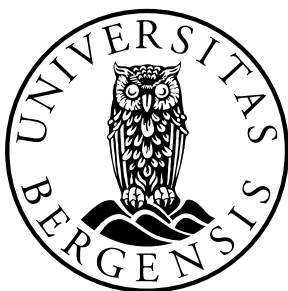


WORKING PAPERS IN ECONOMICS

No. 1/24

Sulagna Dasgupta, Lenka Fiala and
Jantsje M. Mol

For the ‘Greater Good’: Please
Choose A



Department of Economics
UNIVERSITY OF BERGEN

For the ‘Greater Good’: Please Choose A

Sulagna Dasgupta,¹ Lenka Fiala,² and Jantsje M. Mol^{3*}

¹University of Chicago

²University of Bergen

³University of Amsterdam

March 6, 2024

Abstract

How do people trade off individual versus group welfare in the face of uncertainty regarding private benefits of different actions? We propose a partial information revelation (‘recommendation’) policy designed to maximize group welfare, and we show its theoretical robustness to well-documented behavioral deviations from the risk neutral, Bayesian, and self-interested benchmark. In a large-scale online experiment with 2600 subjects, we then show that this policy fails to improve upon a full information benchmark even when individual and group objectives are aligned, as the recommended course of action is not followed often enough. In a setting where individual and group interests clash, the recommendation is followed less often, largely by subjects who misunderstand the policy. This provides suggestive evidence in favor of simplicity in information design in multi-agent strategic settings.

JEL Classification: C78, C91, D82

Keywords: online experiment, matching, imperfect information, information design

*Corresponding author: L. Fiala (l.fiala@uib.no). We thank Yan Chen, Yinghzi Liang, OSub Kwon, Andrew McClellan, Sota Ichiba, Ritwik Banerjee, Alex Imas, Manshu Khanna, Andreas Ziegler, and Jan Potters for very helpful comments. We also thank participants at the M-BEES, KVS New Paper Sessions, ESA World Meetings, Young Economists’ Meeting in Brno, and the Annual Meeting of the Norwegian Economists conferences, the University of Chicago experimental lunch group, Lisbon Microeconomics Group, and seminar attendees at the University of Bergen, Corvinus University of Budapest, and CREED for their feedback. Morgane Bakehe and Augustinas Milius provided excellent research assistance. Funding from the University of Chicago, Nova School of Business and Economics, University of Amsterdam, and University of Bergen is gratefully acknowledged.

1 Introduction

Consider the problem when agents are asked to choose between objects or actions with uncertain private costs or benefits, and their choices influence which options remain available for others. For example, suppose there is a limited supply of vaccines which need to be allocated within a country. Individual factors that may or may not be known (e.g., pre-existing conditions) introduce uncertainty to people's decision whether to take the vaccine, and this decision then affects the availability of the vaccine to others. Suppose further that experts have access to research regarding the disease, and can thus predict private benefits of the vaccine for the individuals with greater precision than the people themselves (e.g., by correctly evaluating risk factors). Finally, similarly to the Covid-19 pandemic, suppose nobody can be forced to receive a vaccine, and a market for the vaccine is not permissible. In such a situation, we ask: How much information regarding private benefits of the vaccine should the experts reveal to the people if their objective is to maximize societal welfare (i.e., make sure the vaccine is taken up by those who benefit the most from receiving it, while others patiently wait their turn)?¹

Building on the matching literature ([Gale and Shapley, 1962](#); [Abdulkadiroğlu and Sönmez, 1998](#); [Bogomolnaia and Moulin, 2001](#)), we employ a model where agents in the society are to be allocated, without monetary transfers, one of two options.² Each option can accommodate a limited number of agents. The allocation process is the following: the agents are randomly sorted into a queue, and then each agent (he) is allowed to pick whatever option he likes among the remaining capacity of each option once agents who are ahead of him in the queue have made their choice. This is known as the *random serial dictatorship* mechanism ([Abdulkadiroğlu and Sönmez, 1998](#)) as each agent at the moment of choosing acts as a dictator regarding his choice. Such a procedure can be thought of as analogous to people waiting for their turn to choose whether to receive a vaccine in a pandemic.³ While each agent knows the distribution of payoffs from the options in the society, he cannot determine

¹The essence of this problem, allocation of scarce objects among imperfectly informed decision makers who could benefit from advice of an expert, extends to many other contexts: For example, a school counsellor can advise parents which school would be a good fit for their child given similar cases the counsellor observed in the past, or a researcher may suggest good fits between research interns and tasks given her superior understanding of the nature of the tasks that need to be performed.

²This makes our model particularly applicable to contexts where markets are impossible, unethical, or impractical to set up. However, as shown by [Azevedo and Leshno \(2016\)](#), matching situations can be modelled like "traditional" markets with demand and supply, allowing for a derivation of comparative statics.

³By using a random ordering of agents rather than allowing the social planner to sort the agents in the queue we are essentially requiring that our mechanism is robust to people attempting to skip the queue, analogous to people in the pandemic using their personal connections to medical professionals to access the vaccine earlier than their designated risk group.

precisely which option is better for him personally. A benevolent social planner (she) knows the best option for each agent - analogous to medical experts identifying risk groups more precisely than the average citizen.⁴ She designs an information policy - a *signal structure*, in the language of information design (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2016) - to release just enough information to each agent, so as to persuade him to pick the option she knows is the best for the group (which may or may not align with what is best for him). This model is a simplified version of the model in Dasgupta (2020).

As that paper shows, in our setting when the goals of individuals and the group are aligned, which happens when agents do not have a strong pre-existing preference for one of the two options, it is theoretically possible to achieve the planner’s *first best* aggregate welfare. This can be achieved by simply recommending each agent to pick the social welfare maximizing option. Each agent follows this recommendation even though it *need not be individually the best* for him ex-post, as it maximizes his *interim* expected payoff, based on the limited information conveyed to him by the recommendation. On the other hand, such an informational intervention is theoretically completely ineffective, i.e., does no better than not sharing any information at all, when individual and group incentives are in conflict. Conversely as before, that happens when agents have strong ex-ante preferences for one of the options.

We evaluate the robustness of these predictions both theoretically and experimentally to known behavioral deviations from the baseline (risk neutral, Bayesian, purely self-interested) model: We show that our proposed recommendation policy remains welfare-maximizing when individual and group interests are aligned even for risk averse or altruistic agents as long as they are not overly prior-biased in their updating following the recommendation. When individual and group interests are not aligned, we specify parameter bounds for which the policy remains welfare-maximizing.

Our experiment is as follows: on Prolific, we randomly divide 2600 participants into groups of four to play the aforementioned allocation game. There are two options for subjects to choose from, each of which can be chosen by precisely two people. Subjects know their own payoff associated with one option, and the probability distribution over own payoffs associated with the other. In addition, the subjects potentially receive more information about the second option, depending on which information treatment they were allocated into: *Full Info*, *Partial Info* or *No Info*. In our main treatment of interest, *Partial Info*, the subjects additionally receive a computerized recommendation on which option to choose. Subjects know that this recommendation is calculated with the intention to allocate people to op-

⁴In our baseline model, the social planner is assumed to have perfect information about the agents’ preferences. However, our results extend to the case where she observes them with noise, as long as the noise is not too large. See the end of Appendix section C.2.

tions such that the aggregate social welfare (here: sum of total payoffs within their group) is maximized. The other two between-subject treatments serve as natural benchmarks: in *Full Info*, the subjects know precisely their payoffs from choosing one or the other option, whereas in *No Info* the subjects receive only the basic information described above, i.e., their own payoff associated with one option, and the probability distribution and corresponding own payoffs associated with the other option.

Subsequently, each subject reports their preferred option to the computer.⁵ The four options are then allocated among the four participants according to the random serial dictatorship mechanism described above. We are interested in how the participants use potentially imperfect information to make their choices. By using within-subject variation of the payoffs associated with the second option, we show the effects of this information setting when individual and group objectives are aligned versus in conflict at the *interim* stage, i.e., once the subject sees their recommendation.

We find that in general, around 71.5% of subjects choose the object they were recommended; this number is higher whenever the recommended option is less risky, or when it is both individually and collectively advantageous to follow the recommendation. However, even in cases when individual and group objectives are aligned, the recommendation-following rate is below that predicted by the theory.

In cases where individual and group interests conflict, we still observe 38.0% of subjects follow a recommendation that is not in their individual interest. Interestingly, this brings the resulting social welfare to the level of what is achievable by simply giving each agent full information about his payoffs, which is more than the level achieved when no information is shared.

Looking at why such a large share of subjects follow a recommendation that favors group over individual welfare when these two are in conflict, we find that this is largely attributable to the subjects' misunderstanding of the game that manifests as mistakes on our comprehension quiz, and, in some cases, can be identified from the subjects' self-reported strategies. However, neither mistakes nor other motives we identify (such as desire to deliberately help one's group at own expense) are frequent enough to induce sufficient compliance with the optimal recommendation, and so the aggregate welfare is not significantly higher in *Partial Info* than in the benchmark case of *No Info*.

This motivates our final point: In a strategic setting, an information policy may be *designed* to be optimal with respect to some objective (such as maximizing social welfare), but agents might misunderstand the policy, resulting in individually sub-optimal choices. Given the observed frequency of these misunderstandings, such

⁵For simplicity, our subjects make their decisions simultaneously and cannot observe the choices of others. As such we therefore do not allow for any learning from others or coordination of actions.

a policy then neither improves the aggregate well-being, nor benefits individual agents. This suggests that simplicity properties of information policies may be worth studying theoretically and empirically as part of the information design toolkit, analogous to similar properties of mechanisms which are well-studied (Li, 2017; Börgers and Li, 2019; Li and Dworzak, 2021).

Related literature. Our results contribute to three broad strands of literature: the experimental literature on matching, information design for policy and the role of advice in strategic environments. First, we provide a novel behavioral extension to a model of matching with information design, and show how empirical regularities such as risk aversion, imperfect Bayesian updating, or social preferences affect decision-making in this setting. Most of the traditional matching experiments literature assumes perfect information on the part of the agents about their own preferences and focuses on comparisons across mechanisms in terms of strategy, stability and welfare, among others (Chen and Sönmez, 2006; Calsamiglia et al., 2010; Klijn et al., 2013; Echenique et al., 2016; Castillo and Dianat, 2016). Hakimov and Kübler (2021) provide a recent survey of this literature, focusing on school choice and college admissions applications. Several papers consider the case of incomplete information about *others'* preferences. For example, Pais and Pintér (2008) and Ding and Schotter (2019) study the impact of agents having varied levels of information about others' strategies on truth-telling and welfare in two-sided matching settings. However, these papers do not consider the case where agents face uncertainty over their *own* payoffs. Closest to our paper is the work of Chen and He (2021), who compare how different school choice mechanisms incentivize students' information acquisition when facing uncertainty about their own preferences. Examples of theoretical works studying similar questions include Immorlica et al. (2020) and Artemov (2021). Also related are Neilson et al. (2019a) who, using a field experiment, show that while more information on the schools shifts parents' choices towards better schools, capacity constraints reduce the positive impacts of this informational intervention. This is fully consistent with our theoretical predictions, though our experiment was not designed to capture this effect.

Second, we contribute to the behavioral and experimental information design literature. While the theoretical literature on information design is now quite rich, with very few exceptions, the notion is understudied in the lab. Relevant works include Aristidou et al. (2019), who compares the mechanism and information design approaches in the lab, and finds the information design approach to be more effective. Incorporating reciprocity into the Bayesian persuasion model, Au and Li (2018) show, using both theory and experiments, that when the prior belief indicates that a good state is more likely, agents are more difficult to be persuaded, implying that the designer's optimal persuasion strategy involves more informative disclosure.

Most of the above literature focuses on the standard Bayesian persuasion set-

ting featuring a single receiver, unlike our paper, which studies information design in *strategic* environments. A notable exception is Ziegler (2023). He studies information design in a game in the lab and focuses on the effectiveness of public vs private signals. In settings where players’ actions are strategic substitutes, the optimally crafted private signal fails to realize its theoretically predicted gains over the public signal, because it is followed less often than the latter. Ziegler (2023) identifies “aversion to complexity and differential treatment” created by the private signal as one of the channels contributing to the unwillingness to follow it. At a high level, this result bears similarities to our finding that while theoretically, potentially different, privately communicated recommendations are supposed to generate the greatest gains, in the lab, the gains from such recommendations are about the same as that from the more transparent signal.

In a separate experiment, Ziegler (2023) also studies sender behavior and finds that it is quite similar to receiver behavior. Examples of other works within the experimental information design literature which do the same include Nguyen (2017); Fr chet te et al. (2022); Kwon (2020). In contrast, our “sender” is computerized and not played by participants.

Finally, this paper also speaks to the experimental literature on the role of advice in strategic environments (Koutout et al., 2021; Zhu, 2015; Guillen and Hakimov, 2018; Masuda et al., 2022; Braun et al., 2014). Closest to our work within this literature is Guillen and Hing (2014), who evaluate the effects of both correct and incorrect advice on participants’ strategies in a matching setting, and find that the rate of following the correct strategy is highest when no advice is given and falls with *both* correct and incorrect advice. Also closely related is Koutout et al. (2021), who experimentally demonstrate the improvements strategic advice can bring about in the same setting. Finally, Braun et al. (2014) study “strategy coaching” in their lab experiment. The focus of much of this literature has been on the role of advice in mitigating the *strategic mistakes* participants may make in these environments. Our paper introduces a novel type of advice to this literature — namely one strategically designed to help the *group*, not necessarily the individual. In fact, the heart of our research problem is the non-trivial task of the participants of figuring out the connection between these two.

2 Theoretical background

Now let us turn to our model.

2.1 Baseline model

We use a standard model of allocating $2n, n \geq 1$, agents between $m = 2$ options, without monetary transfers (see Hylland and Zeckhauser (1979), more recently e.g., He et al. (2018)), where the agents only know the joint distribution of their cardinal preferences over these two options rather than the preferences themselves. Let us call the set of agents $I := \{1, \dots, 2n\}$ and that of the objects $H := \{A, B\}$. Each option has a capacity to accommodate n agents, i.e., the two options in total have exactly enough capacity to accommodate all agents.⁶

Let $u_{i,h}$ denote agent i 's utility from object h . Let u_i denote his utility vector. Each agent's i 's utility vector lies in some compact set $\mathcal{U}_i = [\underline{u}, \bar{u}]^2, \underline{u} \geq 0$. The space of cardinal preference profiles is $\mathcal{U} \equiv \times_i \mathcal{U}_i = [\underline{u}, \bar{u}]^{2 \times 2n}$, with its typical element denoted by u . u_i s are distributed with joint prior probability measure μ over \mathcal{U} . μ is common knowledge.

The agents are to be allocated across the two options using the random serial dictatorship (Abdulkadiroğlu and Sönmez, 1998) mechanism. Under this mechanism, a ranking of all agents is chosen uniformly at random, and then each agent is allowed to pick his favorite option from the remaining options after all agents ranked ahead of him have already made their pick. Mathematically, this mechanism can be represented as a mapping from agents' jointly reported *ordinal* preference profiles - i.e., profiles of their reports of which of the two options they like more, if any - to distributions over joint allocations.

2.1.1 The planner's problem

A benevolent planner or *designer* (she) wants to maximize the society's expected aggregate welfare, which we model as the sum of the expected utilities of the agents. But - in a departure from the standard in the matching literature - she takes the aforementioned allocation mechanism as given.

Instead of tweaking the mechanism, the way the planner achieves her objective is by optimally choosing an *information policy* - a policy about how much information to reveal to each agent about his and others' preferences. Formally, we model the planner as an information designer with full commitment,⁷ who chooses a distribution over a set of *signals* for each realization of the preference profile u .

⁶The theoretical results used in our experimental design hold for a more general model with any finite set of objects and agents, and where options could have varying capacities. We present a simpler version of the model, since that is the set up we use for our experiment. See Dasgupta (2020) for the general model.

⁷The commitment assumption is standard in the information design literature. Empirical evidence suggests that, in school choice contexts, families largely see information about schools shared by authorities as reliable, or true (Neilson et al., 2019b). We see this observation as lending support our commitment assumption. Theoretically, foundations for this assumption in organizational contexts has been provided by Deb et al. (2023).

Because this distribution depends on the actual preference profile, it conveys information about it without necessarily fully revealing it. In our context, the strategy of the designer is to use this partial revelation of information to steer agents towards joint reports she would like them to make to achieve her objective of maximizing aggregate welfare.

For our baseline model we assume that the planner can choose the signal to be as precise as she likes. We discuss in the Appendix (C.2) that our predictions hold even if her designed signal is noisy, as long as the noise is below an upper bound, which we also characterize.

2.1.2 The optimal information policy: The “recommendation” signal

We ask: What is an aggregate welfare-maximizing information policy in this setting, and does it work as intended in practice?

If there exists an information policy that implements the pointwise maximum aggregate welfare, i.e., a policy that can persuade agents to make joint reports to the mechanism which maximizes aggregate welfare at *each* realized preference profile, then such a policy is clearly optimal for the planner. We call the pointwise maximum allocation the *first best* allocation.

It can be shown that such a policy exists if and only if agents do not have strong opinions about the options a priori. Moreover, in that case the planner can achieve the first best by simply privately recommending to each agent the option *she* would like him to pick, at each realized preference profile - an information policy we call the *first best recommendation signal*, hereafter called the *recommendation*.⁸

As mentioned above, the key to the first best being generally implementable is that the agents must be *sufficiently suggestible*. For example, if an agent knows a priori with probability one that object *A* is better for him than object *B* - even though he does not know for sure the cardinal strength of his preference - no signal can change his posterior preference to object *B*. This is an example of having very strong a priori preferences - a condition which renders the recommendation uninformative, and therefore the first best not (generally) implementable. We formalize this idea below.

Let $\hat{u}_i \equiv u_{iA} - u_{iB}$ denote agent *i*'s *relative* preference for *A* over *B*. For a given utility profile $u \in \mathcal{U}$, let $\text{rank}(\hat{u}_i)$ denote the rank of agent *i* when all agents in *I* are sorted from highest to lowest according to \hat{u}_i , i.e., $\text{rank}(\hat{u}_i) = r$ means there are $r - 1$ other agents in the economy whose relative preference for *A* over *B* is at least as strong as *i*'s. Suppose further that the prior measure is atomless. In this setting, agent *i* is said to have a **strong a priori preference** for object *A* if $\mathbb{E}(\hat{u}_i | \text{rank}(\hat{u}_i) \geq n) > 0$. In other words, agent *i* has a strong a priori preference for *A* if, even when he knows that his relative preference for *A* over *B* is not among

⁸A version of this signal is called the *Object Recommendation (OR)* signal in Dasgupta (2020).

the top half of the population, he still believes A is better for him than B . The definition of a strong a priori preference for B is reciprocal.⁹

With the above background, Dasgupta (2020) shows the following:

Proposition 1 (Adapted from Corollary 4, Dasgupta (2020)). *Suppose agents' preferences are iid and the prior is atomless. First best is achievable by random serial dictatorship if and only if agents have no strong a priori preferences, in which case it can be achieved using the recommendation.*

Clearly, in our experimental set up, the prior is not atomless – it has a finite support. However, it can be shown that Proposition 1 holds for arbitrary priors, with a slightly more generalized definition of *strong a priori preferences* (see Appendix D.1). That definition boils down to the one presented above when the prior is atomless.

2.2 Parametrization of the baseline model

In what follows, we fix the remaining model parameters such that each of the two options has capacity $n = 2$, which are to be allocated among a set of four ($2n$) agents.

Option A is commonly known to have a value of 300 to all agents, whereas option B may take values 100, 300, or 500 with probabilities 20%, 60%, and 20% respectively (Scenario 1, where agents are *sufficiently suggestible*), or it may take values 0, 200, or 400 with probabilities 20%, 60%, and 20% respectively (Scenario 2, where agents have *strong a priori preferences*).

We consider three information settings: *No Info*, *Partial Info*, and *Full Info*.

Under *No Info*, agents have no further information about their realized value of option B prior to making their decision which option they prefer.

Under *Partial Info*, agents receive a Recommendation, i.e., information on which option the social planner recommends them to pick given that it is known that she aims to maximize aggregate social welfare. The agents also know that when the planner is indifferent among multiple allocations - i.e., multiple allocations all generate the same maximum aggregate payoff - she picks one uniformly at random and recommends it to agents.

Finally, under *Full Info* agents know their own realized value of option B .

2.3 Theoretical predictions

In this section, we state the hypotheses we test in our experiment and provide some intuition behind them. Formal proofs are provided in Appendix D.2.

⁹Agent i is said to have a strong a priori preference for object B if $\mathbb{E}(\hat{u}_i | \text{rank}(\hat{u}_i) \leq n) < 0$.

Let W_t^s denote the aggregate payoff (i.e., the sum of individual payoffs) of all four participants in Scenario s and information setting t , i.e., $s \in \{1, 2\}$ and $t \in \{\text{No Info}, \text{Partial Info}, \text{Full Info}\}$. Also, let W^{s*} denote the maximum possible aggregate payoff possible in Scenario s . For each of our hypotheses, we make the assumption of risk neutral, Bayesian, and purely self-interested agents.

Hypothesis 1 (Payoffs in Scenario 1). In Scenario 1, the aggregate payoff is the lowest under *No Info*. It is strictly higher under *Full Info*. Finally, it is the highest under *Partial Info*, which is equal to the theoretical maximum possible aggregate payoff. That is, $W_{NoInfo}^1 < W_{FullInfo}^1 < W_{PartialInfo}^1 = W^{1*}$.

The intuition for why *Full Info* does better than *No Info* in Scenario 1, is as follows: Under *Full Info*, at least two out of the four participants can maximize their own payoffs – those ranked 1 and 2 under random serial dictatorship. In contrast, under *No Info*, the expected payoff is constant across agents and objects. Hence, no action by any agent can improve their own – or the group’s – interim aggregate welfare.

As for the second part of the hypothesis why *Partial Info* achieves the first best: Note that in Scenario 1, each agent is *ex-ante indifferent* between the two options, because the expected payoff from $B = \frac{1}{5} \times 100 + \frac{3}{5} \times 300 + \frac{1}{5} \times 500 = 300$, which is the same as the payoff from A . In other words, agents have no strong a priori preference for either option, in the sense described in Section 2.1.2. Thus, in this Scenario, if an agent knows that his allocation under one of the utilitarian welfare maximizing allocations is A , and updates his expected utilities based on Bayes rule, a posteriori, he prefers A over B , as outlined by Proposition 1. More intuitively, due to the *high suggestibility* of each agent in this scenario, as captured by his ex-ante indifference, this minimal “good news” about A is sufficient to tilt his posterior preference in favor of A . This channel works for both objects. Therefore, per Proposition 1, in this case the recommendation theoretically achieves the first best as agents follow their Recommendations.

Hypothesis 2 (Payoffs in Scenario 2). In Scenario 2, the aggregate payoff is the lowest under *No Info*, which is equal to that under *Partial Info*. It is strictly higher under *Full Info*, which is, in turn, below the theoretical maximum possible aggregate payoff. That is, $W_{NoInfo}^2 = W_{PartialInfo}^2 < W_{FullInfo}^2 < W^{2*}$.

The intuition for Hypothesis 2 is as follows: In Scenario 2, the agents’ ex-ante expected payoff from object $B = \frac{1}{5} \times 0 + \frac{3}{5} \times 200 + \frac{1}{5} \times 400 = 200 < 300$, i.e., agents have an a priori preference for object A . In fact, calculations show that this a priori preference is strong enough so that the recommendation provides *too little* useful information to agents. That is, even when an agent knows that he has been allocated B under the utilitarian welfare-maximizing rule, this is not sufficient “good news” to shift his a posteriori preferences to B . He still prefers A to B a posteriori,

which is the same as his prior preference. Therefore, equilibrium outcomes with the recommendation are the same as without any information. It follows that the ex-ante social welfare should be equal in the *No Info* and *Partial Info* cases. As for *Full Info*, it maximizes welfare for at least two out of the four participants, as in Scenario 1, and thereby improves welfare over *No Info*.

Hypothesis 3 (Agent choices in Scenario 1 vs 2 in *Partial Info*). For each participant, the acceptance rate of the recommendations in Scenario 1 is 100% regardless of which object is recommended. In Scenario 2, it is 100% when the recommended object is *A* and 0% when it is *B*.

Hypothesis 3 simply summarizes the predictions for individual behavior, which follow directly from the arguments outlined above.

2.4 Behavioral extensions

Recognizing that our prior assumption of risk neutral, Bayesian, and purely self-interested agents is unlikely to hold in practice, we outline several behavioral extensions of the baseline model and use them to qualify our three key hypotheses. Specifically, we consider risk attitudes, imperfect Bayesian updating, and social preferences. We also run simulations to show the magnitude of the impact of these deviations from the baseline model on group welfare.

We show that the agents' risk preferences do not change our baseline predictions for a wide range of parameter values. However, risk preferences do play a role when agents are additionally prior-biased; specifically prior bias cannot exceed 60% for our baseline predictions to hold across all risk preference profiles. Finally, when agents have other-regarding concerns, we derive the conditions when they would be willing to follow the recommendation even in Scenario 2, altering our Hypothesis 3 and, as a consequence, Hypothesis 2, as higher aggregate social welfare would be reached in Scenario 2 under *Partial Info* in that case.

The details are in Appendix C.

3 Experimental design

To test our theoretical predictions, we conducted an online experiment. In this section we first describe its general structure, and then detail all the tasks our subjects completed. Implementation and power calculations are described in the final two subsections.

3.1 Experiment structure

Our experiment consists of two key stages: In the first stage, the subjects familiarize themselves with and complete the main task of interest: a choice between two objects, A and B , in a matching setting as described in section 2.2. Between-subject, we vary the amount of information (information setting treatment) the subjects receive about the two options. Within-subject, we vary the possible payoffs associated with option B (Scenarios 1 and 2). In the second stage, the subjects complete a sequence of tasks designed to measure their underlying preferences and background information to help us establish the reasons for their possible deviations from predicted behavior.

All our subjects complete the experiment independently and without feedback between tasks, and are only matched into groups of four ex post. As a consequence, we can treat individual decisions as independent observations for the purposes of Hypothesis 3, and group outcomes as independent observations for Hypotheses 1 and 2.

3.2 Experimental tasks

Upon entering the online interface, subjects are automatically randomized into one of the three between-subject treatments: *No Info*, *Partial Info*, or *Full Info*. These treatments only differ in the amount of information that is provided to the subjects on the main task of interest.

Subjects start the experiment by reading instructions regarding their choice between objects A and B . In the experiment, we refer to this main task as a “game”, and frame the choice as if the subjects were choosing between two different tasks as workers.¹⁰ The subjects are shown a detailed example describing how the random serial dictatorship allocation mechanism works in a group of four, and they complete a practice decision with four comprehension questions. Subjects who make a mistake on comprehension questions are nudged towards the part of instructions that explains that particular concept, and are asked to resubmit their answer. The number of mistakes on the comprehension quiz as well as the time spent on different parts of the instructions is tracked.

The three treatments correspond to our theoretical benchmarks and provide the following information to the subjects:

- **No Info:** Subjects know their payoff from choosing option A , and know the probability distribution and the associated possible payoffs from choosing option B .

¹⁰Recent evidence suggests that the choice of frame does not affect outcomes to a large extent (Abbink and Hennig-Schmidt, 2006; Engel and Rand, 2014).

- **Partial Info:** Subjects have the same information as those in No Info, plus receive a recommendation generated by the computer. The subjects know that the computer calculates the set of utilitarian welfare maximizing allocations given their realized payoffs from option B , picks one of them uniformly at random, and makes the corresponding recommendations to all agents.¹¹
- **Full Info:** Subjects know their exact payoffs from choosing tasks A and B .

After the subjects submit their chosen object, the computer randomly sorts them into groups of four, orders them in a queue, and then assigns them their preferred object in the order they were sorted as long as their preferred object is still available.¹²

The subjects make two decisions (Scenarios) with the following parameters (also discussed in section 2.2):

- **[Scenario 1] Ex-ante indifferent prior:** If a subject is assigned object A at the end of the scenario, they get 300 points. If they get object B , they get 100 or 500 points, each with probability 20%, or 300 points with probability 60%.
- **[Scenario 2] Strong preference:** If a subject is assigned object A at the end of the scenario, they get 300 points. If they get object B , they get 0 or 400 points, each with probability 20%, or 200 with probability 60%.

The preference distributions are i.i.d. across agents in both cases.

Without any feedback,¹³ the subjects proceed with follow-up tasks (see Figure 1).

The subjects' payoffs are revealed at the end of the experiment, and payment takes place electronically within two days of each experimental session.

¹¹By design, we ensure that these recommendations are always correctly calculated. For a discussion to what extent our theoretical predictions would change if errors were possible, see section C.3.3.

¹²Notice that the subjects do not know their ranking at the time of making their decision; this is done in order to maximize statistical power, as this way everybody's decision potentially matters for outcomes and truthful reporting of preferences is incentive-compatible. If the subjects did know their ranking, all subjects ranked fourth in each group would know that their decision does not matter for group outcomes, which could introduce (additional) noise into their decisions.

¹³This is done in order to minimize concerns about possible order effects: This way subjects have no way of learning anything from Scenario 1 that they could use in Scenario 2. Likewise, we prevent hedging across these decisions by making it clear to the subjects that only one decision from the experiment is going to be paid.

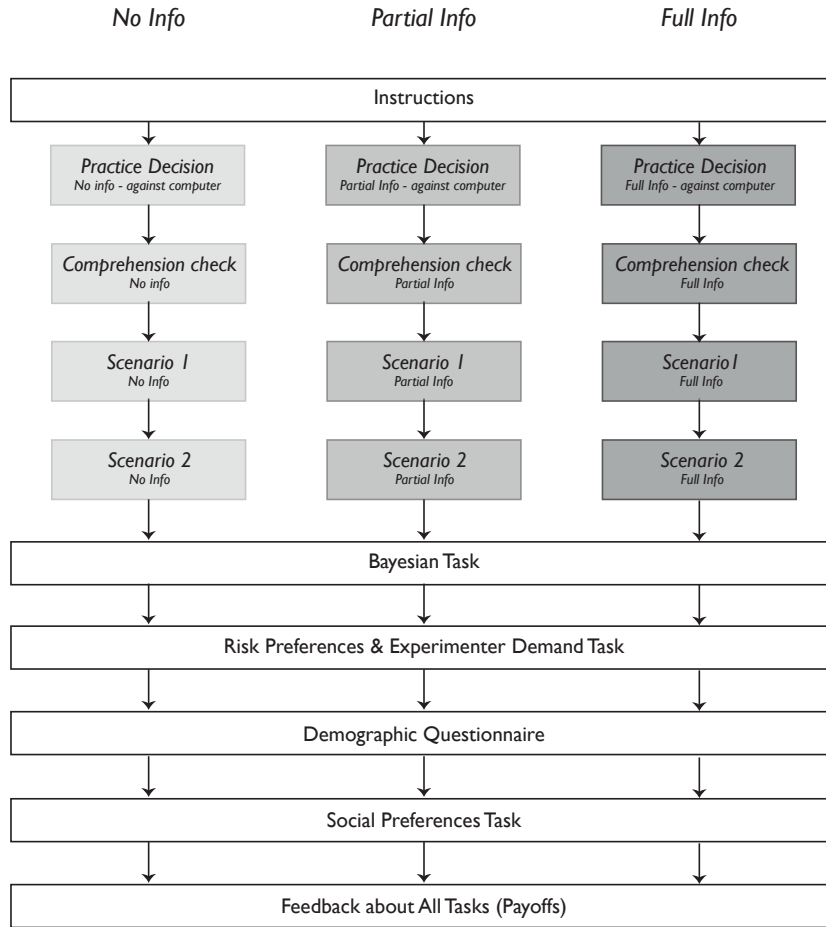


Figure 1: Sequence of tasks. Shaded columns indicate (between-subject) treatments.

The subjects first complete a Bayesian updating task (Mellers et al., 2017)¹⁴ and the Bomb Risk Elicitation Task (BRET) (Crosetto and Filippin, 2013). We ask the subjects to complete the BRET twice, and the second time we provide an explicit request how we would like the subjects to behave: we consider this as an upper bound on experimental demand our subjects succumb to. Next, the subjects complete a demographic questionnaire with an attention check, and a social preference elicitation: We offer the subjects either a bonus for themselves, or for everybody *else* in their group. By the nature of this elicitation, in what follows we refer to subjects who opt for the bonus for others rather than themselves as *altruists*. Finally, the subjects receive information about their earnings.

¹⁴We slightly changed the wording of the Bayesian updating task to minimize the probability that subjects would find the answer by a Google search (Ludwig and Achtziger, 2021).

3.2.1 Purpose of ancillary tasks

While our main hypotheses are written assuming agents are risk neutral, perfectly Bayesian, and completely self-interested, we acknowledge that they are unlikely to be (Brase and Hill, 2017; Harrison and Swarthout, 2022; Cochard et al., 2021; Fromell et al., 2020). Guided by the experimental literature, we propose four main channels that may be driving behavior that align with our theoretical discussion in Section C: (1) risk aversion combined with lack of/insufficient Bayesian updating, (2) social preferences, (3) experimenter demand, and (4) mistakes and misunderstanding by subjects.

First, as discussed in sections C.1 and C.2, if people are risk averse *and* they update their priors insufficiently, our prediction for Scenario 1 changes such that subjects start ignoring the recommendation and choose the safe option instead. The literature shows that while many people update their beliefs roughly in line with Bayes' rule (Coutts, 2019; Barron, 2021), there is a lot of individual heterogeneity (Holt and Smith, 2009) and thus scope for deviations from optimal play. For this reason, we measure the both subjects' risk preferences and their ability to update based on new information.

Second, as proposed in section C.3, since high levels of altruism can affect subjects' behavior in the experiment, particularly in Scenario 2, we measure subjects' willingness to sacrifice their own payoff in order to benefit others in their group.

Third, subjects might deviate from "optimal" behavior due to experimenter demand effects (Zizzo, 2010): We would expect some subjects to follow the recommendation in an attempt to please the experimenter regardless whether they understand the task, are able to update their beliefs, or care about the well-being of others.¹⁵ Alternatively, the subjects may defy the experimenter on purpose, doing exactly the opposite of what they are asked. To bound the likely size of these effects, we repeat our risk preference task (BRET) with direct experimenter instructions to behave in a particular way, reasoning that subjects eager to defy or to please are likely to do so across tasks.

Finally, like in any human activity, it is possible that our subjects make mistakes. Broadly, we can attribute mistakes to two possible reasons: inattention, and misunderstanding. Fortunately, in a controlled experimental environment, we can measure both: As a direct verification whether our subjects pay attention, we incorporate an attention check at the end of the experiment. This is because we would expect subjects' attention to decrease over time, and, to prevent any possible retali-

¹⁵More broadly, one could think of our recommendation as the experimenter's way of communicating a certain *norm* for behavior. However, as shown by Arroyos-Calvera et al. (2023), conveying a behavior norm is not essential for a recommended action to influence behavior. This is consistent with what our subjects self-report in their explanation of what motivated their choices: experimenter demand or any reference to an expected 'norm' of behavior were rare in our experiment.

ation by subjects when they realize they are being “checked” whether they are doing a good job (Paas and Morren, 2018). However, even if subjects do pay attention, they might not fully understand the experimental instructions. For this reason, we measure how many times subjects fail our comprehension quiz, and which parts of the instructions they misunderstand.

3.3 Procedures

We obtained IRB approval from the University of Chicago (IRB21-1695), the University of Amsterdam, and Nova School of Business and Economics, and pre-registered the experiment at AsPredicted.org (<https://aspredicted.org/yu65p.pdf>). Specifically, we pre-registered our hypotheses, outcome variables, and analyses of the main hypotheses, and mechanisms we explore to study behavior deviating from the theoretical predictions.

To check that our software was working properly, we ran a pilot session with 42 subjects in total; we do not use this data in our analysis. We conducted the experiment on Prolific in August 2022, selecting subjects to be US nationals with a Prolific approval rating of at least 90%. The analysis code can be found at https://www.jantsje.nl/files/analysis_matching.html.

3.4 Power calculations

In our power calculations, we determined the minimum necessary sample sizes to detect the average effect size in Scenario 1 in case the *Partial Info* treatment outperforms the *Full Info* treatment with 80% power.

Under standard assumptions ($\alpha = 0.05$, a two-tailed Wilcoxon rank sum test), our setup thus requires at least 2220 subjects across the three treatments. For details on how we averaged the effect sizes of interest, and auxiliary assumptions, please see Appendix A.

4 Experimental results

In total, we collected data for 650 groups (2600 subjects),¹⁶ with 89 groups in *No Info*, 281 in *Partial Info*, and 280 in *Full Info*. As pre-registered, we drop all groups consisting of only subjects who did not complete the main part of the experiment (Scenarios 1 and 2). Since dropping groups that contain a combination of some drop-outs and some participants who finished would decrease our power under 65%, then,

¹⁶We recruited more subjects than our pre-registered minimum because we expected some subjects to drop out, as is common in online experiments with the general population.

as pre-registered, we keep these groups for the analysis.¹⁷ For the main analysis we are therefore using data from 603 groups, split 80-262-261 between the *No-Partial-Full Info* treatments, which is slightly more groups than the pre-registered desired minimum. Appendix G.2 shows that the results do not change if we focus on groups where all participants completed the experiment.

4.1 Descriptive statistics

Table 1 provides the summary statistics for the experiment. Across all treatments, groups achieved higher social welfare in Scenario 1 than in Scenario 2, both in absolute and relative terms, but this difference was the largest in *Partial Info*. The majority of subjects chose the safe option A, with more choosing it in Scenario 2, but this share was higher in *No Info*. In the *Partial Info* treatment, the majority of subjects followed the Recommendation, but fewer did so in Scenario 2.

Looking at subject characteristics, around 42.3% of our subjects were women, and 7.6% students. The experiment took approximately 14 minutes, and average earnings were 2.36 GBP.¹⁸ Around 9.4% of subjects re-checked the instructions after reading them for the first time, but even then more than 63.3% of subjects failed at least one comprehension check. Almost 49.0% of subjects failed our attention check at the end of the experiment.¹⁹ Most subjects were moderately risk averse, but were willing to increase their exposure to risk by almost one third (corresponding to about 12 boxes) when prompted to do so in our experimental demand task. Most subjects also made mistakes on the Bayesian task, deviating from the correct answer by 13.4 percentage points (which is 22.3% of the correct answer, 60) on average. Less than a third of the subjects (around 32.0%) behaved altruistically, i.e., chose to give money to others in their group rather than keep it for themselves.

The Balance Table (G1) is reported in the Appendix.

¹⁷Following our pre-registration, if a subject dropped out of the experiment before completing one or both of the Scenarios of interest, their choice is randomly determined by the computer (so that group outcomes can be calculated). These randomly assigned choices are not used for individual-level analysis (Hypothesis 3). From now on, we refer to these computer-assigned choices as ‘bots’.

¹⁸At the time of the experiment, this average payment corresponded to 140% of the Prolific minimum-wage rules for experiments.

¹⁹We consider this an upper bound on the number of subjects not paying careful attention for two reasons: First, the attention check question was not incentivized, unlike the rest of the experiment. Second, the attention check was at the end of the experiment (in order not to anger subjects that they are “monitored”), and attention is likely to decrease over the course of the experiment.

Table 1: Summary Statistics

	Full Info		Partial Info		No Info	
	Scn 1	Scn 2	Scn 1	Scn 2	Scn 1	Scn 2
Main outcomes:						
Aggregate welfare	1278	1077	1262	1031	1195	1010
(sd)	(182)	(186)	(168)	(184)	(183)	(168)
... as % of max	93.92	92.47	93.15	89.02	89.70	87.89
(sd)	(8.15)	(10.65)	(9.12)	(12.88)	(11.09)	(11.43)
% choosing A	59.68	76.24	55.78	78.79	66.42	86.35
% following recommendation	-	-	74.13	67.02	-	-
Subject characteristics:						
Completion time (min)	13.26		14.20		14.14	
Risk aversion	58.14		59.22		60.70	
Experimenter demand	11.73		11.65		13.52	
Bayesian deviation	13.00		13.50		14.82	
% female	42.28		43.12		39.61	
% student status	6.13		9.10		7.06	
% instructions check	10.41		8.13		10.20	
% instructions failure	60.57		65.76		63.92	
% attention failure	47.75		49.73		50.59	
% altruist	28.06		29.90		28.78	

Note: Main outcomes of interest are reported for every within- and between-subject treatment (Scenarios 1 and 2, and *Full/Partial/No Info*). Aggregate welfare refers to group outcomes, whereas shares of subjects choosing a specific strategy are calculated on an individual level. Individual-level statistics do not include choices determined by the computer for subjects who dropped out.

Subject characteristics are reported for every between-subject treatment on an individual level. Completion time refers to the total number of minutes a subject took to complete the experiment, and is reported only for subjects who completed the experiment. Share of instructions check, instructions failure, and attention check refers to the share of participants who went back to re-read the instructions, who failed the instructions comprehension quiz, or who failed the attention check, respectively. Risk aversion refers to (100 - the number of boxes collected on the BRET task), such that higher values mean higher risk aversion. Experimenter demand is measured as the total increase in the number of boxes collected on the BRET task with explicit experimenter instructions as compared to our first task without any nudges about “correct” behavior. The deviation from Bayesian updating on our cookie task is listed as the absolute difference in probability reported and the Bayesian estimate. Finally, we report the percentage of subjects who chose to sacrifice their own payoff and instead benefit others in their group (altruist).

4.2 Main results

Our analysis matches our pre-registration unless noted otherwise. Exploratory (non pre-registered) analyses are listed in the next section.

Since our Hypotheses 1 and 2 involve a sequence of inequalities to compare social welfare levels across treatments, we run the Jonckheere-Terpstra trend test, and verify every pairwise inequality with the Mann-Whitney-U test (see Table 2). As that table shows, we find partial support for our Hypothesis 1 with the aggregate welfare being statistically indistinguishable in the the *Partial* and *Full Info* treatments in Scenario 1, but significantly lower in the *No Info* treatment. Additionally, we find convincing support for our Hypothesis 2 with the *Full Info* treatment outperforming both *Partial* and *No Info* in Scenario 2, which are in turn yielding equal aggregate welfare. These results are visualised in Figure 2.

To compare welfare levels to the specified theoretical maximum, we use the one-sample version of the Wilcoxon sign rank test (see Table 3). Contrary to our theoretical predictions, the *Partial Info* treatment does not reach the first best in Scenario 1. Likewise, the *Full Info* treatment does not reach the first best in Scenario 2, in line with our Hypothesis 2. Both of these results can be clearly seen in Figure 2, with the first best benchmark being represented by the dashed line.

Table 2: Treatment Effects on Social Welfare

	All comparisons (Jonckheere-Terpstra)	Full vs. Partial (Mann-Whitney-U)	Partial vs. No (Mann-Whitney-U)	Full vs. No (Mann-Whitney-U)
H1: Partial > Full > No	0.082	0.218 [0.096]	0.002** [0.005]	0.000*** [0.002]
H2: Full > Partial = No	0.000***	0.005** [0.005]	0.381 [0.146]	0.005** [0.005]
N (groups)	603	523	342	341

Note: The first column lists p-values from the Jonckheere-Terpstra trend test for the ordered aggregate social welfare levels, and columns 2-4 list p-values for two-sided pairwise comparisons using the Mann-Whitney-U test.

* p-val < 0.05, ** p-val < 0.01, *** p-val < 0.001

Sharpened false discovery rate q-values for the six pairwise tests (Anderson, 2008) are in brackets.

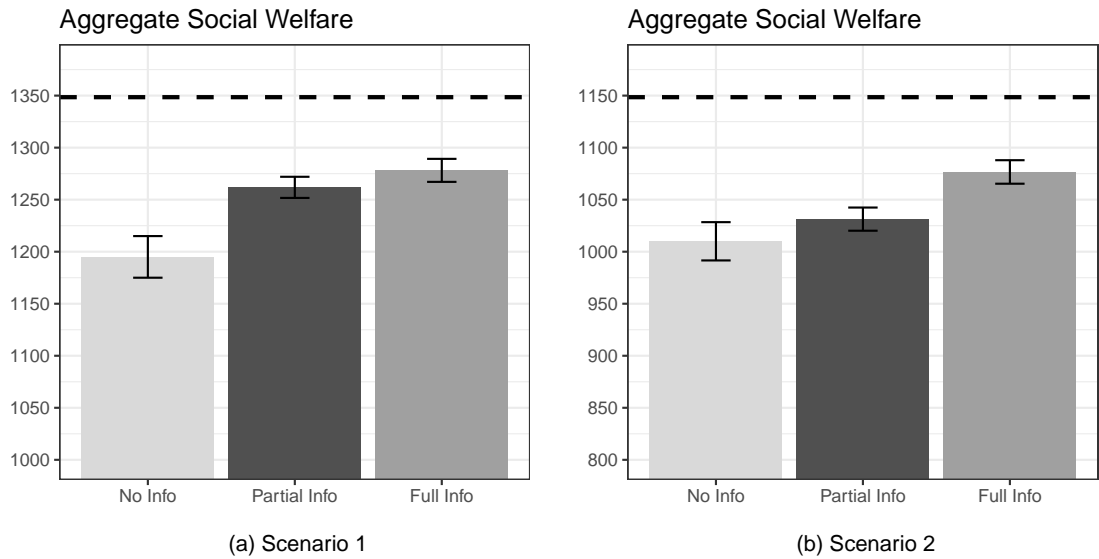


Figure 2: The figure plots the aggregate social welfare reached in Scenarios 1 and 2 with 95% confidence intervals. The dashed line indicates the theoretical maximum aggregate welfare (first best) that can be achieved in expectation.

Table 3: Reaching first best

	Scenario 1	Scenario 2
H1: $W^* = \text{Partial}$		H2: $W^* > \text{Full}$
	0.000***	0.000***
	[0.001]	[0.001]
N (groups)	262	261

Note: The table lists the p-values for the one-sample two-sided Wilcoxon sign rank test, comparing the treatment that was hypothesized to equal (Scenario 1) or fail to reach (Scenario 2) the aggregate social welfare optimum to the theoretical first best.

* p-val < 0.05, ** p-val < 0.01, *** p-val < 0.001

Sharpened false discovery rate q-values for the two tests (Anderson, 2008) are in brackets.

As a robustness check, we randomly reshuffle subjects in groups such that within each group, we allow for all possible orderings of subjects and thereby all possible orderings whose decisions are consequential for the group outcomes. (Recall that the subject ranked 4th has no power to influence own or group outcomes.²⁰) We plot the aggregate social welfare levels for each reordering (see Figure 3), showing that our results are generally stable with one exception: in the majority of alternative subject reshufflings under *No Info* in Scenario 2 the social welfare is lower than in our

²⁰This method is in the spirit of Mullin and Reiley (2006).

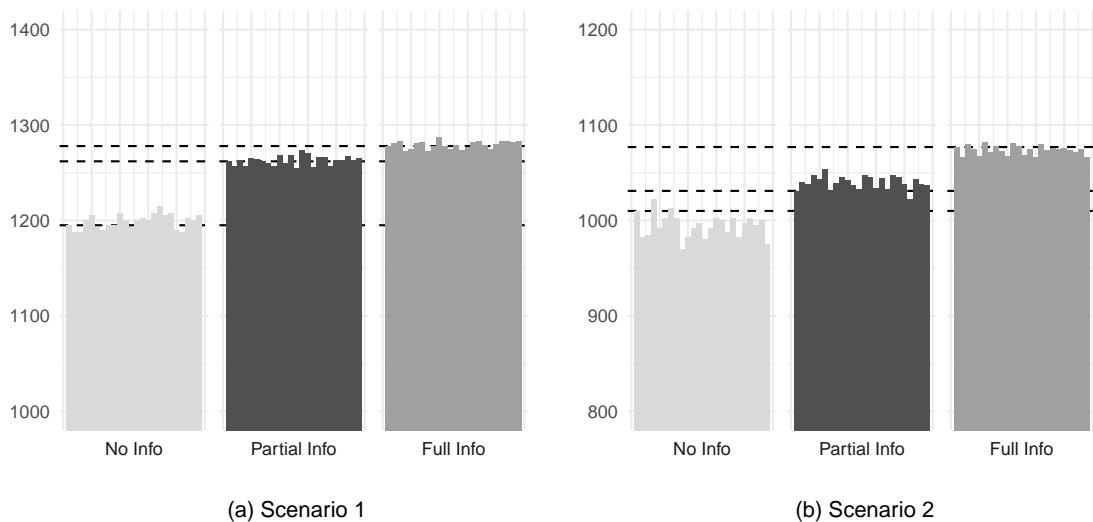


Figure 3: The figure plots the aggregate social welfare reached in Scenarios 1 and 2 under different subject reshufflings within groups. The dashed lines indicate the average social welfare levels achieved in our default ordering of subjects.

original dataset. In the 23 additional pairwise (Mann-Whitney-U) tests where we compare the aggregate social welfare in Scenario 2 under partial vs. no information, we find a significant difference in 4 reshufflings (17%). This is more than would be expected by pure chance (5%), and so we see it as a qualification of our result from Table 2, suggesting that there might be groups that slightly benefit from partial information as compared to no information even in Scenario 2.

Based on the above results, we conclude that in Scenario 1, where following the recommendation is theoretically optimal, the *Partial Info* and the *Full Info* treatments result in the same level of aggregate social welfare, which is below the first best. Both of these treatments outperform the *No Info* benchmark. In Scenario 2, where subjects have a strong a priori preference for object *A*, *Full Info* outperforms both the *Partial* and *No Info* treatments, both of which result in the same aggregate welfare levels, albeit we have suggestive evidence that the *Partial Info* treatment may be marginally better for subjects than the *No Info* treatment. Finally, also in this case, the *Full Info* treatment does not reach the first best.

Moving on to our Hypothesis 3 regarding individual behavior, we use the one-sample Mann-Whitney-U test to see whether the share of people following each type of recommendation corresponds to the predicted share. Since this analysis is on the individual level, we directly exclude subjects who did not complete one or both of the treatment scenarios and their choices were replaced by a random choice ('bots'), since we maintain sufficient power. To establish whether hint following behavior differs across scenarios, we use the within-sample Wilcoxon sign-rank test. All of these results are reported in Table 4.

Table 4: Following the Recommendation

	Scenario 1	Scenario 2	Scenario 1 vs. 2
	H3: always follow	H3: only follow A	
% follow	74.13		
	0.000***		
	[0.001]		
% follow A	79.62	95.38	-15.76
	0.000***	0.000***	0.000***
	[0.001]	[0.001]	[0.001]
% follow B	68.52	38.12	30.40
	0.000***	0.000***	0.000***
	[0.001]	[0.001]	[0.001]
N (subjects)	943	943	943

Note: The table lists the average share of subjects who follow the Recommendation in the Partial Info treatment and the associated p-values for comparisons w.r.t. theoretical benchmarks (columns 1 and 2) and between scenarios (column 3). We do not run the aggregate test for hint following (row 1) in Scenario 2 since the theoretical prediction depends on the realized share of subjects receiving each recommendation; we therefore only test conditional on realized recommendations (rows 2 and 3).

* p-val < 0.05, ** p-val < 0.01, *** p-val < 0.001

Sharpened false discovery rate q-values for the seven tests (Anderson, 2008) are in brackets.

As the above table makes clear, subjects do not always behave in line with the theoretical predictions: In Scenario 1 and for cases of *A* recommendation in Scenario 2, they do not (sufficiently) follow the recommendation even though it is in their interest, whereas in Scenario 2 in case of a *B* recommendation they follow the recommendation even when it is not in their private interest. However, recommendation-following is in the aggregate higher in Scenario 1, and this is driven by subjects following recommendation *B*. Interestingly, recommendation-following of *A* is higher in Scenario 2, even though it should theoretically be equal in both Scenarios.

Finally, we move on to the mechanisms why subjects behave this way. First, we investigate which strategies the subjects pursue given the information/recommendations they received. In Table 5 below, the actions taken by the subjects (one in each Scenario) are indicated by columns, while the information received by them are along the rows. In case of *Full Info*, instead of recommendations received, we break the sample down based on which action would result in higher expected payoff for the subject, since the exact payoffs each option offers are known. In case of equal payoff for options *A* and *B*, we use =. Individually rational

payoff-maximizing choices, assuming no social preferences, are highlighted in **bold**.

Table 5: Strategies Played

	Actions taken			
	AA	AB	BA	BB
No Info:	164	16	70	21
Partial Info: Recommendations received				
AA	225	1	43	9
AB	96	57	35	10
BA	64	2	122	10
BB	66	15	92	96
Full Info: Highest expected payoff				
AA	113	16	6	6
AB	7	29	1	1
BA	15	5	128	8
BB	1	4	6	36
=A	266	18	119	18
=B	32	49	15	31
N (subjects)	1049	212	637	246

As Table 5 shows, the majority of subjects behave in accordance with the theoretical predictions of Bayesian, self-interested profit maximizers. In the *Full Info* treatment, most (82.9%) subjects choose the individually payoff-maximizing object. In contrast, 56.7% of subjects do so in the *Partial Info* treatment. In *No Info*, the individually optimal payoff-maximizing option depends on risk preferences: a full 86.3% of subjects play strategies consistent either with risk aversion or risk neutrality (*AA* or *BA*), which is largely consistent with past experimental research on risk preferences.

In order to investigate why some subjects deviate from our theoretical predictions in the *Partial Info* treatment, we estimate a logit model to see how our proposed channels influence decisions.²¹ As Table 6 shows, by far the strongest predictor of subjects' choices was the recommendation that they received. In Scenario 1,

²¹In the Appendix Table G2, we examine the results as a sequential logit to see how the our proposed channels influence decisions at different stages (Buis, 2013). The results are similar to the simple logit presented here.

those who were recommended A were almost twice as likely to make the theoretically optimal decision, i.e., follow their recommendation. This effect was even more pronounced in Scenario 2 where only following the recommendation of A was individually optimal: those who received the recommendation of A were now *fourteen* times more likely to follow it. Once we include controls for our proposed mechanisms, we observe that subjects who made mistakes on the game comprehension quiz are only about half as likely to make the optimal decision in Scenario 2. The raw data confirm the pattern that those who misunderstood the instructions are less likely than those who understood (76.9% versus 84.8%) to make the optimal decision in Scenario 2 (Fisher’s Exact Test, $OR = 1.67$, $p = 0.005$). There is no difference between those who misunderstood (57.0%) and those who did not (53.5%) in optimal decisions in Scenario 1 (Fisher’s Exact Test, $OR = 0.87$, $p = 0.328$). This is relatively unsurprising, since the optimal strategy in this Scenario is arguably more complex than in Scenario 1 (i.e., involves following some recommendations and not others).

In an exploratory analysis (see columns 5 and 6 of Table 6), we specifically show that the sub-optimal choices in Scenario 2 are primarily taken by subjects who made errors on comprehension questions 3 and 4 prior to the experiment; i.e., they either do not understand that it is in their interest to truthfully report their preferences (Q3), or do not understand the information setting (treatment) they are assigned to, and thus likely misunderstand the information they are given about option B. This is also apparent in the raw data, as reported in Table 7.

Additionally, we observe that altruistic subjects are also less likely to make the optimal decision in Scenario 2, which is consistent with our theoretical discussion in Section C.3. However, we do not find robust support for altruism driving optimal behavior in either scenario based on raw data (Fisher’s Exact Test, $OR_{Scenario1} = 1.03$, $p_{Scenario1} = 0.935$, $OR_{Scenario2} = 1.13$, $p_{Scenario2} = 0.487$). For this reason we caution against placing too much emphasis on the significant coefficient of altruism in Table 6.

Quite remarkably, we find no disproportionate preference for the safe object A among those subjects who are both risk averse and bad at Bayesian updating, even in Scenario 2.

Finally, we find no evidence that either experimenter demand or inattention are driving our results, as these do not seem to influence the probability that a subject makes the optimal decision.

Table 6: Optimal Choices across Scenarios: Partial Info

	Scenario 1		Scenario 2			
Recommended A	1.745*** (0.158)	1.697*** (0.159)	14.391*** (0.267)	16.065*** (0.266)	15.564*** (0.276)	16.725*** (0.274)
Risk averse		1.005 (0.006)		1.015 (0.008)		1.014 (0.008)
Non Bayesian		1.025 (0.016)		1.043 (0.027)		1.042 (0.028)
Risk averse × non-Bayesian		1.000 (0.0003)		0.999 (0.0004)		0.999 (0.0004)
Inattention		0.949 (0.159)		0.904 (0.194)		0.941 (0.197)
# attempts comprehension Qs		1.328 (0.163)		0.526** (0.215)		
Altruist		1.013 (0.173)		0.637* (0.207)		0.629* (0.209)
Experimenter demand		1.006 (0.005)		1.001 (0.006)		1.000 (0.005)
Failed comprehension Q1					1.034 (0.192)	1.080 (0.159)
Failed comprehension Q2					0.904 (0.190)	0.885 (0.195)
Failed comprehension Q3					0.655** (0.138)	0.634** (0.141)
Failed comprehension Q4					0.632* (0.188)	0.659* (0.195)
N (total)	869	869	869	869	869	869

Note: The table shows the odds ratios for making the theoretically optimal choices in the Partial Info treatment. Optimal strategies follow the Recommendation in the first Scenario, and select A in the second Scenario. Only subjects who completed all ancillary tasks are included.

Q1 = *suppose you would have chosen the other task, how much would you have earned?*, Q2 = *suppose two players ahead of you would have chosen task A too, would you have been allocated A?*, Q3 = *could player 4 improve their payoff?*, Q4 = *I will know the exact payoff of B (correct answer is treatment-dependent)*

Simple logit estimation.

Robust standard errors are in parentheses. Clustering on individual level.

* p-val < 0.05, ** p-val < 0.01, *** p-val < 0.001

Table 7: Choices of subjects who misunderstand in *Partial Info*

	Wrong at least once	Of those with a wrong answer...	
		chose A in Scenario 1	chose A in Scenario 2
Q1	12.2%	49.6%	82.3%
Q2	23.9%	59.3%	78.7%
Q3	36.9%	57.8%	74.8%
Q4	29.0%	56.0%	74.6%
Overall	63.3%	56.0%	76.2%

Note: Table indicates the percentage of subjects who answered specific comprehension questions wrong (first column) and decisions of those subjects in Scenario 1 (second column) and Scenario 2 (third column).

Q1 = *suppose you would have chosen the other task, how much would you have earned?*, Q2 = *suppose two players ahead of you would have chosen task A too, would you have been allocated A?*, Q3 = *could player 4 improve their payoff?*, Q4 = *I will know the exact payoff of B (correct answer is treatment-dependent)*

4.3 Exploratory analysis: Self-reported strategies

To shed more light on the subjects’ motives, we provide a descriptive analysis of their self-reported strategies in the experiment.

Following Scenario 2 in the experiment, we asked our subjects to explain in words their reasoning for the choices they made. We hired two research assistants, blind to our hypotheses, to manually classify²² their responses into categories that correspond to our pre-registered mechanisms of interest for the *Partial Info* treatment:

- **Risk preferences:** Differentiate whether subject chose the safe option (suggestive of risk aversion), the risky option (suggestive of risk seeking), or engaged in a risk/benefit calculation (suggestive of some sophistication in decision-making).
- **Altruist:** Indicate whether subject made a decision with the intention to benefit others.
- **Experimenter demand:** Differentiate whether subject made a decision with the intention of pleasing the experimenter (suggestive of positive experimenter demand) or indicated opposing the recommendation given in an effort of making their “own” decision (suggestive of negative experimenter demand)
- **Mistake/misunderstanding:** Indicate whether the subject’s response suggests that the person misunderstood the rules, was incorrectly using the in-

²²The research assistants independently hand-coded the self-reported reasons of subjects for choosing objects *A* or *B*; they discussed cases of disagreement in their coding afterwards. The analysis in this section uses their collective agreed-upon classification.

formation provided (e.g., was misinterpreting the recommendation), or was making a similar type of error.

- **No explanation:** Indicate whether the subject failed to provide an explanation for their choice.
- **Other:** Indicate whether subject listed other reasons for their choice (e.g., chose an option randomly).

Overall, 62.5% of the subjects explained their decision depended on their risk preferences; of these, the majority reported they weighed the risks and benefits of the two options (57.7%), while about a third (37.5%) indicated that they simply chose the safe option. Only a small share of subjects indicated they simply chose the risky option (4.8%). Recall from Section 2.4 our theoretical prediction that risk preferences matter only for (sufficiently) prior-biased agents; however, we find no support in the data that these subjects who considered their risk preferences are any better at the Bayesian updating task than the other subjects (Wilcoxon rank sum test, $p = 0.754$). We therefore point to these self-reports as symptomatic of a particular type of mistake: either not updating correctly, or not even realizing that one should update following the recommendation.

The second most common category of reasons given was “other” (10.9%); this included explanations such as “choosing the best option” without explaining what “best” meant, or relying on “gut feeling”. Other reasons provided were relatively uncommon, such as appeals to altruistic motives (0.7%) or experimenter demand (0.3%). 1.8% of the subjects failed to provide an explanation for their choice.

Finally, only (0.8%) of subjects provided an explanation that could be directly identified as a reasoning mistake or confusion. Unsurprisingly, all of these subjects also failed at least one question in our comprehension test, with Q4 being the most common mistake. However, their responses do not provide us enough detail to say anything further about what specifically they do not understand.

On top of these categories, the research assistants also indicated whether the subject mentioned they followed their recommendation (26.6%), took it into consideration (21.3%), or ignored it (49.7%). In cases the subjects did not mention the recommendation at all, the research assistants coded such responses as ignoring the recommendation.

Of those subjects who mentioned they followed the recommendation they were given, the majority (82.3%) did not mention any other reason for their choice. Of the subjects who mentioned at least one reason for their choice, by far the most common argument mentioned was that the subjects weighed the risks and benefits associated with each option. Other reasons for making choices were rare, including altruism and experimenter demand. Notice that these numbers are conservative since we did

not provide the subjects with any possible options to explain their choices; these are reasons the subjects came up with on their own.

Taken together, we believe this provides additional nuance to our earlier findings: While there is a share of subjects who simply follow the recommendation they are given without feeling the need to come up with their ‘own’ reasons for an action; most subjects think carefully about the problem and try to weigh the risks and benefits associated with each choice. However, we have few reasons to believe that the subjects engage in “correct” analysis (from the perspective of Bayesian updating and our model). Rather, the data and the self-reports point to the fact that subjects take mental shortcuts and do not properly internalize all information available to them.

5 Discussion

In this study, we have investigated, both theoretically and empirically, how people process and use information signals in an environment where individually and societally optimal actions do not necessarily align. We make three key observations: First, in contrast to theoretical predictions, a recommendation provided by a knowledgeable social planner in an incomplete information setting does not achieve first best aggregate welfare, even in the case where individual and societal objectives are aligned. However, in this case a recommendation signal improves social welfare over a no information setting. Second, in situations where individual and societal objectives clash, a substantial share of subjects follow the recommendation they received, benefiting their group. These decisions do not seem to be driven by social preferences, inattention, or experimenter demand; rather, it is subjects who struggle to understand the game who make the individually sub-optimal choice. And third, full information provision does no worse than the recommendation signal when individual and societal objectives are aligned, and outperforms it otherwise.

Our findings are relevant to settings where experts decide on a public advisory policy geared to aid the allocation of scarce resources (e.g., vaccines) among people with heterogeneous private benefits, in the absence of a market (e.g., for ethical reasons). They suggest that, while full information revelation does result in some inefficiency due to limited resources, an optimally crafted recommendation policy may not be able to counter this inefficiency enough, as it is not followed by a large enough share of the population.

The most relevant channel for the aforementioned deviation from the theoretical prediction that we find, is the subjects’ misunderstanding of the game. In particular, those who made mistakes on our comprehension quiz were less likely to make the optimal decision in the cognitively more demanding case (Scenario 2) where individual and societal objectives clash. However, in light of the known result that in

strategic environments, even correct strategic advice often fails to convince players to follow it (Guillen and Hing, 2014; Guillen and Hakimov, 2018; Braun et al., 2014; Koutout et al., 2021), we recognize that there may be additional channels driving the deviations (e.g., see Rees-Jones et al., 2020), which represent promising avenues for future research.

Our main insight from the above finding, however, is as follows. The notion that not all incentive compatible mechanisms are easily understood to be so by real world players is well established in the mechanism design literature, both theoretically and experimentally (Kagel et al., 1987; Chen and Sönmez, 2006; Klijn et al., 2013; Li, 2017; Li and Dworzak, 2021). Various more restrictive notions of incentive compatibility which accommodate for this limitation have been suggested, such as *obvious strategyproofness* (Li, 2017) and *simplicity* (Börgers and Li, 2019; Li and Dworzak, 2021). Our experimental findings provide evidence in favor of an analogous insight for information design – in strategic environments, all *obedient* (Bergemann and Morris, 2016) information policies need not be understood as such by real world agents. As an avenue for future research, it would be interesting to explore simplicity-focused theoretical notions of obedience.

Appendix

A Power calculations

We use simulations to guide our sample size choice for the experiment. Our key starting point is that for the recommendation to be a relevant policy, it needs to deliver welfare improvements over the next best-case alternative, which in this case is *Full Info*.

Building on our theoretical background, we propose three types of common behavioral deviations from optimal (Bayesian, risk neutral, self-interested) strategy in the two scenarios of interest:

1. *Always follow*: Subjects who always follow the recommendation, regardless of the payoff distributions they face. As argued in Sections C.3 and C.3.3, such behavior could be consistent with high levels of altruism or experimenter demand.
2. *Always opposite*: Subjects who always do the opposite of what the recommendation suggests. Analogously as above, this could result from reverse experimenter demand effect – a tendency to defy the experimenter on purpose – or negative altruism (spite).²³

²³As we note in Section C.3 in the Appendix, spite must be sufficiently high – in a sense made precise in Section C.3 – for this effect to come into play.

3. *Always safe*: Subjects who always select the certain/safe option.²⁴ As shown in Section C.2, such behavior would be consistent with a high degree of prior bias combined with risk aversion.

We proceed by simulating the behavior of 2500 groups (10 000 agents) assuming that different shares of the agents play either the (Bayesian, risk-neutral, self-interested) optimal strategy, or one of the three alternatives above. For example, we thus simulate the resulting aggregate social welfare when 20% of the agents play the optimal strategy, and 80% play *always safe*. We contextualize these by also showing the aggregate welfare under *Full Info*, and the minimum and maximum welfare that can be reached by forcing the agents to accept the welfare-maximizing/minimizing allocation.

Based on these simulations (depicted in Figures C2, C4, and C5), *Partial Info* outperforms *Full Info* in terms of aggregate welfare if *sufficiently many* subjects follow the recommendation. This is true even in Scenario 2, where subjects have a strong preference for the safe object. Depending on which type of “alternative strategy” we look at, the key threshold is around 70% or 80% of the agents following the recommendation.

Since we are primarily interested in Scenario 1 where the use of the recommendation is desirable, we average the effect sizes across the three cases at points where either 80% or 100% of subjects follow the recommendation. Based on this average effect size, we calculate the required number of groups in the *Partial* and *Full Info* treatments. We use GPower (Faul et al., 2009), and assume $\alpha = 0.05$, power of 80%, a two-tailed Wilcoxon rank sum test, and a normal parent distribution. Taken together, we calculate the target effect size to equal 0.25, giving us a required N=257 groups of subjects per each of these two treatments.

To calculate the required sample size for the *No Info* treatment, we proceed in three steps: First, we simulate subjects’ behavior in the *No Info* treatment, varying the share of subjects who choose the safe option vs. those who maximize expected value. (Notice that in Scenario 1, this implies choosing randomly, since both options have the same expected value.) This is shown in Figure A1. Second, to be conservative, we aim for 80% power to detect whichever effect size is *smaller*: that comparing *Full Info* and *No Info*, or that comparing *Partial Info* and *No Info*. Since in our cases of interest *Partial Info* outperforms *Full Info*, we therefore calculate the effect size of interest based on the 80% and 100% cases for *Full Info* and *No Info*. This gives an average effect size of 0.495. Third, we take into account that we are already using 257 groups for *Full Info*, and we can thus allow for a smaller number of groups in *No Info*. Setting the ratio of sample sizes to 1:5, we end up with 41

²⁴Note that theoretically, we could also distinguish an *Always risky* type, i.e., subjects who always select the risky option; however, given that most subjects tend to be risk-averse (Harrison and Rutström, 2008), we do not consider this a likely situation.

groups in the *No Info* treatment.²⁵

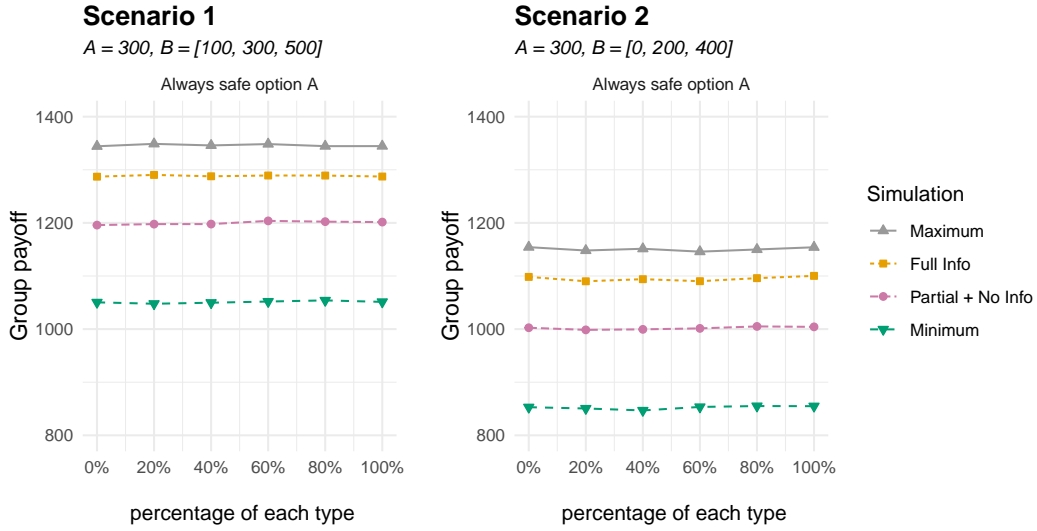


Figure A1: Average group welfare across scenarios and types of agents, *No Info*. It varies shares of subjects who choose to always select the safe option in pink (circles). We plot the *Full Info* in orange (squares), and the group maximizing (grey, upward pointing triangles) and minimizing (green, downward pointing triangles) outcomes.

Finally, we again consider the three cases of interest with either 80% or 100% of subjects behaving optimally, and we find the average effect size for Scenario 2. Taking the number of groups per treatment as given, we calculate the implied power to equal 99.97% for *Full Info* vs. *Partial Info*, and 77.34% for *Full Info* vs. *No Info*.

Taken together, we aim for at least 2220 subjects across the three treatments, with 257 groups in *Full Info* and *Partial Info* and 41 groups in *No Info*.

²⁵We end up having more power than 80% in practice, since we end up with more groups in *Full Info* than the calculation requires.

B Experimental instructions

Welcome

You will be paid £1.50 for participation.

The experiment consists of three parts, in which you can earn a bonus on top of this participation fee. The average participant will earn £0.64 in bonuses, based on decisions and luck (minimum £0, maximum £1.98).

You will be asked to make several different decisions. In the first part you will play a game in a group of four participants, but everyone can finish at their own pace. The other parts concern individual decisions.

The computer will randomly select one of the decisions for final bonus payoff.

In case this is a decision in the game, your bonus will be determined once all participants in your group have finished.


[Next](#)

Welcome


Welcome page, including information about payoffs and the random selection mechanism.

Instructions for Part 1 (game)

Let us now explain the game in greater detail. You will be randomly matched into groups such that each group consists of **four participants** (workers). You will stay in the same group for the entire part. Your payoff for a given decision depends on your decisions and the decisions of the other three participants in your group. You and everybody else in your group are receiving the same instructions, information, and are facing the same decision.



For these four workers, there are four tasks available, two of type A and two of type B. Every available task is linked to a potential payoff. These payoffs can vary within the group. Imagine the payoff as a representation of how much you enjoy being selected for a certain task.



Before we start, the computer will **randomly rank you and your three other group members** such that one of you is ranked first, one is second, one is third, and one is fourth. This ranking affects how the computer allocates you to the task of your choice. You will **not know your ranking** at the time of making your decision.

[Next](#)

Instructions (1)

First set of instructions, participants can navigate with *next* and *previous* buttons.

How to choose a task?

- If you are assigned to Task A, your payoff is fixed at w .
- If you are assigned to Task B, your payoff is 20% likely to be x , 20% likely to be y and 60% likely to be z .

The exact values of w , x , y , and z will differ from one decision to another, and you and your group members will always know what they are. For each of you, the computer will randomly select which of these three values applies in the given decision, and will do so such that your value of Task B does not depend on the values drawn for the other participants.

[Previous](#) [Next](#)

Instructions (2)

Text in grey bar differs across treatments. Here: *Partial Info treatment*

You will be asked to report which task you prefer. The computer will give you a recommendation about which task you should choose, such that the group payoffs are maximized. This gives you a hint about your true value of Task B: likely to be high or low. The recommendation will be displayed to you and all of your other group members.

After you submit your choice (which task you prefer):

- The participant ranked 1 is allocated to the task of their choice.
- The participant ranked 2 is allocated to the task of their choice.
- Therefore, after both participants 1 and 2 have made their choices, there are two open slots left.
- If both of the remaining slots are in the *same* task, participants ranked 3 and 4 get assigned to this task.
- If the remaining slots are in *different* tasks, the participant ranked 3 is assigned to their preferred task.
- The sole remaining slot is given to participant 4.

You will **not get any feedback** about your group members' task B values, assigned task, or earnings until the end of the experiment. In total, we will ask you to make two task choice decisions, each time with different values of w , x , y , and z .

[Previous](#) [Next](#)

Instructions (3)

Text in grey bar differs across treatments. Here: *Partial Info treatment*

No Info: "The computer will give you no information about your or anyone else's value."

Full Info: "The computer will give you private information about your true value of Task B. You will not learn your group members' values."

A bit more information about the recommendation

The computer is programmed to **maximize the total payoff to workers** by finding the best task allocation for the group. The computer knows the task B values (whether **x**, **y**, or **z** is selected) of each participant.

For example, the computer could conclude that participants 1 and 2 should be allocated to Task A, and participants 3 and 4 to Task B. This information is then shown to all of you.

The recommendation is giving you **partial information** about your own value of Task B. Consider: If the recommendation links you to Task A, is your value of B more likely to be high or low? Notice that regardless of what the computer shows in the recommendation, you are still free to choose either Task A or B.

[Read more](#)

In case there are multiple possibilities that would maximize the total payoff of your group, the computer will show you one possibility, picked at random.

For example, suppose that participant 1's value of Task B is 100, while the three other group members value Task B at 500. In this example, the value of Task A is always 300 for each participant. In that case, there are three possibilities how to allocate workers to tasks that maximize the group's total earnings:

- Participants 1 & 2 are assigned to Task A, participants 3 & 4 to Task B.
- Participants 1 & 3 are assigned to Task A, participants 2 & 4 to Task B.
- Participants 1 & 4 are assigned to Task A, participants 2 & 3 to Task B.

Since the computer can only display one option, it will choose one of the above three at random.

[Previous](#) [Next](#)

Instructions (3a)

Only shown in the *Partial Info* treatment.

Second half of page is only shown when *Read more* button is clicked.

An example of task choice

We will go through an example to illustrate how the task choice and allocation works. This example has the same number of workers tasks, and required workers per task as the actual decisions you will make.

Suppose that the computer ranks the workers as follows: 1 - 2 - 3 - 4 and that the workers submit the following:

Worker (rank)	1	2	3	4
Most preferred task	A	B	A	A

Allocation step 1: Participants ranked 1 and 2 get assigned to their favorite task:

Assigned task	A	B	-	-
---------------	---	---	---	---

Allocation step 2: Since the remaining slots are in *different* tasks, the participant ranked 3 gets to choose their favorite task between these:

Assigned task	A	B	A	-
---------------	---	---	---	---

Allocation step 3: The sole remaining task is assigned to participant 4:

Assigned task	A	B	A	B
---------------	---	---	---	---

[Previous](#) [Next](#)

Instructions (4)

Would it help to select B if you actually prefer A?

Short answer: **no**.

[check what happens if you would](#)

- If you are a participant ranked 1 or 2, and you reported Task B as your most preferred task, you get assigned to task B, and therefore do not get the task you most wanted - Task A.
- If you are the participant ranked 3, and the two slots available after participants ranked 1 and 2 have made their picks are in *different* tasks, if you report Task B as your most preferred task, you get assigned to Task B, and therefore do not get the task you most wanted - Task A.
- If you are the participant ranked 3, and the two slots available after participants ranked 1 and 2 have made their picks are in the *same* task, your report has no bearing on the task you are assigned to, so you do not get an advantage by reporting B as your most preferred task when you really want A.
- If you are the participant ranked 4, your report has no bearing on the task you are assigned to, so you do not get an advantage by reporting B as your most preferred task when you really want A.

[Previous](#) [I understand, start the Practice Decision](#)

Instructions (5)

Final page of instructions, including an answer to a frequently asked question.

More information (4 bullet points) only shows when the *check what happens if you would* button is clicked.

Practice Decision

[check instructions](#)

- There are four tasks available (two of type A and two of type B) for a total of four workers.
- You will be selected for only one task.
- All workers will be randomly ranked, but you currently do not know your rank.

The payoff of Task A is always 50 points.

Note that **for you**, the potential payoff of Task B is either 20, 40 or 60 points, with different probabilities, as given below:

20 points	40 points	60 points
20%	60%	20%

The following table summarizes the decision. Please click on one of the buttons to indicate your preference.

	Task A	Task B
Potential Payoff	50 points with 100%	20 points with 20% 40 points with 60% 60 points with 20%
Your preferences	<input type="button" value="I prefer A"/>	<input type="button" value="I prefer B"/>

Practice Decision

A pop-up screen with instructions is always available via the right top button.

No Info

HINT: The computer recommends that you select Task A

The following table summarizes the decision. Please click on one of the buttons to indicate your preference.

	Task A	Task B
Potential Payoff	50 points with 100%	20 points with 20% 40 points with 60% 60 points with 20% (but check the recommendation)
Your preferences	<input type="button" value="I prefer A"/>	<input type="button" value="I prefer B"/>

Partial Info treatment

	Task A	Task B
Potential Payoff	50 points	20 points
Your preferences	<input type="button" value="I prefer A"/>	<input type="button" value="I prefer B"/>

Full Info

Results of the Practice Decision

[check instructions](#)

- The potential payoff of Task B was 20 points (20%), 40 points (60%) or 60 points (20%).
- The computer selected a payoff of **20 points for Task B for you**.
- You were ranked **third**.
- The payoff of **Task A was 50 points**.
- You entered that you prefer Task A.

The computer selected the following allocation:

Rank	Player	Preference	Allocation
1	player 1 (computer)	B	allocated B
2	player 2 (computer)	A	allocated A
3	player 3 (you)	A	allocated A
4	player 4 (computer)	A	allocated B

Comprehension Questions

In the Practice Decision, you selected A and you earned 50 points.

[Show questions](#)

Comprehension Question 1/4

- Suppose you would have chosen task B. How much would you have earned?

20 points ▾

Comprehension Question 2/4

- Suppose the two players ahead of you in the ranking both selected Task A, just like you. Would you have been assigned task A?

No ▾

Comprehension Question 3/4

- Players 1 and 2 chose B and A. Could player 4 improve their payoff by reporting something other than what they reported (A)?

No ▾

Comprehension Question 4/4

- I will know the exact payoff of Task B in points.

False ▾

Results of the Practice Decision

Showing the preferences and allocations of the other players (computerized) and the participant (player 3).

Comprehension Questions are shown one-by-one on the same page after button click.

Comprehension Questions

Software counts the number of attempts. Correct answer currently selected, except for question 4/4, which depends on treatment:

- *No Info*: The distribution (possible payoff + likelihood of being selected)
- *Partial Info treatment*: No info + the computer will give a recommendation
- *Full Info*: The exact payoff in points

Decision 1

[check instructions](#)

Please state your preferences for Task A and Task B in the following form. Please note:

- There are four tasks available (two of type A and two of type B) for a total of four workers.
- You will be selected for only one task.
- All workers will be randomly ranked, but you currently do not know your rank.
- Each point translates to £0.0032 of bonus payment, so you can earn between £0.32 and £1.38 for this decision.

The payoff of Task A is always 300 points.

Note that **for you**, the potential payoff of Task B is either 100, 300 or 500 points, with different probabilities, as given below:

100 points	300 points	500 points
20%	60%	20%

The following table summarizes the decision. Please click on one of the buttons to indicate your preference.

	Task A	Task B
Potential Payoff	300 points with 100%	100 points with 20% 300 points with 60% 500 points with 20%
Your preferences	<input type="button" value="I prefer A"/>	<input type="button" value="I prefer B"/>

HINT: The computer recommends that you select Task B

The following table summarizes the decision. Please click on one of the buttons to indicate your preference.

	Task A	Task B
Potential Payoff	300 points with 100%	100 points with 20% 300 points with 60% 500 points with 20% (but check the recommendation)
Your preferences	<input type="button" value="I prefer A"/>	<input type="button" value="I prefer B"/>

The payoff of Task A is always 300 points.

The potential payoff of Task B is either 100, 300 or 500 points. The computer has selected the following value for you: **500 points**.

The following table summarizes the decision. Please click on one of the buttons to indicate your preference.

	Task A	Task B
Potential Payoff	300 points	500 points
Your preferences	<input type="button" value="I prefer A"/>	<input type="button" value="I prefer B"/>

Decision 1

After Decision 1, a large yellow-pop up indicates that a new independent decision is coming up.

Decision 2 is almost identical to Decision 1, except for lower valuations [0, 200, 400] versus [100, 300, 500].

No Info

Partial Info treatment

Full Info

Some feedback

Please explain how you made your decisions in the task choice.

Strategy

Self reported strategy. Under Partial Info treatment, one sentence is added:
For example, how did you use the recommendation (if at all)?

Bonus task

Consider the following situation:

Ben has two plastic storage containers full of cookies. The blue container has ten chocolate-covered cookies and thirty regular ones. The red container has twenty cookies of each type. Ben picks one of the containers at random. Then without looking, he randomly takes a cookie from that container. The cookie is a regular one. Ben wonders which of the two containers he originally selected. The regular cookie is one piece of evidence.

Given that the cookie he selected was regular, what is the probability that the cookie came from the blue container?

If your answer is correct and this task is selected for payment, you will earn a £1.60 bonus. Please give us your best estimate.

I think that the probability is %

I don't know how to estimate this and I don't want to guess, please take me to the next task.

Bayesian task

Correct answer: 60%.

Based on Mellers et al. (2017)

<http://sjdm.org/journal/17/17408/jdm17408.pdf>

Instructions Box Collecting Task

In this next task, you will be asked to collect boxes. There are a total of 100 boxes that you can choose to collect. Every box that you collect will earn you additional £0.02, with one exception: **there is one bomb hidden in one of the boxes**, and you do not know where. If you collect that box, it will explode, and you will lose all your earnings from this task.

You will learn whether you collected the bomb **at the end of the experiment**. Example: Suppose that you choose to collect 50 boxes. The computer will randomly select 50 boxes for you. If the bomb is hidden in any of these fifty boxes, you will earn nothing. If the bomb is not hidden in any of your boxes, you will earn $0.02 \times 50 = £1.00$.

Notice that collecting all 100 boxes results in zero earnings from this task for sure; you know that there is a bomb in one of the boxes, and so if you collect all of them, you are guaranteed to find the bomb.

Instructions Risk Elicitation Task

Bomb Risk Elicitation Task for oTree by Holzmeister & Pfurtscheller (2016)

<http://dx.doi.org/10.1016/j.jbef.2016.03.004>

Box Collecting Task

Instructions

Here are 100 boxes that you can choose to collect, and one of them contains a bomb. Every collected box earns you £0.02 unless you collect the box with the bomb – in that case the bomb explodes, and you earn nothing.

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

How many boxes would you like to collect?

Confirm

Risk Elicitation Task

Participants can use the arrows or simply type to select a number of boxes to be opened.

Box Collecting Task 2

We will play the Box Collecting Task once again. Before you make your decision **in this round** about the number of boxes you would like to collect, we would like to ask you to **please consider collecting as many boxes as you dare**, as this would be really helpful for our research.

Experimenter Demand Task

Participants are asked to do the Box Collecting Task again and to open as many boxes as possible. When they open more than in the previous round, we take this as a signal of experimenter demand.

Final Questionnaire

What is your occupation?

- Full-time job
- Part-time job
- Student
- Retired
- Caretaker / voluntary work
- Other:

What is your gender?

- Male
- Female
- Non-binary
- Rather not say

It is important for research that participants are paying attention. If you are reading this, please select Australia, regardless of your country of birth.

- United States
- United Kingdom
- South Africa
- Australia
- New Zealand

This is the end of the survey. In case you have comments, please leave them here.

Final Questionnaire

Questions: occupation, gender, attention check, final comments.

Additional bonus

The payoffs of Decision 1 and 2 will be calculated later, when your group **of four participants** is complete. The earnings of the group members may vary, depending on decisions and random selection of the rankings by the computer.

Thank you for completing the experiment. As a final thank you from us, we are offering you an additional bonus. Please choose which do you prefer:

Additional £0.20 for **myself**

Additional £0.10 for each of the **others** in my group (£0.30 in total)

Social Preferences Task

Payoffs

We are almost done with the experiment, thank you for participating! Here is a summary of your bonus earnings in each task. The task indicated in green has been randomly selected for payment. You will also get a show-up fee of £1.50 and an additional bonus of £0.20 (possibly more, depending on choices of others).

Bonus Task	Payoff
Decision 1	will be determined once all participants finished the game (ranges between £0.32 and £1.28)
Decision 2	will be determined once all participants finished the game (ranges between £0 and £1.60)
Cookie question	£0.00 (correct answer was 60%)
Box collecting task 1	£0.10 (5 boxes, no bomb)
Box collecting task 2	£0.04 (2 boxes, no bomb)

Next

Payment Overview

One task has been randomly selected for payment (indicated in green).

Thank you for participating

Your final payoff is £1.50 (show-up fee) + £0.20 (additional bonus) + bonus, which will be determined once all participants finished the game

Back to Prolific

End of experiment

C Behavioral extensions

In this section we show how risk attitudes, imperfect Bayesian updating, and social preferences influence agents' behavior in our game. We also run simulations to show the magnitude of the impact of these deviations from the baseline model on group welfare.

C.1 Risk preferences

Suppose agents can be risk averse, risk neutral or risk loving, with their utility function following the standard CRRA form given below. x denotes the monetary value of the agents' earnings, and $\gamma \in [-1, 1]$ is the relative risk aversion coefficient:

$$u(x; \gamma) = \begin{cases} \frac{x^{1-\gamma}-1}{1-\gamma}, \gamma \in [-1, 1) \\ \log(x), \gamma = 1. \end{cases} \quad (1)$$

It can be shown that risk attitudes – as long as they are within a reasonable range one would expect based on past experimental research (see e.g., [Holt and Laury \(2002\)](#); [Anderson and Mellor \(2008\)](#); [Andersen et al. \(2008\)](#); [Bombardini and Trebbi \(2012\)](#); [Charness et al. \(2020\)](#))²⁶ – do not affect predicted behavior in the *Partial* and *Full Info* settings.

Proposition 2. *Theoretical predictions in the Partial and Full Info treatments are unaltered when the subjects' utility function is given by (1), for all $\gamma \in [-1, 1]$.*

See Appendix [E.1](#) for proof.

It is worth noting that in the *No Info* treatment, risk preferences do alter predicted reports by agents in Scenario 1: In this case, the expected earnings from options B and A are the same. Therefore, agents prefer A (B) if and only if they are risk averse (risk loving). But since under *No Info*, the predicted reports by all agents are the same and independent of realized preferences, this makes no difference in agents' *expected* earnings, and therefore in our hypotheses.

C.2 Non-Bayesian updating: Prior bias

In this subsection we analyze how imperfect Bayesian updating can affect agents' choices under the *Partial Info* treatment. We assume that imperfect belief updating

²⁶Note that there is another popular specification of the CRRA utility function with $1 - \gamma$ as the coefficient of relative risk aversion rather than γ like in our paper ([Wakker, 2008](#)). Of course, once the reader adjusts for the different functional specification, results regarding people's preferences remain similar.

has no effect on agents' choices under the *Full Info* treatment, as they directly observe their valuations for B .²⁷

We use a model of *prior bias* to capture the possibility that agents can use updating rules that are only partially Bayesian. Prior bias, also known as conservative Bayesianism (Edwards, 1968) or inertia, captures inferences drawn in favor of the prior belief, and it belongs under the umbrella of confirmation biases. Related experimental evidence is well documented in the psychology literature (e.g., Pitz et al. (1967), Geller and Pitz (1968), Pitz (1969)).²⁸ We use the following model of a prior-biased updating rule, following Epstein (2006), who provides an axiomatic foundation for prior bias:

$$q(\cdot|s) = (1 - b)p(\cdot|s) + bp_0 \quad (\text{Prior Bias})$$

where p_0 is the prior belief and $p(\cdot|s)$ is the Bayesian update of p_0 after receiving a signal s . In our case, since we consider only the *Partial Info* treatment, s can only be a recommendation of A or B . Hence, $q(\cdot|s)$ captures the prior-biased updated belief where $b \in [0, 1]$ is the degree of prior bias.

We show that as long as the prior bias is not more than 60%, our baseline predictions still hold, as long as the agents' risk aversion is within our assumed range of $\gamma \in [-1, 1]$.

Proposition 3. *For prior bias $b < 60\%$, baseline predictions hold.*

The intuition for this result is as follows: In Scenario 1, risk neutral and risk loving agents weakly prefer B a priori. They naturally prefer it also when they are recommended B . Hence, regardless of the prior bias, such agents accept B when recommended. Risk averse agents, on the other hand, prefer A a priori, but prefer B when B is recommended, when fully Bayes-rational ($b = 0$). Hence these agents reject the recommendation of B if the prior bias is too high. The exactly opposite case arises when the agents are recommended A .

Putting the two cases in Scenario 1 together, we obtain the threshold of prior bias depicted in Figure C1 below. In particular, it plots the maximum prior bias which still allows baseline predictions to hold, as a function of risk preferences.

²⁷Technically, imperfect belief updating can affect agents' choices under the *Full Info* treatment as well if we think of an agent's observed information as only a signal, even if an agent is told that it is supposed to be fully revealing. We assume away this possibility.

²⁸See de Clippel and Zhang (2022) for how prior bias affects the optimal signal in information design settings with a single decision-maker. Our designer does not take potential prior bias into account in her design.

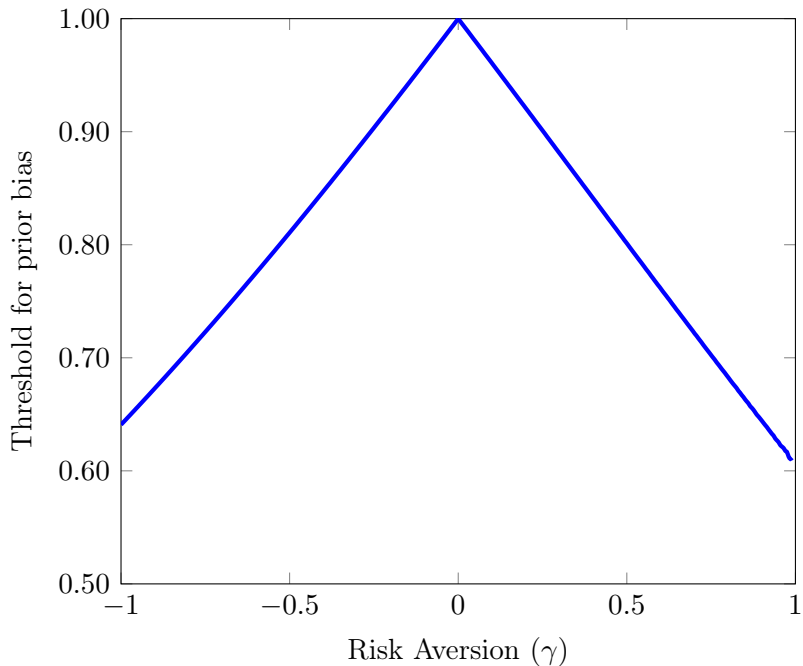


Figure C1: Bounds of prior bias for baseline predictions to hold in Scenario 1. The same bounds apply to the probability of the social planner mistakenly observing preferences of the agents (see below).

In Scenario 2, the prior and posterior preferred object of each agent is the same — object A — regardless of the recommendation, even when they are fully Bayes-rational. Prior bias brings posterior preferences closer to prior preferences, thereby leaving Scenario 2 preferences unchanged. Hence Scenario 2 imposes no upper bound on the tolerance of our model for prior bias. For proof, see Appendix E.2.

Having established that baseline predictions no longer hold if the prior bias is sufficiently high, let us highlight the benchmark extreme case of $b = 1$. In this case the agents do not update their estimates of A and B based on the recommendation at all. In Scenario 1, such agents always choose A (respectively, B) under *Partial Info* if they are risk averse (respectively, risk-loving), based on prior expected utilities. For the same reason, they always choose A (i.e., behave the same way as predicted under the baseline model) in Scenario 2. This is summarized in the proposition below.

Proposition 4. *Fully prior-biased ($b = 1$) agents' predicted choices in the two scenarios are as follows:*

- *Scenario 1: They always choose A (respectively, B) under *Partial Info* if they are risk averse (respectively, risk-loving). Either choice is possible if they are risk neutral.*
- *Scenario 2: They always choose A .*

In summary, sufficiently prior biased agents play the “Always safe” strategy of choosing A in both scenarios, in the usual case when they are weakly risk averse.

To illustrate the impact of the *Always safe* strategy on group welfare, consider Figure C2: It compares the aggregate welfare levels of *Full Info* and *Partial Info* treatments for different shares of agents in *Partial Info* exhibiting the behavior of always choosing the safe option A. Each blue data point (diamond) depicts the average group payoff of a simulation of 2500 groups by 10 000 computerized agents (*bots* in the oTree software). At the 0% mark on the left, all agents are programmed with the optimal strategy: follow the recommendation in Scenario 1 and select the safe option in Scenario 2. The other marks on the x-axis show the results for increasing shares of agents who play *Always safe*. For example, in Scenario 1 under 20% *Always safe*, there are approximately 2000 agents who always choose A, and 8000 agents who follow the recommendation. Since the agents are randomly allocated to groups, there might be some groups in which all agents follow the recommendation, but also some where all choose A.

We can contrast this with the average group payoff based on the same valuations under the *Full Info* treatment, depicted in orange (squares). Here, each computerized agent ‘selected’ the option with the highest payoff and a random choice was implemented in case of a tie. Notice that the orange lines are horizontal by definition.

The grey (green) triangles represent the aggregate group welfare when the group achieves the maximum (minimum) possible payoff by allocating B to the two agents with the highest (lowest) B-valuations, and A to the remaining agents. These lines are likewise horizontal by definition because they do not depend on the agents’ behavior (but note some variation due to the random selection of valuations by the software).

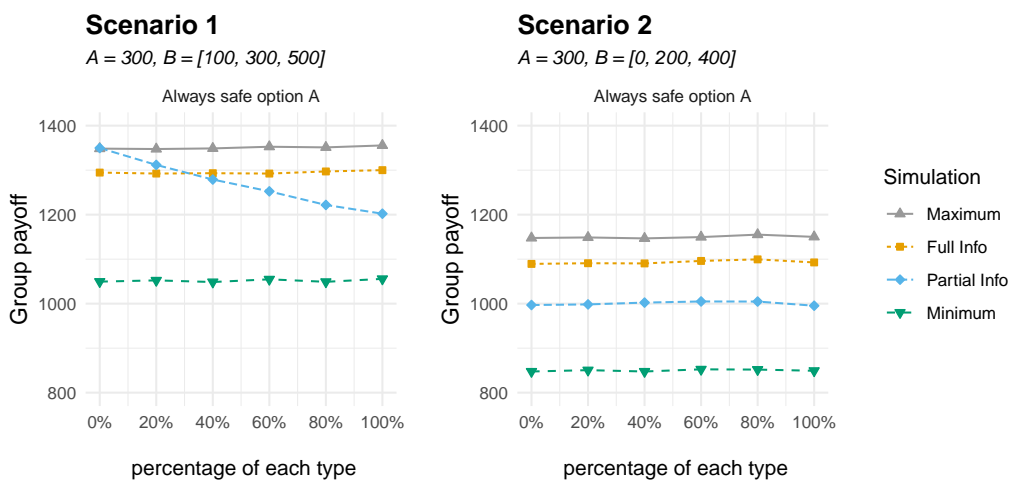


Figure C2: Average group welfare across scenarios and types of agents, *Always safe*

Finally, we note that prior bias could be interpreted as the degree of *mistrust* agents have for an information source (Lee et al., 2023). In other words, the agents

would believe that the planner does not observe the agents’ preferences precisely – with some probability, she observes a potentially different preference profile, chosen uniformly at random from all preference profiles. As a result, their updated beliefs would become biased towards the prior, and thus the bound derived for prior bias above doubles up as the bound for the maximum probability of random mistakes of the planner for which baseline predicted behaviors remain unchanged.

C.3 Social preferences

Departing from the assumption of perfect self interest, we now assume the agents care about a weighted average of their own payoff and their group’s payoff.²⁹

Let u_i denote the utility of agent i purely from the allocation he gets, regardless of others’ allocation. Using v_i to denote the overall utility of agent i , taking into account his regard for others,

$$\begin{aligned} v_i &= (1 - a)u_i + a \sum_j u_j \\ &= u_i + a \sum_{j \neq i} u_j. \end{aligned} \tag{Social Preferences}$$

As given above, a is the weight an other-regarding agent puts on the group’s payoff. In the calculations that follow, we assume other-regarding agents see others as similar to themselves, but only take into account the direct utility obtained by other agents, not their “social” utility. In particular, an agent with a risk aversion parameter γ and prior bias b assumes the same values for these parameters and $a = 0$ for all others (as captured by (Social Preferences)).

In general, we consider $a \in [-1, 1]$. It can be shown that spiteful agents ($a \in (-1, 0)$) behave identically as non-other-regarding agents ($a = 0$) in our setting. For brevity, we omit these details. For the rest of this subsection we focus on the case when agents are altruistic, i.e., $a \geq 0$.

C.3.1 Full Info

Whenever an agent i ’s report affects outcomes, using (Social Preferences) and the fact that agents’ preferences are i.i.d., we have:

$$\begin{aligned} \text{His expected payoff from picking B} &= u_{iB} + a(2u_A + \mathbb{E}u_{iB}), \\ \text{From picking A} &= u_A + a(u_A + 2\mathbb{E}u_{iB}). \end{aligned}$$

²⁹Notice that this could be also interpreted as a preference for “efficiency” which is distinct from social preferences. However, in what follows we do not separate these two interpretations.

Hence:

$$\begin{aligned}
& \text{The agent strictly prefers } B \text{ (} A \text{)} \\
& \iff u_{iB} + a(2u_A + \mathbb{E}u_{iB}) > (<)u_A + a(u_A + 2\mathbb{E}u_{iB}) \\
& \iff u_{iB} > (<)(1 - a)u_A + a\mathbb{E}u_{iB} \qquad \text{(Altruism preferences)}
\end{aligned}$$

Clearly, without social preferences ($a = 0$), we have:

$$\text{The agent strictly prefers } B \text{ (} A \text{)} \iff u_{iB} > (<)u_A \qquad \text{(Baseline preferences)}$$

Comparing (Baseline preferences) and (Altruism preferences), we conclude that (i) if $u_A = u_{iB}$, social preferences do not affect the agent's choices, and (ii) the choices with and without social preferences are different if:

$$(1 - a)u_A + a\mathbb{E}u_{iB} < u_{iB} < u_A \quad \text{or} \quad (1 - a)u_A + a\mathbb{E}u_{iB} > u_{iB} > u_A \qquad (2)$$

The first (second) inequality can hold only if $u_A > u_{iB}$ ($u_A < u_{iB}$).

It can be verified that for $a \in [0, 1]$, the above can hold only in Scenario 2 for risk averse agents with $u_{iB} = 200$. The threshold of altruism (a) for which baseline predictions hold in Scenario 2 under *Full Info* is depicted in Figure C3a. As we see, for highly risk averse agents even a little bit of altruism is sufficient for them to choose B in Scenario 2 when its value is 200, i.e., lower than A.

C.3.2 Partial Info

For simplicity, for belief updating — which is relevant only in the *Partial Info* treatment — we consider only the two extreme models, fully Bayesian (prior bias $b = 0$) and fully prior-biased ($b = 1$).

Intuitively, whenever a purely self-interested agent accepts a recommendation in the *Partial Info* setting, any altruistic agent ($a > 0$) should also do so, because agents know that the recommendations are given in order to maximize aggregate welfare. Hence, in Scenario 1, and in Scenario 2 upon being recommended A - cases where pure rational self-interest is sufficient for Bayesian agents to accept the welfare-maximizing recommendation — the predictions for altruistic agents ($a > 0$) are the same as those in the baseline model ($a = 0$). Fully prior-biased agents do not update their beliefs in response to the partial information signal. So, their assessment of their (and everyone else's) payoffs from each object is the same as their prior assessment. Hence, for each agent, the gain of another agent from him trading A for B (or the reverse) is exactly equal to their own loss from such a trade. Hence such a trade is not profitable to any non-Bayesian agent, as long as $a < 1$. This is summarized in the proposition below.

Proposition 5. *Altruism makes no difference in predicted behavior in the following cases:*

- *In Scenario 1 for fully Bayesian agents.*
- *Regardless of the Scenario for fully prior-biased agents.*

In Scenario 2, where purely self-interested Bayes-rational agents reject one of the welfare-maximizing recommendations, *sufficiently high* altruism may make them accept it instead. We derive thresholds of altruism for changing the Bayesian agents' decision from not accepting to accepting a Recommendation of B in Scenario 2. These are plotted below in Figure C3. In particular, for the risk neutral case, Bayesians in Scenario 2 need a in the above model of social preferences to be at least $\sim 15\%$. The reason the threshold falls from that level for both risk-averse and risk-loving agents is because of our assumption that agents see others as just as risk-averse/risk-loving as themselves. Hence, while the loss from making a sacrifice falls with (signed) risk aversion, so does the gain of the agent who is the beneficiary of such a sacrifice. Given our model parameters, the personal loss falls at a rate *lower* than the societal gain with increasing risk aversion, leading to agents becoming more willing to sacrifice for others *both* with increasing risk aversion and increasing risk-lovingness.

Detailed calculations are in Appendix E.3.

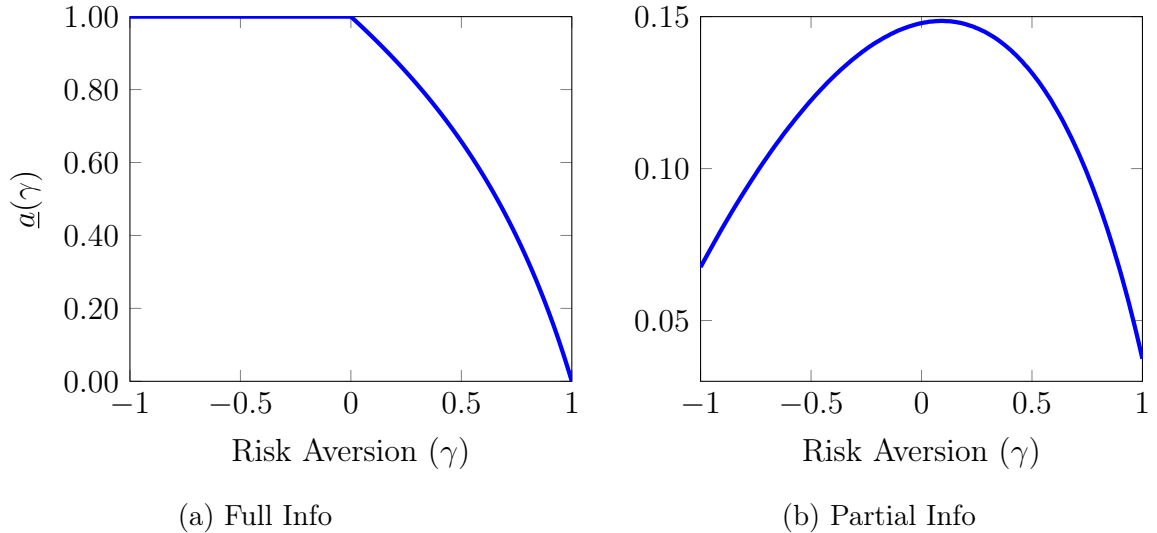


Figure C3: Threshold of altruism to accept B for Bayesians in Scenario 2. For prior-biased agents, the threshold would be higher, for each level of risk-aversion.

In summary, altruism pushes agents to follow the recommendation in both Scenarios. As in subsection C.2, we pair this insight with simulation results. In particular, we compare average aggregate payoffs under *Full* and *Partial Info* while varying

the share of agents who use the decision rule of always following the recommendation. Figure C4 below depicts the results, reflecting the theoretical conclusions of this subsection. Specifically, in Scenario 1 the decision rule of *Always Follow* makes no difference (flat lines in the figure on the left), but in Scenario 2, as the share of always following agents increases, so does the aggregate payoff (upwards-trending blue line). This underscores that altruism-driven collective recommendation following can increase group payoffs all the way until the maximum group payoff.

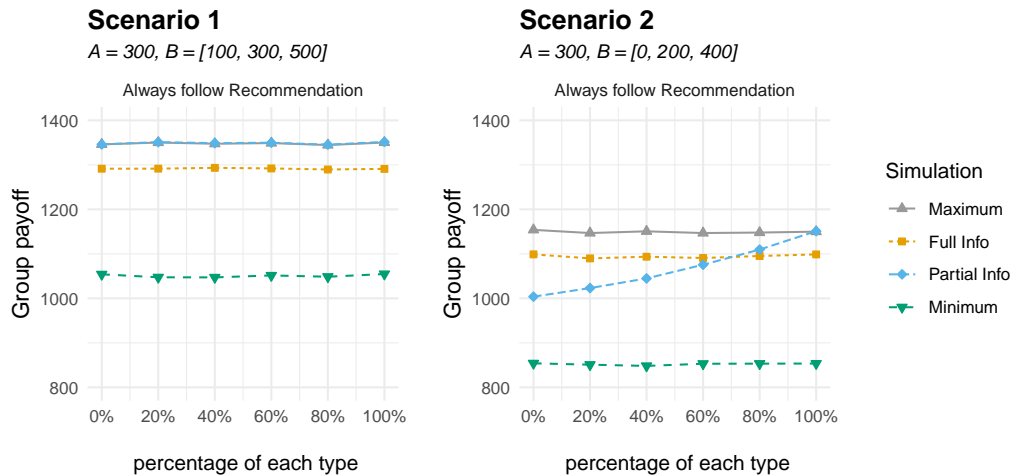


Figure C4: Average group welfare across scenarios and types of agents, *Always follow*

C.3.3 Experimenter demand

If experimenter demand effects (Zizzo, 2010) are present, agents follow the recommendation in an attempt to please the experimenter, regardless of their risk aversion, belief updating rules and social preferences. Therefore, in this sense, experimenter demand operates like social preferences, as delineated at the beginning of Section C.3. Hence, experimenter demand makes no difference to baseline predictions in Scenario 1, and makes agents more likely to accept the recommendation in Scenario 2. More specifically, using any appropriate model of experimenter demand, we can find its threshold for the agent’s acceptance of B in Scenario 2, just like in Figure C3b. Because of this analogy in the function of altruism and experimenter demand, the behavioral rule of *Always follow* depicted in the simulation results of Figure C4 can be also explained by experimenter demand. We omit the straightforward mathematical details for the sake of brevity.

Alternatively, the subjects may display a tendency to defy the experimenter on purpose. While we do not model this explicitly, the simulated average aggregate payoffs for varying percentages of agents always defying the recommendation are depicted in Figure C5 below. As expected, as the share of defying agents increases, the average aggregate payoff drops in both Scenarios. As the recommendation is designed to maximize the group payoff, it is not surprising that the opposite behavior

facilitates sub-optimal object allocations and therefore minimizes group payoffs when all agents follow that strategy.

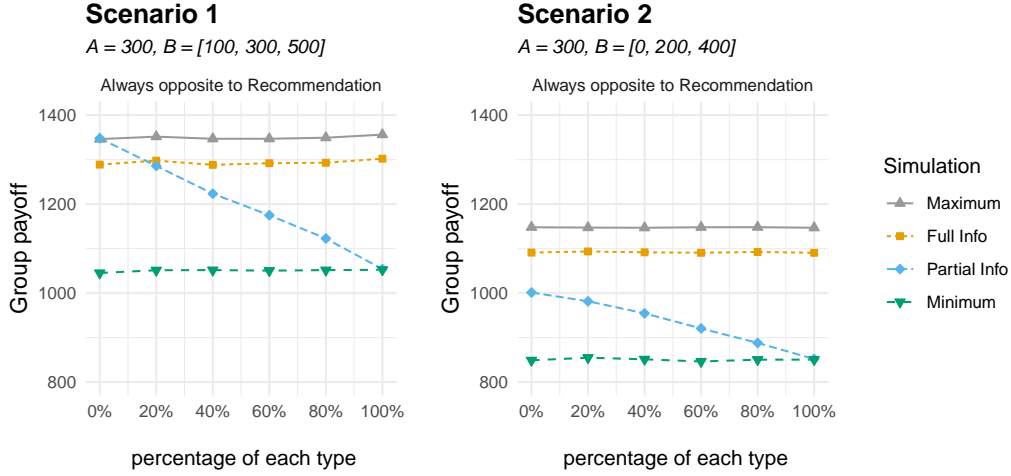


Figure C5: Average group welfare across scenarios and types of agents, *Always opposite*

An alternative explanation that could rationalize agents pursuing the *opposite* of what they were recommended would be a belief that with some probability, the planner mistakenly *swaps* the agent’s recommendations – i.e., recommends him A when his utilitarian allocation should have been B (or vice versa). Hence, unlike in the baseline case where a recommendation of $x \in \{A, B\}$ is always “good news” for his utility from x , in this case it is a mix of good and bad news. This is because a recommendation of, say, B , means that with some probability, the correct recommendation should have been A , which is “bad news” for his utility from B . Hence, if such mistakes are too likely, recommendations end up giving the exact opposite of their intended information to recipients — and hence upend our predictions. The maximum tolerable mistake probability for our baseline predictions to hold, as a function of the risk aversion parameter γ — called $\bar{\delta}_S(\gamma)$ — is plotted in Figure C6.

The intuition behind such a pattern is as follows. An observed recommendation of each of A and B is a mixture of its “actual” recommendation and a recommendation of the opposite object. Hence, unless correct recommendations are *at least* as likely as incorrect ones, the meaning of each recommendation flips and our predictions are flipped as well. This puts an upper bound of 50 % on the maximum tolerable mistake probability, as we want agents to accept recommendations in Scenario 1. In Scenario 2, A is preferable to agents both when it is recommended and when B is recommended. Hence any mixture of these two recommendations makes no difference to the agent, and our predictions continue to hold, for *any* mistake probability. Therefore Scenario 2 imposes no additional restriction on the maximum tolerable mistake probabilities.

In Scenario 1, B has to be favored (respectively, disfavored) compared to A

when it is recommended (respectively, when A is recommended). This can put tighter bounds on the tolerable mistake probability — and does, if and only if the agents are risk averse, as depicted in Figure C6.

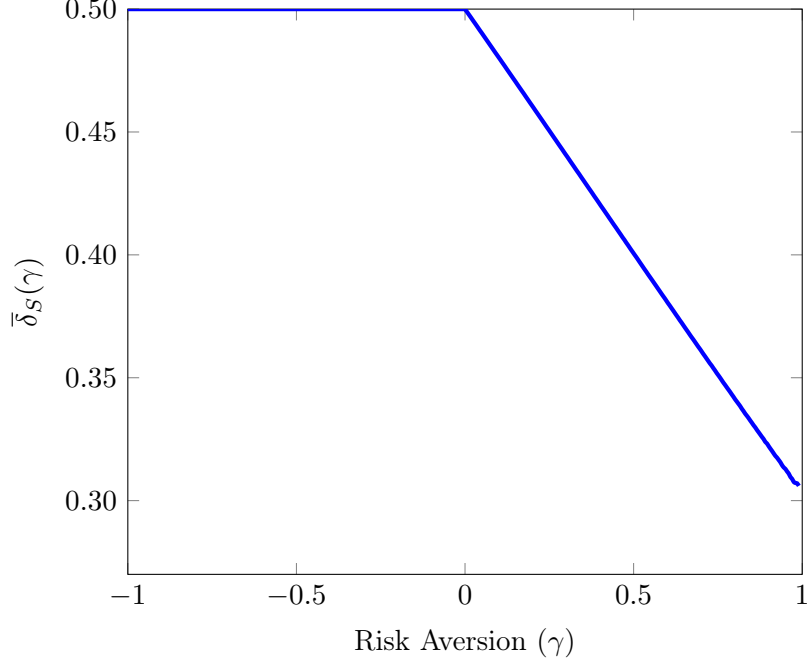


Figure C6: The maximum probability with which swap mistakes can occur for baseline predictions to hold in Scenario 1

D Proofs of main theoretical predictions

D.1 Generalization and proof of Proposition 1

Let us call $v_{ix} := u_{ix} - u_{iy}$ agent i 's *relative preference* for object x where $x, y \in \{A, B\}$, $x \neq y$. As in the main text, we use \hat{u}_i to denote i 's relative preference for A . Hence $v_{iA} = \hat{u}_i$ and $v_{iB} = -\hat{u}_i$ for all i .

We need additional notation to define the generalized notion of *no strong a priori preferences (NSAP)*, introduced in the main body. For all i and $x \in \{A, B\}$, let:

$$\begin{aligned} \mathcal{U}_{+,x,i} &= \{u \in \mathcal{U} : \#\{j : v_{ix} > v_{jx}\} \geq n\}, \\ \mathcal{U}_{0,x,i} &= \{u \in \mathcal{U} : \#\{j : v_{ix} > v_{jx}\} < n, \#\{j : v_{ix} \geq v_{jx}\} \geq n\}, \\ \mathcal{U}_{-,x,i} &= \{u \in \mathcal{U} : \#\{j : v_{ix} < v_{jx}\} \geq n\}. \end{aligned}$$

Note that $\mathcal{U}_{0,x,i} = \mathcal{U}_{0,y,i}$. Let us therefore denote $\mathcal{U}_{0,x,i}$ by $\mathcal{U}_{0,i}$ going forward.

Further, let $F_{=x,i}$ denote the CDF of v_{ix} conditional on $\mathcal{U}_{-,x,i}$, i.e. $F_{=x,i} : V \rightarrow [0, 1], v \mapsto \mu(\{v_{ix} \leq v | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\})$.

In an i.i.d. setting with two options, we say agents have no strong a priori preferences if there exists $q \in [0, 1]$ such that:

$$\mathbb{E}(v_{ix} | \mathcal{U}_{+x,i} \cup \{\mathcal{U}_{0,i}, F_{=x,i}(v_{ix}) = q1_{\{x=A\}} + (1-q)1_{\{x=B\}}\}) \geq 0, x \in \{A, B\}. \quad (\text{NSAP})$$

Agent i is said to have no strong a priori preferences if there exists $q \in [0, 1]$ such that, when he knows his relative preference for object A (B) is “strongly above the median” ($\#\{j : v_{ix} > v_{jx}\} \geq n$ for either x) or “weakly above the median” ($\#\{j : v_{ix} > v_{jx}\} < n$ and $\#\{j : v_{ix} \geq v_{jx}\} \geq n$ for either x) with its value lying in the top q ($1 - q$) *quantiles*, measured according to the prior conditional on lying weakly above the median, he prefers A (B).

Note that if the prior μ is atomless, $\mu(\mathcal{U}_{0,i}) = 0$, hence in (NSAP), $\mathbb{E}(v_{ix} | \mathcal{U}_{+x,i} \cup \{\mathcal{U}_{0,i}, F_{=x,i}(v_{ix}) = q1_{\{x=A\}} + (1-q)1_{\{x=B\}}\}) = \mathbb{E}(v_{ix} | \mathcal{U}_{+x,i})$, i.e., the definition of NSAP boils down to the simpler, rank-based definition given in the main text.

With this generalized definition of NSAP, the statement of Proposition 1 remains the same. We provide its proof below.

Claim 1. *Under any aggregate welfare maximizing signal, i is recommended A (respectively, B), if:*

$$\#\{j : \hat{u}_i > \hat{u}_j\} \geq n \text{ (respectively, } \#\{j : \hat{u}_i < \hat{u}_j\} \geq n),$$

and only if:

$$\#\{j : \hat{u}_i \geq \hat{u}_j\} \geq n \text{ (respectively, } \#\{j : \hat{u}_i \leq \hat{u}_j\} \geq n).$$

Proof. “Only if” part. Suppose agent $i \in I$ is recommended A under the aggregate welfare maximizing allocation. Fix any other agent $j \in I \setminus i$ who has been recommended B . i knows that keeping the allocation of every agent in $I \setminus \{i, j\}$ the same as what they have been recommended, exchanging i and j ’s allocation would weakly reduce aggregate utility. That is,

$$\begin{aligned} u_{iA} + u_{jB} &\geq u_{iB} + u_{jA} \\ \iff \hat{u}_i &\geq \hat{u}_j \end{aligned}$$

The above must hold for every i who has been recommended A and every j who has been recommended B . Hence $|\{j : \hat{u}_i \geq \hat{u}_j\}| \geq n$ for all i who have been recommended A .

“If” part. Conversely, if i is recommended B , by the above result this implies,

$$\begin{aligned}
& \#\{j : -\hat{u}_i \geq -\hat{u}_j\} \geq n \\
& \iff \#\{j : \hat{u}_i \leq \hat{u}_j\} \geq n \\
& \implies \#\{j : \hat{u}_i > \hat{u}_j\} \leq n - 1, \text{ i.e. } \#\{j : \hat{u}_i > \hat{u}_j\} < n.
\end{aligned}$$

The last line comes from the fact that $\#\{j : \hat{u}_i > \hat{u}_j\} + \#\{j : \hat{u}_i \leq \hat{u}_j\} = 2n - 1$. Hence the “if” part follows. \square

Before proceeding further we need to formalize recommendation signals, as defined in the main text. A recommendation signal can be described as a vector-valued function $R : \mathcal{U} \rightarrow [0, 1]^I$ where the i -th component of $R_i(u)$ captures the probability of recommending A to agent i when the realized preference profile is $u \in \mathcal{U}$.

By Claim 1, a recommendation signal must recommend x to i if $(u_i, u_{-i}) \in \mathcal{U}_{+x,i}$. Following the definition of NSAP, let us define a *cutoff* recommendation signal as a recommendation signal for which there exists v and $q \in [0, 1]$ such that, for $(u_i, u_{-i}) \in \mathcal{U}_{0,i}$ it recommends x to i with probability 1 if $v_{ix} > v$, with probability q if $v_{ix} = v$ and with probability 0 otherwise.

Claim 2. *For any recommendation signal R , there exists a cutoff recommendation signal R^c such that:*

$$\mathbb{E}(\hat{u}_i | A, R_i^c) \geq \mathbb{E}(\hat{u}_i | A, R_i) \text{ and } \mathbb{E}(\hat{u}_i | B, R_i^c) \leq \mathbb{E}(\hat{u}_i | B, R_i).$$

Proof. Let the total probability of recommending A and B to i , under the recommendation signal R , conditional on $\mathcal{U}_{0,i}$ be p_A and p_B respectively. Hence $p_A + p_B = 1$. Let:

$$\begin{aligned}
v_A &= \inf\{\tilde{v} : \mu(\{\hat{u}_i > \tilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) < p_A\} \\
v_B &= \sup\{\tilde{v} : \mu(\{\hat{u}_i < \tilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) < p_B\}
\end{aligned}$$

We first claim that $v_A = v_B$. To see why, note that $\forall \tilde{v} < v_B, \mu(\{\hat{u}_i < \tilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) < p_B$, i.e. $\mu(\{\hat{u}_i \geq \tilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) > p_A$. Hence, $\tilde{v} \leq v_A$ for all $\tilde{v} < v_B$. $\therefore v_B \leq v_A$.

Suppose $v_A > v_B$. Hence, for all $\tilde{v} \in (v_B, v_A)$:

$$\begin{aligned}
& \mu(\{\hat{u}_i > \tilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) \geq p_A, \text{ and} \\
& \mu(\{\hat{u}_i < \tilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) \geq p_B.
\end{aligned}$$

Adding the above two inequalities we have:

$$1 \geq \mu(\{\widehat{u}_i > \widetilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) + \mu(\{\widehat{u}_i < \widetilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) \geq p_A + p_A = 1.$$

The above is possible only if, for all $\widetilde{v} \in (v_B, v_A)$, $\mu(\{\widehat{u}_i > \widetilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) + \mu(\{\widehat{u}_i < \widetilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) = 1$, i.e. $\mu(\{\widehat{u}_i = \widetilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) = 0$. Moreover, $\mu(\{\widehat{u}_i > \widetilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) = p_A$, and $\mu(\{\widehat{u}_i < \widetilde{v} | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) = p_B$. This implies:

$$\begin{aligned} \mu(\{\widehat{u}_i \in (v_B, v_A) | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) &= 0, \\ \mu(\{\widehat{u}_i \geq v_A | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) &= p_A, \\ \mu(\{\widehat{u}_i \leq v_B | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) &= p_B. \end{aligned}$$

If $v_A > v_B$, pick any cutoff $v \in (v_B, v_A)$ and $q \in [0, 1]$. If $v_A = v_B$, there exists $q \in [0, 1]$ such that:

$$\mu(\{\widehat{u}_i > v_A | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) + q\mu(\{\widehat{u}_i = v_A | (u_i, u_{-i}) \in \mathcal{U}_{0,i}\}) = \int_{\{u \in \mathcal{U}_{0,i}\}} R_i^c(u | \mathcal{U}_{0,i}) d\mu(u | \mathcal{U}_{0,i})$$

The cutoff recommendation signal, as constructed above - let us call it R^c - has the property that A and B are each recommended with the same probability on $\mathcal{U}_{0,i}$ under R^c as under R .

Clearly, by construction, $\mathbb{E}(\widehat{u}_i | \mathcal{U}_{0,i}, R^c) \geq \mathbb{E}(\widehat{u}_i | \mathcal{U}_{0,i}, R)$ and $\mathbb{E}(\widehat{u}_i | \mathcal{U}_{0,i}, R^c) \leq \mathbb{E}(\widehat{u}_i | \mathcal{U}_{0,i}, R)$.

$$\begin{aligned} \mathbb{E}(\widehat{u}_i | A, R^c) &= \frac{\mathbb{E}(\widehat{u}_i | \mathcal{U}_{+A,i})\mu(\mathcal{U}_{+A,i}) + \mathbb{E}(\widehat{u}_i | \mathcal{U}_{0,i}, R^c) \int_{\{u \in \mathcal{U}_{0,i}\}} R_i^c(u | \mathcal{U}_{0,i}) d\mu(u | \mathcal{U}_{0,i})}{\mu(\mathcal{U}_{+A,i}) + \int_{\{u \in \mathcal{U}_{0,i}\}} R_i^c(u | \mathcal{U}_{0,i}) d\mu(u | \mathcal{U}_{0,i})} \\ &\geq \frac{\mathbb{E}(\widehat{u}_i | \mathcal{U}_{+A,i})\mu(\mathcal{U}_{+A,i}) + \mathbb{E}(\widehat{u}_i | \mathcal{U}_{0,i}, R) \int_{\{u \in \mathcal{U}_{0,i}\}} R_i^c(u | \mathcal{U}_{0,i}) d\mu(u | \mathcal{U}_{0,i})}{\mu(\mathcal{U}_{+A,i}) + \int_{\{u \in \mathcal{U}_{0,i}\}} R_i^c(u | \mathcal{U}_{0,i}) d\mu(u | \mathcal{U}_{0,i})} \\ &= \frac{\mathbb{E}(\widehat{u}_i | \mathcal{U}_{+A,i})\mu(\mathcal{U}_{+A,i}) + \mathbb{E}(\widehat{u}_i | \mathcal{U}_{0,i}, R) \int_{\{u \in \mathcal{U}_{0,i}\}} R_i(u | \mathcal{U}_{0,i}) d\mu(u | \mathcal{U}_{0,i})}{\mu(\mathcal{U}_{+A,i}) + \int_{\{u \in \mathcal{U}_{0,i}\}} R_i(u | \mathcal{U}_{0,i}) d\mu(u | \mathcal{U}_{0,i})} \\ &= \mathbb{E}(\widehat{u}_i | A, R). \end{aligned}$$

Similarly, it follows that $\mathbb{E}(\widehat{u}_i | B, R^c) \leq \mathbb{E}(\widehat{u}_i | B, R)$. □

Completing the proof. By Claim 2, the first best is implementable if and only if it is implementable by a cutoff recommendation signal. Recalling our definition of NSAP, the result follows.

D.2 Calculations underlying main hypotheses

The payoff from of B can take three values. In order to be able to use the same notation for both scenarios, let us call them H, M, L , going from the highest to the lowest.³⁰ For the reader's convenience, the distribution of B-values for both Scenarios is re-iterated below.

B-values (u_{iB})	H	M	L
Probabilities	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$

D.2.1 Basis for theoretical predictions (Section 2.3)

In this section, we describe the algebraic steps taken to arrive at the hypotheses in Section 2.3.

Fix any Scenario, 1 or 2. The distribution of B pins down the distribution of the $3^4 = 81$ possible cardinal preference profiles. Fix any of these preference profiles. Now let us consider the information treatments one by one.

- **Full Info:** Given any preference profile, we can calculate the allocation for each of the $4! = 24$ priority rankings of agents and their corresponding aggregate welfare, assuming that agents simply choose the option that gives them a higher private payoff. Since all rankings are equally likely under random serial dictatorship, we take their uniform average to compute the expected aggregate payoff.
- **Partial Info:** In Scenario 1, the recommendation, which is aggregate payoff maximizing, should be accepted (we show this in the next subsection, D.2.2). Therefore, for each realized preference profile, the aggregate payoff is the maximum possible aggregate payoff, across all possible allocations.

In Scenario 2, each agent reports A . Hence, we can compute the aggregate payoff for each agent ranking, and therefore the expected aggregate payoff, exactly like in *Full Info*.

- **No Info:** In Scenario 1, under our baseline assumption of risk-neutrality, each player is indifferent between A and B . We assume, for each ranking, each player is equally likely to report A and B . Under this assumption, we can calculate the expected aggregate payoff for each ranking, and then take the uniform average across all rankings to generate the expected aggregate payoff for the realized preference profile. We also repeat the above calculation assuming each player reports A and each player reports B . This does not alter our baseline hypotheses reported in Section 2.3.

³⁰So, in Scenario 1, H, M and L denote 500, 300 and 100 respectively, and in Scenario 2, they denote 400, 200 and 0 respectively.

In Scenario 2, each agent strictly prefers A . Hence for this Scenario, our calculations are identical to that under *Partial Info*.

The above calculations allow us to compute the distribution of aggregate payoffs for each Scenario and information treatment. An example is given below.

Table D1: Distribution of aggregate payoffs in Scenario 1

Probability	Aggregate payoff
0.0016	800
0.0192	1000
0.3952	1200
0.4032	1400
0.1808	1600

D.2.2 Posterior distributions of B conditional on recommendations

Let R_i denote the agent's recommendation, so $R_i \in \{A, B\}$ for all $i \in I$. Let $u_{-i,B}$ denote the set of B-values of all (three) agents other than i . Upon observing a recommendation of B under the partial information treatment, the agent updates his posterior distribution of B-values using Bayes rule as follows:

$$Pr(u_{iB}|R_i = B) = \frac{Pr(R_i = B|u_{iB})Pr(u_{iB})}{\sum_{u_{iB} \in \{H,M,L\}} Pr(R_i = B|u_{iB})Pr(u_{iB})}, \quad (\text{Bayes})$$

$$\text{where } Pr(R_i = B|u_{iB}) = \sum_{u_{-i,B}} Pr(R_i = B|u_{iB}, u_{-i,B})Pr(u_{-i,B})$$

where in the last line we use the fact that $Pr(u_{-i,B}|u_{iB}) = Pr(u_{-i,B})$, by independence across agents.

Recall that when there are multiple allocations which all maximize aggregate payoff, the planner chooses one uniformly at random, and jointly recommends it (privately) to each agent. Using this information, the distribution of B-values for the population and the corresponding probability of i being recommended B conditional on each of *his* possible B-values, is given below.

Table D2: Distribution of recommendation B

		$Pr(R_i = B \mid u_{-i,B}, u_{iB})$			$Pr(R_i = B, u_{-i,B} \mid u_{iB})$		
$u_{-i,B}$	$Pr(u_{-i,B})$	$u_{iB} = H$	$u_{iB} = M$	$u_{iB} = L$	$u_{iB} = H$	$u_{iB} = M$	$u_{iB} = L$
HML	18/125	1	1/2	0	18/125	9/125	0
HHM	9/125	2/3	0	0	6/125	0	0
HHL	3/125	2/3	0	0	2/125	0	0
MMH	27/125	1	1/3	0	27/125	9/125	0
MML	27/125	1	2/3	0	27/125	18/125	0
LLH	3/125	1	1	1/3	3/125	3/125	1/125
LLM	9/125	1	1	1/3	9/125	9/125	3/125
HHH	1/125	1/2	0	0	1/250	0	0
MMM	27/125	1	1/2	0	27/125	27/250	0
LLL	1/125	1	1	1/2	1/125	1/125	1/250

The table entries are understood as follows. Fix an agent i . The first two columns of Table D2 capture the joint distribution of the other three agents' valuations of B (The first column captures the joint valuations and the second, their probabilities). For example, the valuation set $u_{-i,B} = \{M, M, L\}$ can arise in 3 ways - with L being assigned to any of the three agents other than i . The probability of each of these three possible joint assignments occurring is $\frac{3}{5} \cdot \frac{3}{5} \cdot \frac{1}{5}$. Hence the total probability of the event $u_{-i,B} = \{M, M, L\}$ is $3 \times \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{1}{5} = \frac{27}{125}$, as noted in the second column of the fifth row.

The next three columns of the table capture the probability of i being recommended B , given a particular joint realization of others' valuations, depending on the row, and i 's own. For example, When the others' joint assignment is $\{M, M, L\}$ and i 's own is M , note that there are three possible aggregate payoff maximizing allocations, in two of which i is allocated B . Hence the entry of 2/3 in the fourth column of the same row.

The last three columns of the table simply convert the above distribution of the recommendation of B for i , conditional on the joint valuation profile of all the four agents, to the joint distribution of a recommendation of B for i and a joint valuation profile of the other three agents, conditional on each possible realization of i 's own value.

We can use (Bayes) with values from Table D2 to calculate the posterior distribution of each agent upon receiving each recommendation, as given in Table D3 below. Note that the probability of the recommendation A is just the complement of that of B, for any given realized valuation.

Table D3: Distribution of B values conditional on the recommendation

Value of B	L	M	H
Value in Scenario 1	100	300	500
Value in Scenario 2	0	200	400
Probability (on being recommended B)	$\frac{9}{625}$	$\frac{375}{625} = \frac{3}{5}$	$\frac{241}{625}$
Probability (on being recommended A)	$\frac{241}{625}$	$\frac{375}{625} = \frac{3}{5}$	$\frac{9}{625}$

Using the above, we obtain the values given in the text for the case when the agents are risk-neutral.

E Proofs for behavioral extensions

Note that by strategyproofness of random serial dictatorship, as long as subjects understand the game, under any information treatment, they report an object — A or B — if and only if its posterior expected utility is higher than that of the other one. The bounds derived in this section draw on this observation.

E.1 Risk attitudes

To summarize our arguments for Proposition 2, we show that in Scenario 1 an agent with risk aversion parameter γ accepts a recommendation of B for all $\gamma \in [-1, 1)$. Similar calculations show that he also accepts a recommendation of A in both Scenarios and rejects a recommendation of B in Scenario 2, also for all $\gamma \in [-1, 1)$. The case of the Full Info treatment is obvious. For this subsection we assume agents are perfectly Bayesian.

Proof of Proposition 2. For acceptance of a recommendation of B in Scenario 1 we need,

$$\mathbb{E}(\hat{u}_{i,(b,a)} | \text{rank}(\hat{u}_{i,(b,a)}) \leq 2) > 0, \text{ for any } \gamma \in (0, 1].$$

For any $\gamma \in [-1, 1]$, utility from B takes values $\{u(v_i)\}_{i=1}^3$, where $\{v_i\}_{i=1}^3$ are as given in Table D3 and the utility function $u(\cdot)$ is as defined in (1). Letting $f_L := \frac{9}{625}$, $f_M := \frac{3}{5}$, $f_H := \frac{241}{625}$ and $v_A := 300$, upon being recommended B , the agent accepts the recommendation if and only if:

$$\underbrace{u(L)f_L + u(M)(1 - f_L - f_H) + u(H)f_H}_{\text{Expected utility from } B \text{ when recommended } B} \geq \underbrace{u(v_A)}_{\text{Utility from } A, =u(M) \text{ in Scenario 1}} \quad (3)$$

In Scenario 1, (3) is equivalent to (using Table D3):

$$\frac{f_L}{f_H} \leq \frac{u(H) - u(M)}{u(M) - u(L)} = \frac{500^{1-\gamma} - 300^{1-\gamma}}{300^{1-\gamma} - 100^{1-\gamma}} \quad (4)$$

The right hand side of (4) is a decreasing function of γ for $\gamma \in [-1, 1]$. Therefore suffices to show that $\lim_{\gamma \uparrow 1} \frac{500^{1-\gamma} - 300^{1-\gamma}}{300^{1-\gamma} - 100^{1-\gamma}}$ exists and $\frac{f_L}{f_H} \leq \lim_{\gamma \uparrow 1} \frac{500^{1-\gamma} - 300^{1-\gamma}}{300^{1-\gamma} - 100^{1-\gamma}}$.

Using L'Hospital's rule, $\lim_{\gamma \uparrow 1} \frac{500^{1-\gamma} - 300^{1-\gamma}}{300^{1-\gamma} - 100^{1-\gamma}} = \lim_{\gamma \uparrow 1} \frac{500^{1-\gamma} \ln 500 - 300^{1-\gamma} \ln 300}{300^{1-\gamma} \ln 300 - 100^{1-\gamma} \ln 100} = \ln_3 5 - 1 \approx 0.4605$. $\frac{f_L}{f_H} = \frac{9}{241} < \ln_3 5 - 1$.

Therefore in Scenario 1, agents choose B when recommended regardless of risk parameter γ , as long as $\gamma \in (-1, 1]$.

Similarly, in Scenario 2, (3) is equivalent to:

$$f_H(u(H) - u(M)) - f_L(u(M) - u(L)) \geq u(v_A) - u(M)$$

Similarly, using values from Table D3 for Scenario 2, it is easy to verify that this is never satisfied for $\gamma \in [-1, 1)$. Hence the (Bayesian) agent in Scenario 2 never accepts a recommendation of B - similarly as in the baseline risk-neutral case - for $\gamma \in [-1, 1)$.

Similar calculations show that a recommendation of A is accepted if and only if:

$$u(L)f_H + u(M)(1 - f_L - f_H) + u(H)f_L \leq u(v_A)$$

Using values from Table D3, it can be verified that this is satisfied in both scenarios, for all $\gamma \in [-1, 1]$.

Hence the agents accept a recommendation of A in both scenarios, for any level of risk preferences within the range $\gamma \in [-1, 1]$.

E.2 Non-Bayesian updating: Prior bias

In the model of prior bias, (Prior Bias), the posterior expected utility of an agent from object B upon receiving any signal $s \in \{A, B\}$ is a convex combination of the Bayesian posterior expected utility and the prior expected utility.

Let us denote the posterior expected utility from object B of an agent i with prior bias b , upon receiving recommendation $s \in \{A, B\}$ as $\mathbb{E}(u_{iB}|s, b)$. Then, by our model (Prior Bias) we have:

$$\mathbb{E}(u_{iB}|s, b) = \mathbb{E}(u_{iB}|s, 0)(1 - b) + b\mathbb{E}(u_{iB})$$

Clearly, $\mathbb{E}(u_{iB}|s, 0)$ is the Bayesian posterior expected utility.

Hence,

$$\begin{aligned}
\text{B is chosen after recommendation } s &\iff \mathbb{E}(u_{iB}|s, b) \geq u_{iA} \\
&\iff \mathbb{E}(u_{iB}|s, 0)(1 - b) + b\mathbb{E}(u_{iB}) \geq u_{iA}
\end{aligned} \tag{5}$$

Note that in Scenario 1, for risk neutral or risk loving agents, each term in the average on the left hand side above is weakly greater than $u(v_A)$. Hence, this holds for any $b \in [0, 1]$ for risk neutral or risk loving agents.

For risk averse agents, calculating $\mathbb{E}(u_{iB}|A, 0)$ and $\mathbb{E}(u_{iB}|B, 0)$ as we did in the previous section, and using the same notation, we have, the recommendation $s = B$ is accepted in Scenario 1 if and only if:

$$\left(u(L)f_L + u(M)(1 - f_L - f_H) + u(H)f_H \right)(1 - b) + b \left(u(L)\frac{1}{5} + u(M)\frac{3}{5} + u(H)\frac{1}{5} \right) \geq u(v_A)$$

Simplifying, we have, this holds if and only if:

$$b \leq \frac{1}{1 - \frac{\frac{1}{5}((u(H) - u(M)) - (u(M) - u(L)))}{f_H(u(H) - u(M)) - f_L(u(M) - u(L))}} \tag{6}$$

Let $\bar{b}_s^i(\gamma)$ denote the maximum b as a function of risk aversion for which agents behave as predicted in the baseline case in Scenario $i \in \{1, 2\}$ when recommended $s \in \{A, B\}$.

Then, summarizing the above reasoning, we have:

$$\bar{b}_B^1(\gamma) = \begin{cases} 1, \gamma \leq 0 \\ \frac{1}{1 - \frac{\frac{1}{5}((u(H) - u(M)) - (u(M) - u(L)))}{f_H(u(H) - u(M)) - f_L(u(M) - u(L))}}, \gamma \in (0, 1]. \end{cases}$$

Similar calculations for recommendation A in Scenario 1 lead to the following thresholds:

$$\bar{b}_A^1(\gamma) = \begin{cases} \frac{1}{1 + \frac{\frac{1}{5}((u(H) - u(M)) - (u(M) - u(L)))}{f_H(u(M) - u(L)) - f_L(u(H) - u(M))}}, \gamma \in [-1, 0]. \\ 1, \gamma \geq 0 \end{cases}$$

In Scenario 2, when B is recommended, note that in (5), both expressions on the left hand side - the expected utility from B with and without a recommendation of B — are lower than the left hand side — the expected utility from A, for all $\gamma \in [-1, 1]$, as shown in the previous subsection. Hence (5) never holds in Scenario 2 for $s = B$. Hence $\bar{b}_B^2(\gamma) = 1$ for all $\gamma \in [-1, 1]$. If B is not accepted when B is recommended, it is not accepted when A is recommended, so $\bar{b}_A^2(\gamma) = 1$ for all $\gamma \in [-1, 1]$ too.

The common threshold for this type of mistake probabilities, for any risk aversion $\gamma \in [-1, 1]$, is then $\bar{b}(\gamma) := \min\{\bar{b}_A^1(\gamma), \bar{b}_B^1(\gamma)\}$, which is plotted in Figure C1 in the main text. As we see, $b < 0.6$ is sufficient to uphold our baseline predictions.

E.3 Social preferences

Proof of Proposition 5. Each agent’s reasoning of whether to change his report due to altruistic considerations, is as follows: Depending on agent i ’s position in the ranking of serial dictatorship and other agents’ reports, agent i ’s report may or may not affect his allocation. In particular, agents ranked 1 and 2 always get the object they ask for, the report of the agent ranked 3 affects his allocation (which is equivalent to him getting what he asked for) if and only if those ranked 1 and 2 made different choices, and the report of the agent ranked 4 never affects his allocation. Based on the logic of serial dictatorship, an agent knows that when his report affects his allocation, changing it from $x \in \{A, B\}$ to $y \neq h$ is equivalent to trading h for y with *some* other agent, leaving the other two agents’ allocations unchanged. His individual loss from changing his report from h to y due to altruistic motivations is the absolute difference in his a posteriori expected utilities from A and B , calculated per his posterior. This, as well as his assessment of the gain to others varies depending on the Scenario and how he updates his beliefs, as detailed below. As mentioned in the main text, we only consider the two extreme types of updating of beliefs — fully Bayesian and fully prior-biased.

- Bayesian, Scenario 1: Each agent knows that the recommendation is intended to maximize aggregate social welfare. As we have shown, even a purely self-interested Bayes rational agent ($a = 0$) finds it optimal to accept each recommendation - A or B - in Scenario 1. A positive a pushes any agent to make choices more aligned to the social good. Therefore in Scenario 1, the Bayesian agent’s choices for $a = 0$ and $a > 0$ are the same.
- Fully prior-biased, Scenario 1 and 2: Such agents ignore the recommendation. Therefore, their own expected utilities from choosing A and B are their prior expected utilities. By ignoring the recommendation, these are also the expected utilities of any potential beneficiary of the fully prior-biased agent’s sacrifice if he chooses to change his report. Therefore, from the fully prior-biased agent’s point of view, the expected gain of another agent from his sacrifice is equal to his loss from the same sacrifice. Therefore, such a trade-off makes sense to him only if $a \geq 1$, which we rule out as unrealistic.

Calculations for Figure C3 (Bayesians in Scenario 2). As we have shown, in Scenario 2 it is optimal for a self-interested Bayes rational agent to ignore the recommendation and always choose A . An altruistically motivated Bayes rational agent may, therefore, find it optimal to change his report to B when recommended B for a suitably high value of $a > 0$. When an agent is recommended A , he knows this is the best choice both for him individually, and collectively, therefore his decision to report A when recommended A remains unchanged.

We claim that when he is recommended B , he knows if he picks B and his report is relevant, someone who has been recommended A will get A .

The reasoning is as follows. The agent's report can be relevant if and only if either (i) he is ranked in the top two or (ii) he is ranked third and one of the agents ranked ahead of him has reported B and the other has reported A . In the first case (i), the agent knows that if he has been recommended B , at least one of the agents ranked in the last two has been recommended A . When he chooses to report B , he trades A with one of these last two ranked agents for B , i.e., someone who is recommended A gets it.

Relevant for case (ii), recall that per our assumption, a Bayes rational and altruistic agent assumes others are Bayes rational *but not altruistic*. Therefore, each agent knows that if he is ranked third, his report is not relevant because those ranked ahead of him have both picked A . This is because, recall, in Scenario 2, no purely self-interested agent will report B , regardless of the recommendation.³¹

Of course, the agent does not know his rank. But since he trades A for B with someone who is recommended A in each of the possible cases, he knows this to be the only possibility even without knowing his rank.

The change in the expected utility of an agent who has been recommended B from swapping out B for A can be calculated using values from Table D3. Therefore, a for the altruistic agent must be high enough so that his individual loss is compensated by the gain of the beneficiary of his sacrifice. Therefore, in Scenario 2 for an agent to accept a Recommendation of B we must have,

$$\begin{aligned} & (\mathbb{E}(u_{iB}|\text{recommendation } B) - u_{iA}) + a(u_{iA} - \mathbb{E}(u_{iB}|\text{recommendation } A)) \geq 0 \\ \iff a \geq & \frac{300^{1-\gamma} - \left(\left(\frac{3}{5}\right) 200^{1-\gamma} + \left(\frac{241}{625}\right) 400^{1-\gamma}\right)}{300^{1-\gamma} - \left(\left(\frac{3}{5}\right) 200^{1-\gamma} + \left(\frac{9}{625}\right) 400^{1-\gamma}\right)} =: \underline{a}(\gamma) \end{aligned}$$

The last line comes from using values from Table D3.

F Simulations for policy analysis

A careful reader might point out that our recommendations rely on the unrealistic assumption of risk neutral, self-interested, and perfectly Bayesian agents, and are thus severely misaligned with the agents' actual preferences. Let us address this concern by now supposing that the social planner could obtain additional (truthful)

³¹It is worth noting that the prediction of reporting B when recommended B for altruistic agents in Scenario 2 holds even if we assume that the agent assumes others to be Bayes rational and altruistic. In this case, if he is ranked third, he knows his report can be relevant only if at least one of the top two ranked agents reported B , which – per his assumption of altruism – means that agent was recommended B . Therefore, he knows that the last ranked agent has been recommended A . Therefore, in this case also he trades A for B with someone who is recommended A .

information about the subjects' preferences *prior* to generating her recommendation, and could thus incorporate it in her calculations.³² If this were the case, by how much would the recommendations change, and what would be the resulting effect on social welfare?

Due to the complexity of the problem, we focus entirely on social preferences. We consider this a realistic assumption, since administrative data can provide a lot of information on this matter, be it in the form of charitable giving, volunteering work, care for elderly or disabled relatives, etc. In our setting, we assume that the social planner would use social preferences of agents to break ties, i.e., choose between otherwise interchangeable allocations. Specifically, all else equal, in situations where multiple agents have the same valuations and some of them have to accept a lower-valued object, the planner assigns this lower-valued object to the altruist, reasoning that of all the agents, he is made the *least unhappy* with this allocation. Given that we do not have robust evidence that altruists behave differently in our experiment, we do not adjust the subjects' behavior in response to this change in the recommendation generation rule.

Under these assumptions, using our experimental data on the subjects' social preferences, we find that 6 recommendations across 3 groups (1.15%) would change in Scenario 1 and 160 recommendations across 80 groups (30.5%) would change in Scenario 2. Since we do not know whether a given subject would follow their new recommendation, we compare group outcomes for two benchmarks: First, we assume that all subjects who previously followed the recommendation continue doing so, while those who did not follow it stick with their original choice. And second, we calculate the hypothetical scenario if all subjects followed the original versus the new recommendation.

³²We thank John A. List for this suggestion.

Table F1: Welfare changes under alternative recommendations

	Original decisions	Always follow
Scenario 1		
groups with different welfare	0 (0%)	0 (0%)
individuals with different welfare	4 (0.38%)	6 (0.57%)
increased welfare non-altruists	2 (0.19%)	3 (0.29%)
increased welfare altruists	0 (0%)	0 (0%)
decreased welfare non-altruists	0 (0%)	0 (0%)
decreased welfare altruists	2 (0.19%)	3 (0.29%)
Scenario 2		
groups with different welfare	16 (1.53%)	0 (0%)
individuals with different welfare	90 (8.59%)	160 (15.27%)
increased welfare non-altruists	39 (3.72%)	76 (7.25%)
increased welfare altruists	6 (0.57%)	4 (0.38%)
decreased welfare non-altruists	5 (0.48%)	11 (1.05%)
decreased welfare altruists	40 (3.72%)	69 (7.25%)

Note: Table reports the number (%) of groups and individuals for which welfare changes under new recommendations. First column reports changes under original decisions (follow if originally followed, else original decision); second column compares welfare under the new vs original recommendations if these were always followed.

As shown in Table F1, it is rare for the *aggregate* welfare to increase (as this only happens by chance, and only in a small set of cases). Of course, here we do not quantify the potential welfare gains for pro-social agents who might appreciate the new design feature of recommendations: such gains would be proportional to the share of altruists, and could thus be estimated depending on the context one is interested in.

However, the augmented recommendations clearly change the welfare *within* groups, as the welfare of the pro-social agents is in most cases sacrificed for the non-pro-social agents in the group. This main result is robust regardless what we assume regarding the new recommendation following, albeit in general the share of affected agents is small even in the (hypothetical) case where everybody follows their recommendation.

This result also reveals a practical obstacle to using the additional information on social preferences. Unless subjects exhibit an (unrealistically) high level of altruism such that their loss could be compensated by an anonymous other’s gain, the agents are incentivized to hide their pro-sociality, which may have undesirable spillovers in other decision situations.

As a whole, we thus caution against the use of additional data on subjects’

preferences to generate these recommendations, at least in situations with possible behavioral spillovers in other domains, or in repeated interactions.

G Supplementary tables

G.1 Tables complementing main analysis

Here we show that our treatments are balanced on observables. The only control variable that is significantly different across treatments on a 5% level is completion time, which is not particularly concerning given that the longer treatments do not result in greater problems with attention (which is a common concern in long experiments).

Table G1: Balance table

	Full Info	Partial Info	No Info	p-value
Completion time (min)	13.26	14.20	14.14	0.023
Risk aversion	58.14	59.22	60.70	0.141
Experimenter demand	11.73	11.65	13.52	0.224
Bayesian deviation	13.00	13.50	14.82	0.165
% female	42.28	43.12	39.61	0.603
% student status	6.13	9.10	7.06	0.053
% instructions check	10.41	8.13	10.20	0.219
% instructions failure	60.57	65.76	63.92	0.068
% attention failure	47.75	49.73	50.59	0.601
% altruist	28.06	29.90	28.78	0.679

Note: Variables follow the same definitions as in Table 1.

In the main text, we report a simple logit to explore mechanisms driving subjects' choices in the *Partial Info* treatment. Here, we show the results are similar in a sequential logit model: The key advantage of modelling these choices as sequential is that we take into account that while the parameters of the second choice are unknown to the subjects at the moment of making the first choice, the subjects' second choice may depend on what they chose in the first Scenario (e.g., because they concluded that it was initially in their interest to follow the recommendation, and are now relying on it again as a heuristic instead of checking whether it is still in their interest (de Haan and Linde, 2018)).

Table G2: Optimal Choices across Scenarios: Partial Info

	Scenario 1		Scenario 2			
			(following optimal choice)		(otherwise)	
Recommended A	1.795*** (0.151)	1.697*** (0.159)	31.075*** (0.347)	33.454*** (0.362)	3.131* (0.460)	2.909* (0.465)
Risk averse		1.005 (0.006)	1.021* (0.010)	1.020 (0.010)	1.007 (0.014)	1.005 (0.014)
Non Bayesian		1.025 (0.016)	1.067* (0.033)	1.067 (0.034)	0.975 (0.058)	0.969 (0.058)
Risk averse \times non-Bayesian		1.000 (0.0003)	0.999 (0.001)	0.999 (0.001)	1.000 (0.001)	1.000 (0.001)
Inattention		0.949 (0.159)	0.941 (0.228)	0.987 (0.233)	0.680 (0.422)	0.692 (0.405)
# attempts comprehension Qs		1.328 (0.163)	0.542* (0.256)		0.632 (0.466)	
Altruist		1.013 (0.173)	0.729 (0.243)	0.718 (0.245)	0.406* (0.440)	0.423 (0.481)
Experimenter demand		1.006 (0.005)	1.007 (0.007)	1.006 (0.005)	0.978 (0.017)	0.979 (0.005)
Failed comprehension Q1				1.129 (0.190)		0.840 (0.258)
Failed comprehension Q2				0.846 (0.222)		1.355 (0.525)
Failed comprehension Q3				0.650** (0.162)		0.701 (0.342)
Failed comprehension Q4				0.642* (0.226)		0.698 (0.513)
N (chose A)		482		501		195
N (chose B)		387		144		29
N (chose optimally)		645		501		195
N (total)		869		869		869

Note: The table shows the odds ratios from a sequential logit for making the theoretically optimal choices in the Partial Info treatment. Optimal strategies follow the recommendation in the first Scenario, and select A in the second Scenario. Only subjects who completed all ancillary tasks are included.

Q1 = *suppose you would have chosen the other task, how much would you have earned?*, Q2 = *suppose two players ahead of you would have chosen task A too, would you have been allocated A?*, Q3 = *could player 4 improve their payoff?*, Q4 = *I will know the exact payoff of B (correct answer is treatment-dependent)*

Robust standard errors are in parentheses. Clustering on individual level.

* p-val < 0.05, ** p-val < 0.01, *** p-val < 0.001

G.2 Robustness: Group outcomes excluding groups with bots

In this section we re-estimate our main treatment effects for group-level outcomes for only those groups where all participants completed both Scenarios of the main game and thus were not replaced by computerized bots. Notice we are not providing this robustness check for Table 4 because it already excludes bots in the main analysis.

Table G3: Treatment Effects on Social Welfare (Robustness)

	All comparisons (Jonckheere-Terpstra)	Full vs. Partial (Mann-Whitney-U)	Partial vs. No (Mann-Whitney-U)	Full vs. No (Mann-Whitney-U)
H1: Partial > Full > No	0.040*	0.499 [0.091]	0.001*** [0.002]	0.000*** [0.002]
H2: Full > Partial = No	0.000***	0.029* [0.023]	0.059 [0.037]	0.002** [0.003]
N (groups)	395	348	229	213

Note: The first column lists p-values from the Jonckheere-Terpstra trend test for the ordered aggregate social welfare levels, and columns 2-4 list p-values for two-sided pairwise comparisons using the Mann-Whitney-U test.

* p-val < 0.05, ** p-val < 0.01, *** p-val < 0.001

Sharpened false discovery rate q-values for the six pairwise tests (Anderson, 2008) are in brackets.

Table G4: Reaching First Best (Robustness)

	Scenario 1	Scenario 2
H1: $W^* = \text{Partial}$	0.000*** [0.001]	H2: $W^* > \text{Full}$ 0.000*** [0.001]
N (groups)	182	166

Note: The table lists the p-values for the one-sample two-sided Wilcoxon sign rank test, comparing the treatment that was hypothesized to equal (Scenario 1) or fail to reach (Scenario 2) the aggregate social welfare optimum to the theoretical first best.

* p-val < 0.05, ** p-val < 0.01, *** p-val < 0.001

Sharpened false discovery rate q-values for the two tests (Anderson, 2008) are in brackets.

H Data note

Some participants used several attempts to complete the experiment, for example due to technical errors. Since we stored Prolific IDs, we were able to flag these as duplicates. We decided to keep the first attempt if both Scenario decisions were made. If a person participated multiple times but always faced the same treatment, we keep the first completed attempt (e.g., second or third attempt). In case a person got past the first scenario decision but has several incomplete attempts, all attempts were deleted. The full preparation file dealing with duplicates can be found https://www.jantsje.nl/files/preparation_matching.R

References

- K. Abbink and H. Hennig-Schmidt. Neutral versus loaded instructions in a bribery experiment. *Experimental Economics*, 9(2):103–121, 2006.
- A. Abdulkadiroğlu and T. Sönmez. Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica*, 66(3):689–701, 1998.
- S. Andersen, G. W. Harrison, M. I. Lau, and E. E. Rutström. Eliciting risk and time preferences. *Econometrica*, 76(3):583–618, 2008.
- L. R. Anderson and J. M. Mellor. Predicting health behaviors with an experimental measure of risk preference. *Journal of Health Economics*, 27(5):1260–1274, 2008.
- M. L. Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484):1481–1495, 2008.
- A. Aristidou, G. Coricelli, and A. Vostroknutov. Incentives or persuasion? an experimental investigation. *Working paper*, 2019.
- D. Arroyos-Calvera, J. Lohse, and R. McDonald. Beyond social influence: Examining the efficacy of non-social recommendations. *Working paper*, 2023.
- G. Artemov. Assignment mechanisms: common preferences and information acquisition. *Journal of Economic Theory*, 198:105370, 2021.
- P. H. Au and K. K. Li. Bayesian persuasion and reciprocity: theory and experiment. *SSRN working paper*, 2018.
- E. M. Azevedo and J. D. Leshno. A supply and demand framework for two-sided matching markets. *Journal of Political Economy*, 124(5):1235–1268, 2016.
- K. Barron. Belief updating: does the ‘good-news, bad-news’ asymmetry extend to purely financial domains? *Experimental Economics*, 24(1):31–58, 2021.
- D. Bergemann and S. Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016.
- A. Bogomolnaia and H. Moulin. A new solution to the random assignment problem. *Journal of Economic Theory*, 100(2):295–328, 2001.
- M. Bombardini and F. Trebbi. Risk aversion and expected utility theory: an experiment with large and small stakes. *Journal of the European Economic Association*, 10(6):1348–1399, 2012.

- T. Börgers and J. Li. Strategically simple mechanisms. *Econometrica*, 87(6):2003–2035, 2019.
- G. L. Brase and W. T. Hill. Adding up to good Bayesian reasoning: Problem format manipulations and individual skill differences. *Journal of Experimental Psychology: General*, 146(4):577, 2017.
- S. Braun, N. Dwenger, D. Kübler, and A. Westkamp. Implementing quotas in university admissions: An experimental analysis. *Games and Economic Behavior*, 85:232–251, 2014.
- M. Buis. Seqlogit: Stata module to fit a sequential logit model. 2013.
- C. Calsamiglia, G. Haeringer, and F. Klijn. Constrained school choice: An experimental study. *American Economic Review*, 100(4):1860–74, 2010.
- M. Castillo and A. Dianat. Truncation strategies in two-sided matching markets: Theory and experiment. *Games and Economic Behavior*, 98:180–196, 2016.
- G. Charness, T. Garcia, T. Offerman, and M. C. Villeval. Do measures of risk attitude in the laboratory predict behavior under risk in and outside of the laboratory? *Journal of Risk and Uncertainty*, 60:99–123, 2020.
- Y. Chen and Y. He. Information acquisition and provision in school choice: an experimental study. *Journal of Economic Theory*, 197:105345, 2021.
- Y. Chen and T. Sönmez. School choice: an experimental study. *Journal of Economic Theory*, 127(1):202–231, 2006.
- F. Cochard, J. Le Gallo, N. Georgantzis, and J.-C. Tisserand. Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game. *Journal of Behavioral and Experimental Economics*, 90:101613, 2021.
- A. Coutts. Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395, 2019.
- P. Crosetto and A. Filippin. The “bomb” risk elicitation task. *Journal of Risk and Uncertainty*, 47(1):31–65, 2013.
- S. Dasgupta. Optimal information structures in matching markets. *Working paper*, 2020.
- G. de Clippel and X. Zhang. Non-Bayesian persuasion. *Journal of Political Economy*, 130(10):2594–2642, 2022.
- T. de Haan and J. Linde. ‘good nudge lullaby’: Choice architecture and default bias reinforcement. *The Economic Journal*, 128(610):1180–1206, 2018.

- R. Deb, M. M. Pai, and M. Said. *Indirect Persuasion*. Centre for Economic Policy Research, 2023.
- T. Ding and A. Schotter. Learning and mechanism design: An experimental test of school matching mechanisms with intergenerational advice. *The Economic Journal*, 129(623):2779–2804, 2019.
- F. Echenique, A. J. Wilson, and L. Yariv. Clearinghouses for two-sided matching: An experimental study. *Quantitative Economics*, 7(2):449–482, 2016.
- W. Edwards. Conservatism in human information processing. *Formal Representation of Human Judgment*, 1968.
- C. Engel and D. G. Rand. What does “clean” really mean? the implicit framing of decontextualized experiments. *Economics Letters*, 122(3):386–389, 2014.
- L. G. Epstein. An axiomatic model of non-Bayesian updating. *The Review of Economic Studies*, 73(2):413–436, 2006.
- F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160, 2009.
- G. R. Fréchet, A. Lizzeri, and J. Perego. Rules and commitment in communication: An experimental analysis. *Econometrica*, 90(5):2283–2318, 2022.
- H. Fromell, D. Nosenzo, and T. Owens. Altruism, fast and slow? evidence from a meta-analysis and a new experiment. *Experimental Economics*, 23(4):979–1001, 2020.
- D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.
- E. S. Geller and G. F. Pitz. Confidence and decision speed in the revision of opinion. *Organizational Behavior and Human Performance*, 3(2):190–201, 1968.
- P. Guillen and R. Hakimov. The effectiveness of top-down advice in strategy-proof mechanisms: A field experiment. *European Economic Review*, 101:505–511, 2018.
- P. Guillen and A. Hing. Lying through their teeth: Third party advice and truth telling in a strategy proof mechanism. *European Economic Review*, 70:178–185, 2014.
- R. Hakimov and D. Kübler. Experiments on centralized school choice and college admissions: a survey. *Experimental Economics*, 24(2):434–488, 2021.

- G. W. Harrison and E. E. Rutström. Risk aversion in the laboratory. In *Risk aversion in experiments*. Emerald Group Publishing Limited, 2008.
- G. W. Harrison and J. T. Swarthout. Belief distributions, Bayes rule and Bayesian overconfidence. *Working Paper*, 2022.
- Y. He, A. Miralles, M. Pycia, and J. Yan. A pseudo-market approach to allocation with priorities. *American Economic Journal: Microeconomics*, 10(3):272–314, 2018.
- C. A. Holt and S. K. Laury. Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655, 2002.
- C. A. Holt and A. M. Smith. An update on Bayesian updating. *Journal of Economic Behavior & Organization*, 69(2):125–134, 2009.
- A. Hylland and R. Zeckhauser. The efficient allocation of individuals to positions. *Journal of Political Economy*, 87(2):293–314, 1979.
- N. Immorlica, J. Leshno, I. Lo, and B. Lucier. Information acquisition in matching markets: The role of price discovery. *SSRN working paper*, 2020.
- J. H. Kagel, R. M. Harstad, and D. Levin. Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica: Journal of the Econometric Society*, pages 1275–1304, 1987.
- E. Kamenica and M. Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.
- F. Klijn, J. Pais, and M. Vorsatz. Preference intensities and risk aversion in school choice: A laboratory experiment. *Experimental Economics*, 16(1):1–22, 2013.
- K. Koutout, A. Dustan, M. Van der Linden, and M. Wooders. Mechanism performance under strategy advice and sub-optimal play: A school choice experiment. *Journal of Behavioral and Experimental Economics*, 94:101755, 2021.
- O. Kwon. Bayesian persuasion in the lab. *Ohio State University Working Paper*, 2020.
- Y.-J. Lee, W. Lim, and C. Zhao. Cheap talk with prior-biased inferences. *Games and Economic Behavior*, 2023.
- J. Li and P. Dworzak. Are simple mechanisms optimal when agents are unsophisticated? In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 685–686, 2021.

- S. Li. Obviously strategy-proof mechanisms. *American Economic Review*, 107(11): 3257–3287, 2017.
- J. Ludwig and A. Achtziger. Cognitive misers on the web: An online-experiment of incentives, cheating, and cognitive reflection. *Journal of Behavioral and Experimental Economics*, 94:101731, 2021.
- T. Masuda, R. Mikami, T. Sakai, S. Serizawa, and T. Wakayama. The net effect of advice on strategy-proof mechanisms: an experiment for the vickrey auction. *Experimental Economics*, 25(3):902–941, 2022.
- B. A. Mellers, J. D. Baker, E. Chen, D. R. Mandel, and P. E. Tetlock. How generalizable is good judgment? a multi-task, multi-benchmark study. *Judgment & Decision Making*, 12(4), 2017.
- C. H. Mullin and D. H. Reiley. Recombinant estimation for normal-form games, with applications to auctions and bargaining. *Games and Economic Behavior*, 54(1):159–182, 2006.
- C. Neilson, C. Allende, and F. Gallego. Approximating the equilibrium effects of informed school choice. *Working paper*, 2019a.
- C. Neilson, C. Allende, F. Gallego, et al. Approximating the equilibrium effects of informed school choice. Technical report, 2019b.
- Q. Nguyen. Bayesian persuasion: evidence from the laboratory. *Work. Pap., Utah State Univ., Logan*, 2017.
- L. J. Paas and M. Morren. Please do not answer if you are reading this: Respondent attention in online panels. *Marketing Letters*, 29(1):13–21, 2018.
- J. Pais and Á. Pintér. School choice and information: An experimental study on matching mechanisms. *Games and Economic Behavior*, 64(1):303–328, 2008.
- G. F. Pitz. An inertia effect (resistance to change) in the revision of opinion. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 23(1):24, 1969.
- G. F. Pitz, L. Downing, and H. Reinhold. Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 21(5):381, 1967.
- A. Rees-Jones, R. Shorrer, and C. J. Tergiman. Correlation neglect in student-to-school matching. Working Paper 26734, National Bureau of Economic Research, February 2020. URL <http://www.nber.org/papers/w26734>.
- P. P. Wakker. Explaining the characteristics of the power (crra) utility family. *Health Economics*, 17(12):1329–1344, 2008.

- M. Zhu. Experience transmission: Truth-telling adoption in matching. *SSRN Working Paper 2631442*, 2015.
- A. Ziegler. Persuading an audience: Testing information design in the laboratory. *Working Paper*, 2023.
- D. J. Zizzo. Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1):75–98, 2010.