understand the experiment.

Table 3
Results showing differences in participants' choices across conditions

Conditions	# sig.	# inc. total	% sig.
Rep2x: Ind.(C1) vs. Nominal	7	20	35
Rep2x: Ind.(C2) vs. Nominal	12	20	60
Rep3x: Ind.(C1) vs. Nominal	0	17	0
Rep3x: Ind.(C2) vs. Nominal	15	20	75
Rep3x: Ind.(C3) vs. Nominal	9	19	47
Rep2x: Ind.(C1) vs. (C2)	4	19	21
Rep3x: Ind.(C1) vs. (C2) vs. (C3)	7	20	35

Note. Shown is the percentage of matrices ("% sig.") wherein a chi-squared test of independence showed the histograms of participants' choices for each matrix were significantly different (p = .05: uncorrected) across the listed conditions. In more detail: for each matrix, the number of participants who picked each choice was summed, and this set of values was compared across the two experimental conditions listed. If a matrix had any expected value of 0 in the computed χ^2 frequencies, that matrix was removed from the analysis. Note that expected values were often <5. The number of matrices that were significantly different according to the chi-squared test of independence ("# sig.") was divided by the total number of included matrices ("# inc. total"). Note that condition names are shortened: "Rep2x: Ind.(C1)" represents the "Repeated 2x: Independent (Choice 1)" condition.

4.1.2. Stimuli and procedure

The stimuli and procedure were similar to Experiment 1. Participants again viewed each matrix once, but each matrix contained six choices that were drawn from the "summed" utilities from Experiment 2. In Experiment 2, for the "Repeated 2x" experiment, participants had 10 hypothetical choices per round, and in the "Repeated 3x" experiment, participants had 20 hypothetical choices per round. These hypothetical choices (of the form [A's utility, B's utility]) were what we used to generate each matrix in Experiment 3.

To generate each matrix, we first tried to include one choice that maximized each metric. Specifically, we examined all 10 or 20 choices for each round from Experiment 2, and for each metric selected the choice that would be most preferred under that metric (e.g., [4, 4] might be chosen under the *IA* metric, but not the maxsum metric if [8, 4] was also an option in the set). If two metrics had the same best choice (irrespective of order, so [8, 4] was identical to [4, 8]), this choice was only included once. If a metric had several best choices (e.g., under the *IA* metric, [3, 3] would be as good as [4, 4]), then an unused choice was selected at random. (Metrics with one best choice were examined first, and then metrics with multiple choices were examined in random order.) The remaining choices, since six choices were included in each matrix, were selected from the most popular unused choices from Experiment 2.

4.2. Results and discussion

In Experiment 2, we observed that participants appeared to be reasoning over the whole set of repeated choices, and that their behavior was best described by the *maximin* metric across that set ("Repeated: Summed" conditions). To further test this hypothesis, we constructed matrices where participants only had to make one choice, but each choice constituted the sum of previously repeated choices in Experiment 2. Our results corroborated those from Experiments 1 and 2. With these single-choice matrices, participants' behavior could be best explained by the *maximin* metric. In a statistical analysis comparing the empirical choices made by participants to a simulated set of random choices, *z*-scores were highest for the *maximin* metric as compared to the other metrics in the "Follow-Up (2x)" and "Follow-up (3x)" conditions (Table 1). In our MaxEnt model, participant behavior was also best explained by the *maximin* metric in the "Follow-Up (2x)" and "Follow-up (3x)" conditions, to a degree similar as in the "Repeated: Summed (2x)" and "Repeated: Summed (3x)" as well as the "Nominal" conditions (Fig. 5).[3] The similarity of values across the "Follow-Up" conditions and the "Repeated: Summed" conditions support the hypothesis described in Experiment 2, that participants consider the whole set of choices when they make decisions rather than just each choice individually.

Finally, to check that our MaxEnt weight vectors generalized, we used participants' θ vectors to predict held-out data. Training and testing on 50% participants and matrices, the "Follow-Up (2x)" condition had 71.6% predictive accuracy (non-held-out comparison: 71.6%, chance accuracy: 16.7%); the "Follow-Up (3x)" condition had 60.5% predictive accuracy (non-held-out comparison: 66.7%, chance accuracy: 16.7%). We could predict participant responses to the "Follow-Up" conditions well with and without held-out testing sets, indicating the strength of our inferred weight vectors. The high predictive accuracy in the "Follow-Up" conditions compared to the "Repeated: Summed" conditions indicate that the previous low predictive accuracy was a consequence of the large number of possible choices available in the "Repeated" conditions, and the complexity of not seeing the end results directly and having to compute them, rather than a consequence of the utilities of the choices themselves.

5. General discussion

If a schoolteacher is trying to choose a field trip for a group of students with kinetic learners, visual learners, children who do not speak English at home, boisterous students, and students who are scared of new places, what should they do? How should people donate their money, governments choose between aid programs, or assistive robots mediate between family members' preferences? In this work, we examined the broader question of how a decision-maker should act when people have different preferences. Specifically, we asked people what *they* would do in a paradigm where they could take

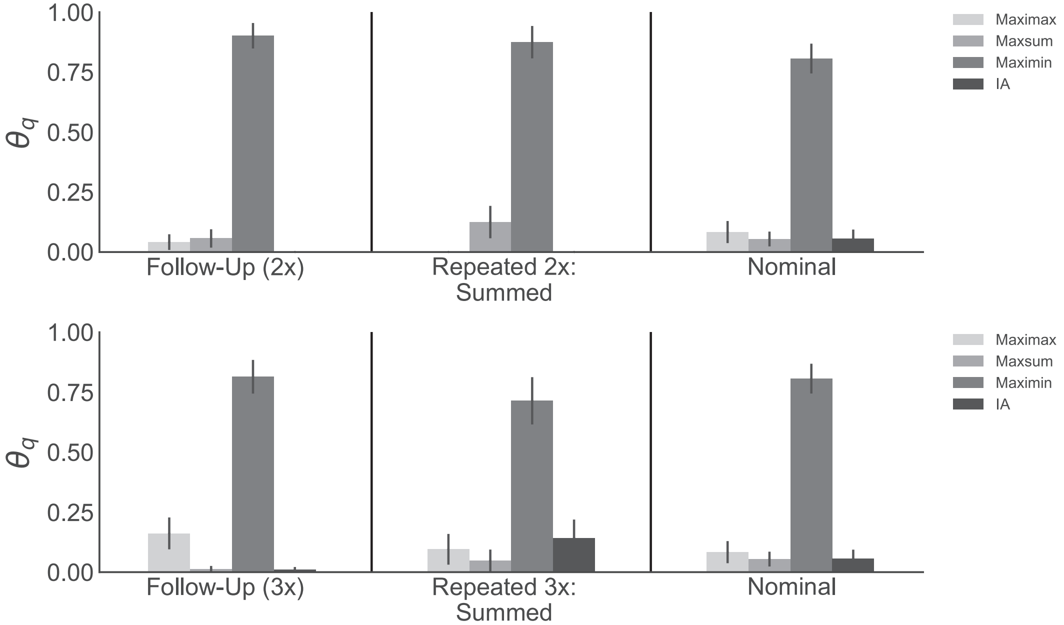


Fig 5. MaxEnt model results for Experiment 3, showing mean weight θ ± *SE* for each metric across participants. (Above) Participants in the "Follow-Up (2x)," "Repeated 2x" (summed choices, where new matrices were calculated according to $\mathcal{F}_q\left(o_1^k + o_2^k + \ldots\right)$), and "Nominal" conditions. (Below) Participants in the "Follow-Up (3x)," "Repeated 3x" (summed choices), and "Nominal" conditions. The *maximin* metric best described participant behavior.

only one action, making a single drink, to the inevitable dismay of some of the people they were serving. While it is not clear that decision-makers should automatically match the behaviors that people intuitively use to make group-level decisions, it is informative to know the ground truth of what people feel are good decisions.

We analyzed participants' choices by assuming their behavior could be captured by a combination of four metrics. Of these metrics—*maximax*, *maxsum*, *maximin*, and *inequality aversion (IA)*—we observed that participants' behavior could be reliably described by the *maximin* metric, the idea of maximizing the utility of the worst-off agent. With respect to our story, the schoolteacher may not know what is the right thing to do, but we at least know how people behave: For each field trip find the child who would have the worst time and choose the field trip wherein that child enjoys the field trip as much as possible.

Participants behaved according to the *maximin* metric despite changes in the agents involved. Specifically, whether considering a robot manager, human manager, or serving friends or strangers, participants' behavior was similar and described by the *maximin* metric (Experiment 1). Participants did tend to move toward behavior more described by the *maxsum* metric when they were told they were beneficiaries of the decision, perhaps reasoning about the task more as a gambling situation and less as one mandating fairness for unknown recipients. In Experiment 2, we found that when participants made repeated decisions, each individual choice therein was characterized by a modest *maximin* preference. However, participants also acted as if they maintained a running total of their choices across repeated decisions. In particular, participants' behavior was best described by the *maximin* metric under the assumption that they were summing utilities across repeated decisions. Experiment 3 provided additional support for the hypothesis that people maintain an overall calculation in repeated decision-making, rather than reasoning over each problem individually. When participants' cumulative choices from Experiment 2 were summed to create a combined set of matrices for Experiment 3, participants' behavior was described by the *maximin* metric to the same degree as was shown in the "Repeated: Summed" and "Nominal" conditions. All of these results suggest that participants keep track of calculations over time and prefer making decisions consistent with the *maximin* metric when allocating indivisible items across agents.

In the remainder of the paper, we discuss the relationship of these results to previous work, and ways in which these results could be extended.

5.1. Relationship to previous work

We asked what a decision-maker should do when it could take only one action—create one resource—to be shared among multiple other agents, and our results support the hypothesis that participants prefer to behave in ways described by the *maximin* metric. How, then, do our results compare to the literature? In the nearby literature of fair allocation, there is not nearly so neat a consensus. However, in most fair allocation studies the participant acts as a self-interested party, weighing the desires of selfishness and "fair" allocation. We are interested in the case of what people consider helpful actions when they have no stake in the outcome. Four studies are similar to our study in that respect.

The first is Herreiner and Puppe (2007), who presented participants with payoff matrices in which agents could receive different “goods.” Unlike our study, in which there was only one item to be created/allocated, the paradigm in Herreiner and Puppe (2007) resembled an estate game in that multiple items, the number of which was often larger than the number of agents, could be distributed to various agents. This added complication to the proceedings, as participants then had large numbers of item combinations to reason over (e.g., their Problem 1 had $3^3 = 27$ different allocations), and they could consider fairness not only over utilities, but over “bundles”—the set (and number) of items allocated. To this end, Herreiner and Puppe (2007) constructed their payoff matrices with an eye to the metric of “envy-freeness,” which describes whether any given agent would be *envious* of any other agent’s bundle.

The results from Herreiner and Puppe (2007) seem to be consistent with a hypothesis that participants prefer acting according to a *maximin* metric, given that in their study this metric was sometimes confounded with different metrics within a single choice. However, in their Problem 1 and Problem 6, Herreiner and Puppe (2007) showed results indicating that participants preferred the *IA* metric over the *maximin* metric. In Problem 1, participants chose to withhold the last item rather than give one agent two items compared to the other one (preferring the two agents to have utilities [49,48] rather than [49,53]), and in Problem 6, participants distributed four items across three agents to maintain exactly equal utilities [45,45,45] rather than choosing the best *maximin* allocation [48,60,52]. Problem 1 is an interesting case in that participants were considering fairness in both utilities *and* item number. Our participants tended to choose according to the *maximin* metric when only considering utilities, but it could be that if we had asked them to reason over both utilities and item number they would have considered choices emphasizing inequality aversion as fairer and better, especially when the difference in utilities was small. In Problem 6, our results suggest that choice complexity could have been influencing the observed results in Herreiner and Puppe (2007), and in future work it would be interesting to check if participants would choose the *IA* solution [45,45,45] rather than the *maximin* solution [48,60,52] if directly presented with those choices, rather than being asked to add utilities from four items across three agents. As a final note on the complexity of addition, Herreiner and Puppe (2007) noted in their discussion that participants’ final solutions were correlated with their reported allocation procedures—for example, the order in which participants assigned items to each agent. Allocation procedures were not inherent to our problem setup but are well-considered within the fair allocation literature (see e.g., Dupuis-Roy & Gosselin, 2011), and perhaps also add implicit utilities to participants’ preferred choices.

In Engelmann and Strobel (2004), the basic structure of the task was very similar to ours, as participants made choices between three items that three agents had utilities over. Though the participant acted as one of these agents, the participant’s utility was always held fixed over all items. Engelmann and Strobel (2004) focused their payoff matrices on distinguishing between two existing models, each of which had one utility function designated as participants’ preferred allocation metric. As such, Engelmann and Strobel (2004) also had many matrices that confounded the metrics we considered, but with this caveat the *maximin* metric could be considered a primary motivation for participants' behavior. Relevantly, Engelmann and Strobel (2004) state that one of the models they evaluated performed well because it captured the *maximin* metric, but that overall a combination of *maximin*, *maxsum* (which they call efficiency), and *selfishness* (which we do not consider) considerations drove participant behavior. Indeed, in several of their payoff matrices, both the best *maxsum* and *maximin* choices were often selected by participants, and the *maxsum* choice often garnered a higher proportion of participants.

Fehr et al. (2006), however, provide a rebuttal to the Engelmann and Strobel (2004) paper, replicating the Engelmann and Strobel (2004) study with non-economics participants and showing that the *maxsum* solution was selected far less by participants who were in different programs. Fehr et al. (2006) did not choose payoff matrices that distinguished between the *IA* and *maximin* metrics, but their results hint that the subject pool can influence preferred allocation metrics. Herreiner and Puppe (2007) tested economics and law student participants, but our study was conducted on Amazon Mechanical Turk and had a broader population than economics undergraduates. In future work, it would be interesting to replicate our study with economics undergraduates, and observe whether this difference is enough to observe participants' preference for the *maxsum* metric over the *maximin* metric. Fairness perceptions have been observed to differ across different populations (see, e.g., Andreoni & Vesterlund, 2001; Croson & Gneezy, 2009; Gaertner & Schwettmann, 2007; Marwell & Ames, 1981; and Camerer, 2011, summarizes some demographic results for behavioral game theory), so this hypothesis would not be improbable, though the review in Konow and Schwettmann (2016) cites that generally demographic variables seem to have relatively small effects on economics experiments.

Yaari and Bar-Hillel (1984) presented several different types of scenarios wherein a third party allocated goods according to the *maximin* metric, *maxsum* metric, and *IA* metric. Participants played three types of games, where they had to allocate fruits according to receiving agents' *needs* or *tastes/utilities*, and also across agents' differences beliefs about the fruits. Participant behavior differed across the varying conditions, shifting according to the tradeoffs presented, as in this work. The authors observed that when choices were presented in terms of needs—one agent needs a certain type of vitamin to be healthy, so they needs a certain type of fruit—82% of participants chose the *maximin* allocation. In our paradigm, agents did not have needs, rather stated utilities (preferences over choices). This was much closer to the second condition in Yaari and Bar-Hillel (1984), in which agents had different "tastes" or stated utilities. Intriguingly, participants did not adhere as strongly to the *maximin* choice in this case, as instead 28% chose the *maximin* solution, and 35% of participants chose the *maxsum* allocation. The distribution of choices also changed in the third condition of Yaari and Bar-Hillel (1984), in which agents had different perceptions of how much value each of the items would give them. Yaari and Bar-Hillel (1984) concluded from their experiments that the *maximin* metric best describes participant behavior, but only when needs are salient.

In our work, we found that participants made choices according to the *maximin* metric outside of needs-based formulations and did so consistently across changes in wording and repetition over time. Our paradigm differed from that of Yaari and Bar-Hillel (1984), however, which could account for the difference in results. The main difference between our work and that of Yaari and Bar-Hillel (1984) is that we asked a different question: Given agents had utilities over a set of four or six options, which option should be chosen to best satisfy those utilities? Yaari and Bar-Hillel (1984) asked a fair allocation question: Given a specific number of items (and agents' preferences over the different items in the "tastes" condition), how should those items be divided between agents? The fair allocation paradigm is necessarily zero-sum, wherein if one agent receives a resource, another loses it. In particular, in question 4 from Yaari and Bar-Hillel (1984), participants were asked to divide 12 grapefruit and 12 avocados between two agents. One agent was stated as hating avocados (utility = 0), but would buy grapefruit if they were priced under $1 per pound. Meanwhile, the other agent liked both avocados and grapefruit and would pay for them if they were priced under $0.50 per pound (half the utility of the first agent for grapefruit). With these agent preferences in mind, participants were now expected to divide a fixed number of grapefruit and avocados. Compare this to our study, in which participants decided which drink to give to two agents to share. This task is not zero-sum —if one agent has high utility for a drink, this does not detract from another agent's utility for that drink. We did, however, have four matrices out of 20 for each experiment in which the joint utilities for each drink were identical, meaning that no matter what drink the participant chose, the agents would only jointly achieve a given happiness. But in this scenario, the question was not "how many of each item should be given to each agent given their utilities" (as in Yaari & Bar-Hillel, 1984), but "given there is a fixed amount of utility to be had, how should that utility be divided between agents?" This is a distinct question, and it is probable that participants reason differently about a zero-sum problem of "distribute 12 pieces of fruit between two people or throw some away" (allocation) compared to "create one shared item that will make two people more or less happy." We may expect that Yaari and Bar-Hillel (1984) did not observe as much *maximin* behavior because they used a task of zero-sum bundle allocation; this hypothesis should be investigated in future work.

Another difference between our study and that of Yaari and Bar-Hillel (1984) was that we probed participants' intuitions of what was fair with 20 questions (4 or 6 choices each) for each experiment, and we determined which metrics best described participants' behavior by accumulating evidence across all of these trials. Alternatively, Yaari and Bar-Hillel (1984) had participants answer one question for each experiment (usually with 5 choices), and we used that single choice to inform which single "mechanism" (analogous to our metrics, but without considering combinations of metrics) best described participant behavior. The authors thus did not have a continuous characterization of how much participants were acting in accordance with different metrics, instead using discrete choices that distinguished "*maximin*," "*maxsum*/utilitarian," and other allocations, which they tested with a single set of choices. With more trials available and a finer-grained measure of how each choice contributed to the various metrics, Yaari and Bar-Hillel (1984) may have observed a more *maximin*-biased distribution, as we did.

As a final aside, with regard to changing in wording, Yaari and Bar-Hillel (1984) found similar distributions of responses when they asked, first, how participants would divide the items, and second, how the two agents would divide the items if the agents were aiming to be just. In our paradigm, we asked only what a third-party decision-maker would do, but this result implies that in our paradigm we would find similar responses if we were asking participants what choices recipients would think were just.

Our study stands in contrast to these four studies in a few ways. First, while our problem formulation can encompass any fairness allocation problem based on our definition of an "action," the specific paradigm we used was specialized to solve the problem of what single item a decision-maker should choose to distribute to a group (a rather straightforward action). This stands in contrast to fair allocation studies, and we described a direct comparison of our study and Yaari and Bar-Hillel (1984) on this axis, with a particular eye to zero-sum choices. Additionally, some differences in our results may be attributable to experimental simplicity. For example, in Herreiner and Puppe (2007), participants may entertain additional implicit utilities when they reason over bundles (such as item number fairness), and there may be difficulties in balancing many possible choices while considering different allocations of 3 to 4 goods. Similarly, Yaari and Bar-Hillel (1984) introduce motivational features not present in our work, like division based on agent needs. A second methodological distinction of our study is that the other studies did not use a continuous measure of behavioral alignment to several metrics, instead using discrete comparisons over metrics that largely correlate, which leaves the option open that finer-grained distinctions may change their details of some of their results. Additionally, a significant advantage of our study is that we employed 20 matrices and could have employed more, whereas other studies had fewer than 12 payoff matrices, and these payoff matrices often confounded different metrics due to the authors' different emphases. Finally, we showed consistent participant behavior aligning with the *maximin* metric, rather than emphasizing the involvement of the *maxsum* and *IA* metrics, and observed this finding over repeated conditions, a contrast which has not been previously conducted to the authors' knowledge.

In summary, our study is aimed at a different question than most in the fairness literature; we aim to study how people would prefer that decision-makers act when they can benefit multiple people. This question is not addressed in previous studies, but is most related to other fairness studies examining uninterested third-party (zero-sum) decisions (note that unlike a fair allocation problem, our problem is not zero-sum since a decision-maker is creating a resource for two people with different but not opposing utilities.) Within this previous work, preference for the *maximin* metric has not been universally shown, and the *maximin* metric is often not directly compared with other potential metrics. Relatedly, previous studies often do not account for the correlations among metrics when describing participant choices. Here, we presented many choices to participants, testing their intuitions many times for each question, and then used a MaxEnt model to disambiguate between the overlapping metrics that described each choice. We additionally probed participants' behavior in repeated, longer-term scenarios than have been evaluated for this problem setting. We distinguished participants' choices representative of the *maximin* metric by direct comparison with competitive strategies, across many question formulations to ensure generalization. We thus provide a novel contribution to the question of what people think a decision-maker should do when faced with helping people with unique utilities.

5.2. Future directions

An interesting extension to this work would be to do the full study that includes extreme comparisons. If participants are faced with the choices [4, 80] and [4, 4] (where agent A receives the first number and agent B receives the second), all participants will likely choose [4, 80], because the tradeoff between maximizing the sum and trying to maintain equality between agents is so extreme. These choices could be varied parametrically, gradually making the tradeoffs less extreme (and more difficult to choose between) with respect to different metrics. Our study lies basically at the center of that parametric descent, where the choices are most difficult. It is difficult to choose whether [4, 8] or [5, 5] is better, and behavior according to both the *maximin* and *maxsum* metrics is competitive. In this range, determining whether a participant was acting more according to a *maximin* or an *IA* metric required accumulating evidence across trials, motivating the use of our MaxEnt model. This is because each choice that a participant made could serve a few possible metrics simultaneously, and in Table 1 we observed this by noting that in the raw data, participants' behavior tended to be described by the *maxsum*, *maximin*, and *IA* metrics rather than a single metric. However, our paradigm would work well in evaluating when participants would change their behavior as tradeoffs become more extreme. If a participant acts according to the *IA* metric until the group utility hits a threshold, that participant should then continue acting according to a *maxsum* metric for all the more extreme choices thereafter.

Tradeoffs like those between equity, the *maxsum* metric, and need (which we do not examine here) have been widely compared in previous studies. Work like Ahlert et al. (2013), Charness and Rabin (2002), Engelmann and Strobel (2004), Faravelli (2007), Fehr et al. (2006), Fisman et al. (2007), Konow (2001, 2003), Konow and Schwettmann (2016), Mitchell et al. (1993), Ordoñez and Mellers (1993), Pelligra and Stanca (2013), Schwettmann (2009, 2012), and Skitka and Tetlock (1992) find that participants do trade off between different principles based on the choices presented, as we see in our work when participants make choices in accordance with several of the metrics. The suggested set of experiments would confirm and expand upon our results, as our stimuli were chosen to isolate which metrics participants thought best when choices were most difficult. Moreover, a parametric study examining these tradeoffs would provide a large and systematic dataset to contribute to the empirical literature on this topic.

Another extension to this study could be to consider whether having verbal problem descriptions, or other representations of utilities, would enhance our paradigm. Hurley, Buckley, Cuff, Giacomini, and Cameron (2011) replicated the study by Yaari and Bar-Hillel (1984) with quantitative problem descriptions, verbal problem descriptions, and both combined. In their verbal descriptions, they instructed participants to create an allocation according to a given metric: "Divide the apples in such a way that the total amount of vitamin F obtained by both Jones and Smith together is as large as possible," is an example of a *maxsum* instruction. The authors report that the results from the verbal problem descriptions were consistent with participants not understanding the relationship between the described metrics and their quantitative allocations, and that participants increased *maxsum* behavior. We used quantitative information in the present study; it would be hard to adapt the flexibility inherent in the different choices we presented to participants in a verbal-only description. Hurley et al. (2011) find that verbal and quantitative descriptions together produce results that more closely match quantitative descriptions alone. Because verbal descriptions do not seem to produce an increased understanding, we expect our paradigm works well with quantitative information, though the question of how to represent utilities and the effects of that choice on participant behavior remains an interesting one.

One limitation of this study is that it asks participants to evaluate between two agents only. The fair allocation literature commonly evaluates over two or three recipient agents; we expect our results to also scale to three agents, but we do not know how robust our findings will be at larger group sizes. This question is incredibly important given that large-scale decision-making impacts many people and so should be done in a way endorsed by many. We do not know that our results will scale: As group size increases, utilities become harder for participants (and people in general) to reason over, so we expect different heuristics will emerge. Fortunately, our experimental paradigm would serve as a good platform for this work given that we can easily generate sets of matrices which require participants to make difficult decisions and can analyze participants' responses with various insertable metrics.

Finally, in future work, we hope to expand the generalizability of our findings. Within our paradigm, we observed consistent and robust results which were supported by using many matrices and conditions. Our paradigm differed from most in the fair allocation literature due to its focus on a more general problem: what a decision-maker should do when it can take one action (in this case making a drink) that will be shared among agents with different preferences. This paradigm itself offers many avenues for expansion—there are many other actions a decision-maker could take besides making an item, including choosing a field trip, donating to organizations, or making governmental policy decisions. We should also consider actions that produce choices with negative utilities, since the dynamics at play in, for example, trolley problems or risk or loss aversion will likely lead to different behaviors. Drawing from the fair allocation literature also shows that this paradigm could be made more complex as multi-step actions are introduced (e.g., an action encompassing bundle distribution), or selfishness, need, desert, and other factors are considered. Konow (2003) presents a pluralistic approach, in which the prime considerations when thinking about justice are a sense of equality and need, utilitarinism and welfare economics, equity and desert, and context (other papers that present pluralistic approaches are Cappelen, Hole, Sørensen, & Tungodden, 2007; Deutsch, 1985; Frohlich & Oppenheimer, 1992; Konow & Schwettmann, 2016; Lerner, 1975; Leventhal, 1976) and all will need to be incorporated into a fuller model of what people consider desirable behavior.

This work provides a general problem framework and quantitative model of what people think a third party should do when taking an action that will bring people different utilities. While the problem we address is distinct from that of fair allocation, it is informed by this literature and can contribute to the discussion of desirable descriptions of helpful behavior. Understanding what actions to take when every choice affects the well-being of many people with disjoint preferences has important bearings on the future, whether in designing decision-making algorithms for artificial intelligence systems, quantitatively evaluating how to help individuals make decisions that are better in line with what people consider to be good choices, or even creating accepted assistive robots. This and future work focusing on making our results more generalizable—and determining whether we would like our decision-makers to adopt people's intuitions of preferred choices—will provide important contributions to this problem.

6. Conclusion

Decision-making would be a much easier problem if everyone had the same preferences. Even having everyone agree about what to do about diverging preferences would help reduce the complexity of reasoning over many diverse perspectives. While that is not quite the spirited environment we live in, in today's world, we have an opportunity to make these complex decision-making tasks easier with artificial intelligence systems. Artificial intelligences can optimize any number of people's preferences once they know what they are, and in this work, we develop a quantitative model of people's intuitive preferences for what to do when making difficult decisions to help people.

Determining what decision-makers should do is a topic that philosophers and governors have wrestled with for centuries: How do we think about inequality within individuals and in society? In this work, we ventured into the morass, developing a general problem framework that let us ask people how they would choose one action whose impact would be different for everyone. Deeper questions remain: What are humanity's morals around inequality, and how does what we do compare to how we think we should be? We do not know what to teach our decision-makers yet, and determining the answers to these questions will require significant collaborative effort. Fortunately, we did determine what drinks to serve at the resulting conferences.

Author note

Special thanks to Professor Anant Sahai for suggesting the conditions in Experiment 1, and we thank various members of the Center for Human-Compatible Artificial Intelligence for helpful comments. This work was funded in part by NSF grant 1456709 to T.L.G. and N.I.H. U.C. Berkeley Neuroscience Training Program Grant to V.G.

Open Research badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/gybd4.

Notes

1. The disadvantage of normalizing the values within each metric was that information about metrics that had particularly high or low values for a given matrix was lost. However, we considered this choice better than the alternative unnormalized option, for which there would be no sense of the alternative options participants were choosing between.
2. As an example, if participants in the “Repeated 2x” condition were analyzing the matrix in Fig. 1, to maximize the *maximin* metric on each choice they would choose the [A: 5, B: 8] cup twice, but if they were maximizing the *maximin* metric across all choices, they would choose the [A: 4, B: 11] cup once and the [A: 12, B: 2] cup once. Note that choosing the [A: 5, B: 8] cup twice is tied for the second-best option for maximizing *maximin* overall, so this set of choices would still contribute to the hypothesis that participants were maximizing the *maximin* metric across all choices, just not as strongly as would be true if the participant had chosen the [A: 4, B: 11] cup once and the [A: 12, B: 2] cup once.
3. Our results from estimating a single θ across all participants, $\theta_{maximin} = 1$ and $\theta_{maximax}$, θ_{maxsum}, $\theta_{IA} = 0$ for the “Follow-Up (2x)” and “Follow-Up (3x)” conditions, also support the conclusion that participants’ behavior was best explained by the *maximin* metric.

References

Ahlert, M., Funke, K., & Schwettmann, L. (2013). Thresholds, productivity, and context: An experimental study on determinants of distributive behaviour. *Social Choice and Welfare*, *40*(4), 957–984. https://doi.org/10.1007/s00355-012-0652-8

Aleksandrov, M. D., Aziz, H., Gaspers, S., & Walsh, T. (2015). Online fair division: Analysing a food bank problem. In Q. Yang, & M. Woolridge (Eds.), *Twenty-Fourth International Joint Conference on Artificial Intelligence* (pp. 2540–2546). Palo Alto, CA: AAAI Press/International Joint Conferences on Artificial Intelligence.

Alesina, A., & Angeletos, G.-M. (2005). Fairness and redistribution. *American Economic Review*, *95*(4), 960–980. https://doi.org/10.1257/0002828054825655

Amanatidis, G., Markakis, E., Nikzad, A., & Saberi, A. (2017). Approximation algorithms for computing maximin share allocations. *ACM Transactions on Algorithms*, *13*(4), 52. https://doi.org/10.1145/3147173

Andersson, F., & Lyttkens, C. H. (1999). Preferences for equity in health behind a veil of ignorance. *Health Economics*, *8*(5), 369–378. https://doi.org/10.1002/(SICI)1099-1050(199908)8:5<369:AID-HEC456>3.0.CO;2-Q

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy*, *97*(6), 1447–1458. https://doi.org/10.1086/261662

Andreoni, J., & Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *Economic Journal*, *103*(418), 570–585. https://doi.org/10.2307/2234532

Andreoni, J., & Vesterlund, L. (2001). Which is the fair sex? Gender differences in altruism. *The Quarterly Journal of Economics*, *116*(1), 293–312. https://doi.org/10.1162/003355301556419

Ashlagi, I., Karagözoğlu, E., & Klaus, B. (2012). A non-cooperative support for equal division in estate division problems. *Mathematical Social Sciences*, *63*(3), 228–233. https://doi.org/10.1016/j.mathsocsci.2012.01.004

Austerweil, J. L., Brawner, S., Greenwald, A., Hilliard, E., Ho, M., Littman, M. L., MacGlashan, J., & Trimbach, C. (2015). The impact of other-regarding preferences in a collection of non-zero-sum grid games. AAAI Spring Symposium 2016 on Challenges and Opportunities in Multiagent Learning for the Real World. Palo Alto, CA: The AAAI Press.

Babcock, L., Loewenstein, G., Issacharoff, S., & Camerer, C. (1995). Biased judgments of fairness in bargaining. *American Economic Review*, *85*(5), 1337–1343.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113* (3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Barman, S., & Krishna Murthy, S. K. (2017). Approximation algorithms for maximin fair division. In Association for Computing Machinery (Ed.), *Proceedings of the 2017 ACM Conference on Economics and Computation* (pp. 647–664). New York: ACM.

Beckman, S. R., Formby, J. P., Smith, W. J., & Zheng, B. (2002). Envy, malice and Pareto efficiency: An experimental examination. *Social Choice and Welfare*, *19*(2), 349–367. https://doi.org/10.1007/s003550100116

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Bergemann, D., & Valimaki, J. (2006). *Efficient Dynamic Auctions. Cowles Foundation Discussion Paper No. 1584*. New Haven, CT: Yale Cowles Foundation. https://doi.org/10.2139/ssrn.936633

Bertsimas, D., Farias, V., & Trichakis, N. (2011). The price of fairness. *Operations Research*, *59*(1), 17–31. https://doi.org/10.1287/opre.1100.0865

Bertsimas, D., Farias, V. F., & Trichakis, N. (2012). On the efficiency-fairness trade-off. *Management Science*, *58*(12), 2234–2250. https://doi.org/10.1287/mnsc.1120.1549

Binmore, K. (1994). *Game theory and the social contract. Vol. 1. Playing Fair (MIT Press Series on Economic Learning and Social Evolution)*. Cambridge, MA: MIT Press.

Bolton, G., Brandts, J., Katok, E., Ockenfels, A., & Zwick, R. (2008). Testing theories of other-regarding behavior: A sequence of four laboratory studies. *Handbook of Experimental Economics Results*, *1*, 488–499. https://doi.org/10.1016/S1574-0722(07)00055-8

Bolton, G., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193. https://doi.org/10.1257/aer.90.1.166

Bosmans, K., & Schokkaert, E. (2004). Social welfare, the veil of ignorance and purely individual risk: An empirical examination. *Research on Economic Inequality*, *11*, 85–114.

Bosmans, K., & Schokkaert, E. (2009). Equality preference in the claims problem: A questionnaire study of cuts in earnings and pensions. *Social Choice and Welfare*, *33*(4), 533. https://doi.org/10.1007/s00355-009-0378-4

Bouveret, S., & Lang, J. (2011). A general elicitation-free protocol for allocating indivisible goods. In T. Walsh (Ed.), *Twenty-Second International Joint Conference on Artificial Intelligence* (pp. 73–78). Menlo Park, CA: AAAI Press/International Joint Conferences on Artificial Intelligence.

Boyd, R., & Lorberbaum, J. P. (1987). No pure strategy is evolutionarily stable in the repeated prisoner's dilemma game. *Nature*, *327*(6117), 58. https://doi.org/10.1038/327058a0

Brams, S. J., Edelman, P. H., & Fishburn, P. C. (2003). Fair division of indivisible items. *Theory and Decision*, *55*(2), 147–180. https://doi.org/10.1023/B:THEO.0000024421.85722.0a

Brams, S. J., & Taylor, A. D. (1996). *Fair division: From cake-cutting to dispute resolution*. Cambridge, UK: Cambridge University Press.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, *97*(3), 818–827. https://doi.org/10.1257/aer.97.3.818

Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., & Tyran, J.-R. (2013). Give and take in dictator games. *Economics Letters*, *118*(2), 280–283. https://doi.org/10.1016/j.econlet.2012.10.030

Carlsson, F., Gupta, G., & Johansson-Stenman, O. (2003). Choosing from behind a veil of ignorance in India. *Applied Economics Letters*, *10*(13), 825–827. https://doi.org/10.1080/1350485032000148268

Cavallo, R. (2008). Efficiency and redistribution in dynamic mechanism design. In *Proceedings of the Ninth ACM Conference on Electronic Commerce* (pp. 220–229). New York: ACM.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. https://doi.org/10.1162/003355302760193904

Chmura, T., Kube, S., Pitz, T., & Puppe, C. (2005). Testing (beliefs about) social preferences: Evidence from an experimental coordination game. *Economics Letters*, *88*(2), 214–220. https://doi.org/10.1016/j.econlet.2005.02.009

Cooney, G., Gilbert, D., & Wilson, T. (2016). When fairness matters less than we expect. *Proceedings of the National Academy of Sciences*, *113*(40), 11168–11171. https://doi.org/10.1073/pnas.1606574113

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*(2), 260–281. https://doi.org/10.1016/S0899-8256(03)00119-2

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–474. https://doi.org/10.1257/jel.47.2.448

Croson, R., & Konow, J. (2009). Social preferences and moral biases. *Journal of Economic Behavior & Organization*, *69*(3), 201–212.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. https://doi.org/10.1007/s00199-006-0153-z

Demers, A., Keshav, S., & Shenker, S. (1989). Analysis and simulation of a fair queueing algorithm. *ACM SIGCOMM Computer Communication Review*, *19*(4), 1–12.

Deutsch, M. (1985). *Distributive justice: A social-psychological perspective*. New Haven, CT: Yale University Press.

Dickerson, J. P., Goldman, J., Karp, J., Procaccia, A. D., & Sandholm, T. (2014). The computational rise and fall of fairness. In *Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 1405–1411). Palo Alto, CA: AAAI Press.

Dreber, A., Fudenberg, D., & Rand, D. G. (2014). Who cooperates in repeated games: The role of altruism, inequity aversion, and demographics. *Journal of Economic Behavior & Organization*, *98*, 41–55.https://doi.org/10.1016/j.jebo.2013.12.007

Dubois, D., Fargier, H., & Prade, H. (1996). Refinements of the maximin approach to decision-making in a fuzzy environment. *Fuzzy Sets and Systems*, *81*(1), 103–122. https://doi.org/10.1016/0165-0114(95)00243-X

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298. https://doi.org/10.1016/j.geb.2003.06.003

Dupuis-Roy, N., & Gosselin, F. (2009). An empirical evaluation of fair-division algorithms. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2681–2686). Austin, TX: Cognitive Science Society.

Dupuis-Roy, N., & Gosselin, F. (2011). The simpler, the better: A new challenge for fair-division theory. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 3229–3234). Austin, TX: Cognitive Society Society.

Ellingsen, T., & Johannesson, M. (2001). Sunk costs, fairness, and disagreement. Stockholm School of Economics Working Paper.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, *94*(4), 857–869. https://doi.org/10.1257/0002828042002741

Escoffier, B., Gourvès, L., & Monnot, J. (2013). Fair solutions for some multiagent optimization problems. *Autonomous Agents and Multi-Agent Systems*, *26*(2), 184–201. https://doi.org/10.1007/s10458-011-9188-z

Faravelli, M. (2007). How context matters: A survey based experiment on distributive justice. *Journal of Public Economics*, *91*(7–8), 1399–1422. https://doi.org/10.1016/j.jpubeco.2007.01.004

Fehr, E., Naef, M., & Schmidt, K. (2006). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *American Economic Review*, *96*(5), 1912–1917. https://doi.org/10.1257/aer.96.5.1912

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, *97*(5), 1858–1876. https://doi.org/10.1257/aer.97.5.1858

Fleurbaey, M. (2008). *Fairness, responsibility, and welfare*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199215911.001.0001

Fong, C. (2001). Social preferences, self-interest, and the demand for redistribution. *Journal of Public Economics*, *82*(2), 225–246. https://doi.org/10.1016/S0047-2727(00)00141-9

Freeman, R., Zahedi, S. M., & Conitzer, V. (2017). Fair social choice in dynamic settings. In C. Sierra (Ed.), *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. Marina del Rey, CA: International Joint Conferences on Artificial Intelligence.

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

Frohlich, N., Oppenheimer, J. A., & Eavey, C. L. (1987). Laboratory results on Rawls's distributive justice. *British Journal of Political Science*, *17*(1), 1–21. https://doi.org/10.1017/S0007123400004580

Gächter, S., & Riedl, A. (2006). Dividing justly in bargaining problems with claims. *Social Choice and Welfare*, *27*(3), 571–594. https://doi.org/10.1007/s00355-006-0141-z

Gaertner, W., Jungeilges, J., & Neck, R. (2001). Cross-cultural equity evaluations: A questionnaire-experimental approach. *European Economic Review*, *45*(4–6), 953–963. https://doi.org/10.1016/S0014-2921(01)00119-2

Gaertner, W., & Schokkaert, E. (2012). *Empirical social choice: Questionnaire-experimental studies on distributive justice*. Cambridge, UK: Cambridge University Press https://doi.org/10.1017/CBO9781139012867

Gaertner, W., & Schwettmann, L. (2007). Equity, responsibility and the cultural dimension. *Economica*, *74*(296), 627–649. https://doi.org/10.1111/j.1468-0335.2006.00563.x

Gollwitzer, M., & van Prooijen, J.-W. (2016). Psychology of justice. In C. Sabbagh & M. Schmitt (Eds.), *Handbook of social justice theory and research* (pp. 61–82). New York: Springer. https://doi.org/10.1007/978-1-4939-3216-0_4

Guo, M., Conitzer, V., & Reeves, D. M. (2009). Competitive repeated allocation without payments. In S. Leonardi (Ed.), *International Workshop on Internet and Network Economics* (pp. 244–255). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-10841-9_23

Harsanyi, J. C. (1975). Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American Political Science Review*, *69*(2), 594–606. https://doi.org/10.2307/1959090

Herreiner, D., & Puppe, C. (2007). Distributing indivisible goods fairly: Evidence from a questionnaire study. *Analyse & Kritik*, *29*(2), 235–258.

Herrero, C., Moreno-Ternero, J. D., & Ponti, G. (2010). On the adjudication of conflicting claims: An experimental study. *Social Choice and Welfare*, *34*(1), 145–179. https://doi.org/10.1007/s00355-009-0398-0

Hoffman, E., & Spitzer, M. L. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *Journal of Legal Studies*, *14*(2), 259–297. https://doi.org/10.1086/467773

Hsu, M., Anen, C., & Quartz, S. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science*, *320*(5879), 1092–1095. https://doi.org/10.1126/science.1153651

Huck, S., & Oechssler, J. (1999). The indirect evolutionary approach to explaining fair allocations. *Games and Economic Behavior*, *28*(1), 13–24. https://doi.org/10.1006/game.1998.0691

Hurley, J., Buckley, N. J., Cuff, K., Giacomini, M., & Cameron, D. (2011). Judgments regarding the fair division of goods: The impact of verbal versus quantitative descriptions of alternative divisions. *Social Choice and Welfare*, *37*(2), 341–372. https://doi.org/10.1007/s00355-010-0487-0

Johansson-Stenman, O., Carlsson, F., & Daruvala, D. (2002). Measuring future grandparents' preferences for equality and relative standing. *Economic Journal*, *112*(479), 362–383. https://doi.org/10.1111/1468-0297.00040

Jungeilges, J. A., & Theisen, T. (2008). A comparative study of equity judgements in Lithuania and Norway. *Journal of Socio-Economics*, *37*(3), 1090–1118. https://doi.org/10.1016/j.socec.2007.04.002

Kalinowski, T., Narodytska, N., & Walsh, T. (2013). A social welfare optimal sequential allocation procedure. In F. Rossi (Ed.), *Twenty-Third International Joint Conference on Artificial Intelligence* (pp. 227–233). Menlo Park, CA: AAAI Press / International Joint Conferences on Artificial Intelligence.

Kash, I., Procaccia, A. D., & Shah, N. (2014). No agent left behind: Dynamic fair division of multiple resources. *Journal of Artificial Intelligence Research*, *51*, 579–603. https://doi.org/10.1613/jair.4405

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, *90*(4), 1072–1091. https://doi.org/10.1257/aer.90.4.1072

Konow, J. (2001). Fair and square: The four sides of distributive justice. *Journal of Economic Behavior & Organization*, *46*(2), 137–164. https://doi.org/10.1016/S0167-2681(01)00194-9

Konow, J. (2003). Which is the fairest one of all? A positive analysis of justice theories. *Journal of Economic Literature*, *41*(4), 1188–1239. https://doi.org/10.1257/002205103771800013

Konow, J. (2009). Is fairness in the eye of the beholder? An impartial spectator analysis of justice. *Social Choice and Welfare*, *33*(1), 101–127. https://doi.org/10.1007/s00355-008-0348-2

Konow, J., & Schwettmann, L. (2016). The economics of justice. In C. Sabbagh, & M. Schmitt (Eds.), *Handbook of social justice theory and research* (pp. 83–106). New York: Springer. https://doi.org/10.1007/978-1-4939-3216-0_5

Kuderer, M., Gulati, S., & Burgard, W. (2015). Learning driving styles for autonomous vehicles from demonstration. In A. Okamura (Ed.), *2015 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2641–2646). Piscataway, NJ: IEEE.

Kurokawa, D., Procaccia, A. D., & Wang, J. (2016). When can the maximin share guarantee be guaranteed? In *Thirtieth AAAI Conference on Artificial Intelligence*. Palo Alto, CA: The AAAI Press.

Kuzmics, C., Palfrey, T., & Rogers, B. W. (2014). Symmetric play in repeated allocation games. *Journal of Economic Theory*, *154*, 25–67. https://doi.org/10.1016/j.jet.2014.08.002

Lerner, M. J. (1975). The justice motive in social behavior: Introduction. *Journal of Social Issues*, *31*(3), 1–19. https://doi.org/10.1111/j.1540-4560.1975.tb00995.x

Leventhal, G. S. (1976). *Fairness in social relationships*. Morristown, NJ: General Learning Press.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, *1* (3), 593–622. https://doi.org/10.1006/redy.1998.0023

Marwell, G., & Ames, R. E. (1981). Economists free ride, does anyone else? Experiments on the provision of public goods, IV. *Journal of Public Economics*, *15*(3), 295–310. https://doi.org/10.1016/0047-2727(81)90013-X

Mitchell, G., Tetlock, P. E., Mellers, B. A., & Ordonez, L. D. (1993). Judgments of social justice: Compromises between equality and efficiency. *Journal of Personality and Social Psychology*, *65*(4), 629. https://doi.org/10.1037/0022-3514.65.4.629

Moulin, H., Brandt, F., Conitzer, V., Endriss, U., Procaccia, A. D., & Lang, J. (2016). *Handbook of computational social choice*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781107446984

Moulin, H., & Stong, R. (2002). Fair queuing and other probabilistic allocation methods. *Mathematics of Operations Research*, *27*(1), 1–30. https://doi.org/10.1287/moor.27.1.1.336

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In P. Langley (Ed.), *Proceedings of the 17th International Conference on Machine Learning* (pp. 663–670). San Francisco, CA: Morgan Kaufmann.

Nord, E., Richardson, J., Street, A., Kuhse, H., & Singer, P. (1995). Who cares about cost? Does economic analysis impose or reflect social values? *Health Policy*, *34*(2), 79–94. https://doi.org/10.1016/0168-8510(95)00751-D

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, *289*(5485), 1773–1775. https://doi.org/10.1126/science.289.5485.1773

Oleson, P. E. (2001). *An experimental examination of alternative theories of distributive justice and economic fairness*. Tucson: University of Arizona.

Ordoñez, L. D., & Mellers, B. A. (1993). Trade-offs in fairness and preference judgments. *Psychological Perspectives on Justice: Theory and Applications*, 138–154. https://doi.org/10.1017/CBO9780511552069.008

Pálvölgyi, D., Peters, H. J. M., & Vermeulen, A. J. (2010). A strategic approach to estate division problems with non-homogenous preferences. METEOR Research Memorandum 10/036. Maastricht.

Pelligra, V., & Stanca, L. (2013). To give or not to give? Equity, efficiency and altruistic behavior in an artefactual field experiment. *Journal of Socio-Economics*, *46*, 1–9. https://doi.org/10.1016/j.socec.2013.05.015

Procaccia, A. D., & Wang, J. (2014). Fair enough: Guaranteeing approximate maximin shares. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation* (pp. 675–692). New York, NY: ACM. https://doi.org/10.1145/2600057.2602835

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Belknap Press.

Rawls, J. (1974). Some reasons for the maximin criterion. *American Economic Review*, *64*(2), 141–146.

Salles, R. M., & Barria, J. A. (2008). Lexicographic maximin optimisation for fair bandwidth allocation in computer networks. *European Journal of Operational Research*, *185*(2), 778–794. https://doi.org/10.1016/j.ejor.2006.12.047

Schäfer, M., Haun, D., & Tomasello, M. (2015). Fair is not fair everywhere. *Psychological Science*, 1252–1260. https://doi.org/10.1177/0956797615586188

Schokkaert, E., & Devooght, K. (2003). Responsibility-sensitive fair compensation in different cultures. *Social Choice and Welfare*, *21*(2), 207–242. https://doi.org/10.1007/s00355-003-0257-3

Schokkaert, E., & Lagrou, L. (1983). An empirical approach to distributive justice. *Journal of Public Economics*, *21*(1), 33–52. https://doi.org/10.1016/0047-2727(83)90072-5

Schokkaert, E., & Overlaet, B. (1989). Moral intuitions and economic models of distributive justice. *Social Choice and Welfare*, *6*(1), 19–31. https://doi.org/10.1007/BF00433360

Schwettmann, L. (2009). *Trading off competing allocation principles: Theoretical approaches and empirical investigations* (Vol. *3343*). Frankfurt, Germany: Peter Lang.

Schwettmann, L. (2012). Competing allocation principles: Time for compromise? *Theory and Decision*, *73*(3), 357–380. https://doi.org/10.1007/s11238-011-9289-9

Skitka, L. J., & Tetlock, P. E. (1992). Allocating scarce resources: A contingency model of distributive justice. *Journal of Experimental Social Psychology*, *28*(6), 491–522. https://doi.org/10.1016/0022-1031(92)90043-J

Thomson, W. (2003). Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: A survey. *Mathematical Social Sciences*, *45*(3), 249–297. https://doi.org/10.1016/S0165-4896(02)00070-7

Traub, S., Seidl, C., Schmidt, U., & Levati, M. V. (2005). Friedman, Harsanyi, Rawls, Boulding–or somebody else? An experimental investigation of distributive justice. *Social Choice and Welfare*, *24*(2), 283–309. https://doi.org/10.1007/s00355-003-0303-1

Walsh, T. (2011). Online cake cutting. In R. Brafman, F. S. Roberts, & A. Tsoukiàs (Eds.), *International Conference on Algorithmic Decision Theory* (pp. 292–305). Berlin: Springer. https://doi.org/10.1007/978-3-642-24873-3_22

Walsh, T. (2015). Challenges in resource and cost allocation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 4073–4077). Palo Alto, CA: The AAAI Press.

Wittig, M., Jensen, K., & Tomasello, M. (2013). Five-year-olds understand fair as equal in a mini-ultimatum game. *Journal of Experimental Child Psychology*, *116*(2), 324–337. https://doi.org/10.1016/j.jecp.2013.06.004

Yaari, M. E., & Bar-Hillel, M. (1984). On dividing justly. *Social Choice and Welfare*, *1*(1), 1–24. https://doi.org/10.1007/BF00297056

Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). *Maximum entropy inverse reinforcement learning*. Chicago, IL: AAAI.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Supplementary Material