
Incentives in Group Decision-Making With Uncertainty and Subjective Beliefs*

Ruggiero Cavallo
Yahoo! Research
111 West 40th Street
New York, NY 10018
cavallo@yahoo-inc.com

Abstract

We address the problem of decision-making in group settings where there is uncertainty and disagreement about the utility that actions will yield. Each individual brings his own private information and subjective beliefs, and a decision-maker aims to arrive at a choice that sensibly aggregates all relevant information to maximize expected social welfare. Agents and the decision-maker revise beliefs based on those held by others; we adopt a weighted averaging model, where the weight one agent assigns to another can be thought of as the agent’s “trust” in the other’s subjective beliefs. If revised beliefs may yield conflicting opinions about which action is optimal, there is a problem of incentives. We provide a payment mechanism that yields implementation of the decision-maker’s desired choice in an ex post equilibrium for arbitrary trust levels. In other words, we solve the disensus problem that arises when agents (and perhaps the decision-maker) disagree about how to weigh each others’ information in aggregating beliefs.

1 Introduction

A group of agents face the task of jointly choosing among a set of boxes, each containing some unknown amount of money. Only one can be chosen, and the money within is to be shared evenly. The agents’ goal, then, is to try to choose the box containing the greatest amount of money. But a problem arises if the agents have varying intrinsic beliefs about the contents of each box, and varying levels of regard for the beliefs of others in the group. How can they coordinate and form consensus?

The key elements of this toy example are shared by a number of more serious dilemmas: managers choosing a project to invest resources into, a chain of restaurants choosing a menu to set for the coming year, directors of the front office of a major league baseball team picking a relief pitcher to trade for in the off-season, etc. The central challenge of this kind of problem is reconciling agent beliefs that may not have any shared foundation. A typical approach is to simply stipulate a shared foundation—in particular, to assume that all individuals are perfect Bayesian decision-makers. If all agents were Bayesian, began with common prior beliefs, and obtained a communicable set of evidence that could be methodically combined with others’, the task would be a simple matter of Bayesian updating. Agents would volunteer to share their private information because they would agree on the most advantageous aggregating approach.

But in the real world this is almost always implausible, as agents may not agree on prior probabilities, conditional probabilities, the set of evidence, etc. Or, more fundamentally, agents may not believe that Bayesian reasoning is the right way of dealing with uncertainty. So here we will take this fact seriously, and consider agent beliefs that are inherently *subjective*. One agent may have knowledge that another does not, so each agent’s opinion may have value to the others in considering the options, but the knowledge cannot be combined to derive an objectively “optimal” decision.

We will model agents as having pairwise “trust” relationships that determine how intrinsic (a priori) beliefs translate to updated beliefs in light of information held by other agents. Specifically, we will define each agent’s (updated) expectation about the value of each alternative to be a trust-weighted sum of his own intrinsic expectation and the intrinsic expectations of the rest of the group. We will (unfortunately) need to assume that agent trust levels are publicly known. We will provide a payment mechanism that achieves truthful revelation of private information (beliefs) in an ex

*Paper date: March 18, 2011. Presented at the *Conference on Uncertainty and Artificial Intelligence (UAI-11)*, Barcelona, Spain, July 17, 2011.

post Nash equilibrium, which then allows for a choice that is optimal according to the decision maker’s aggregation method. Given classic mechanism design results, the problem is not particularly difficult if agents are either completely untrusting (other agents’ beliefs are irrelevant to them) or have “consensus trust” (all agents and the decision maker trust all agents the same). The place where particular care is needed, and where our main contribution lies, is where agents care about others’ beliefs, but to differing degrees. In that case it is not simply a team game—agents may have conflicting preferences about which alternative should be selected—and at the same time the beliefs an agent reports will impact the preferences of others.

The approach we suggest has myriad potential applications, as disagreement in the face of uncertainty is ubiquitous in group decision-making. To give just one example, consider the (often ignored) problem of decision making when an efficient choice function cannot be implemented due to computational intractability. When a Bayesian-optimal policy is not available we are left with a choice among “heuristics”, and agents may have subjective beliefs about the efficacy of each heuristic. Prior work has not addressed the possibility of mechanism design in such settings.

1.1 Background and related work

Trust and information sharing is central to human discovery, deliberation, and choice. The philosopher John Hardwig observes that “trust is often epistemologically even more basic than empirical data or logical arguments: the data and the arguments are only available through trust” [Hardwig, 1991]. Beyond the obvious cases of social and business relations, even in the “objective” domains of mathematics and the natural sciences, trust essentially always underlies new discoveries either through direct teamwork (e.g., co-authored research papers) or work that extends aspects of previous discovery. Taking the centrality of trust in knowledge formation as our starting point, in the current paper we examine its implications in a group decision making environment with self-interest and a priori uncertainty about the results of any chosen action.

There is a rich and growing literature on group belief revision/aggregation in uncertain settings (see [Genest and Zidek, 1986] for an excellent early review). The first question one might ask is why the problem isn’t merely a matter of Bayesian updating. Why don’t agents simply revise beliefs (prior probability distributions) according to Bayes rule in light of new evidence possessed by others? As we touched on above, there are several fundamental practical implausibilities about this idea. First, it may be the case that agents are unable to communicate the evidence underlying their beliefs (e.g., because the beliefs are primi-

tives, or memory is limited), but rather can only communicate the beliefs themselves. With the underlying evidence—and, also importantly, the correlations or overlaps between multiple agents’ evidence sets—unknown, strict-Bayesianism is impossible. Also, in many cases agents are likely to form beliefs based on different initial priors (we return to this topic shortly). Moreover, for “unrepeatable” events the application of Bayes rule to combine various subjective beliefs is controversial even in its definition; one common interpretation is to consider one agent’s beliefs as “evidence” to be used in the Bayesian update of another’s prior beliefs,¹ but why should there be consensus about the prior probability of this kind of “evidence”?

Putting the idealized Bayesian view aside, the most prevalent model of belief aggregation in the literature is that of weighted averaging (also called “proportional averaging” or the “proportional weight view”), in which an agent replaces his initial beliefs with a weighted sum of *all* agents’ beliefs, where weights may correspond to “expertise” or some other measure of reliability. This model has both significant theoretical and empirical justifications. On the theoretical side, Lehrer and Wagner [1981] show that any aggregation method that satisfies two intuitive axioms² must be a weighted averaging method. They further justify the method by showing that an iterative form of the update process achieves convergence to a consensus set of beliefs. See [Lehrer, 1983] for a helpful exposition of these ideas. [Elga, 2007] provides a philosophical defense of the “equal weight view,” a special case of weighted averaging (see also [Fitelson and Jehle, 2009]). Numerous studies have found averaging methods to yield superior predictive performance to that of any individual (see, e.g., [Ariely *et al.*, 2000]).

From a practical perspective, even more relevant is the question of how agents actually behave. Here too the weighted averaging method is supported. [Budescu *et al.*, 2003] finds weighted averaging to fit behavior in an environment where there is asymmetric information and a history of discrepant quality amongst other agents (“advisors” to a decision-maker). Particularly relevant is [Budescu and Yu, 2006], in which the weighted averaging method is experimentally compared with a “naive Bayes” method where agents make Bayesian updates, applying strong independence assumptions. The averaging model was found to generally be a closer fit to the observed behavior. In an-

¹See [Wallsten *et al.*, 1999] for discussion of this issue and also an excellent review of the literature overall.

²The axioms are: 1) “zero unanimity”: if all agents assign an event probability 0, then the aggregated belief does as well; and 2) “irrelevance of alternatives”: the aggregated probability of a statement is constant if the component belief probabilities are constant.

other study, [Yaniv, 2004] adopts the weighted averaging perspective and infers the weights agents assign to others’ advice in a question/answer setting.

In the setting we address in the current paper, a decision is to be made that will impact each of the individuals whose beliefs we wish to aggregate. So another important aspect of the problem is the interplay of beliefs and *utility*. The standard approach to such settings when agents are ignorant of the types (private information) of other players is Bayesian, and was established by Harsanyi [1967–68]. In this approach an agent’s type is taken to include beliefs in the form of a probability distribution over other agents’ types; and importantly, as mentioned above, agents are presumed to “start” (prior to obtaining any information) with common prior beliefs.³ The common prior assumption is very controversial since, for one thing, it leads to the conclusion that rational disagreement is solely the result of asymmetric information. [Morris, 1995] provides a detailed discussion of the assumption and its justifications, such as they are; [Bonanno and Nehring, 1999] provides insight by redescribing the assumption in terms of conditions that must hold when beliefs are taken as primitives; [Dekel and Gul, 1997] also explores its justifications and limitations.

To predict agent behavior in strategic settings so modeled, we look to equilibria. Specifically, a *Bayes-Nash equilibrium* is defined as a profile of strategies such that, given the profile of types, no agent’s expected utility can be improved by unilateral deviation. But this equilibrium notion is fragile because of the strength of the assumptions it must make: the common prior, plus the requirement that all agents believe that other agents hold the specified prior and will play according to the equilibrium strategy profile.

In this paper we will be able to move away from the assumptions that define the Bayes-Nash solution concept by adopting a distinct model—supported by the behavioral literature cited above—of how agents form expectations. In our model agent types specify *intrinsic subjective beliefs* about the results actions will bear,⁴ and expectations are formed by combining these

³Harsanyi also considers the case of subjective priors, but notes that for their to be a Bayesian game representation beliefs must be “consistently describable” as the posterior of some common prior distribution.

⁴While our model deals explicitly with expectations, allowing us to sidestep analysis of any structure that may underlie them, there is a consonance with the subjectivist view of probability advocated, e.g., in the foundational works of [Ramsey, 1931] and [de Finetti, 1964]. Ramsey takes pride in the fact that his theory can accommodate “a probable belief founded not on argument but on direct inspection”. With respect to the priors central to a Bayesian approach, Ramsey says “what are the absolutely

intrinsic beliefs with those of other agents. Essentially, we adopt a personalized proportional weight view of belief revision, where each agent assigns weights according to his “trust” of each agent (including himself),⁵ and the decision-maker also assigns his own distinct trust weights to each agent in determining how to reach a choice. The paper is proscriptive rather than merely descriptive because, when trust levels are too asymmetric or small to achieve consensus (i.e., to bring truthful revelation into equilibrium), to solve that problem we must rebalance the incentives. We are able to accomplish this with a payment mechanism, and this is the heart of our contribution.

In our solution truthful revelation of beliefs will be achieved in an ex post equilibrium, which is crucial since one agent’s beliefs form the partial basis for another’s expectation. So, unlike in the case of Bayes-Nash, the equilibrium we achieve stipulates nothing about the nature or genesis of the probability distributions that may underlie agents’ intrinsic beliefs about the world, but rather only that they will communicate such beliefs truthfully: no agent can benefit from lying as long as others are truthful, regardless of their types. The concept of *trust* thus comes in not with respect to agents’ truthfulness, but rather with respect to the perceived reliability (i.e., predictive power) of their beliefs. One agent’s trust in another is manifest in how expectations are updated in light of the other’s intrinsic beliefs.⁶

2 Setup

There is a set of agents $I = \{1, \dots, n\}$, a set of actions A , a set of outcomes O , and joint typespace $\Theta = \Theta_1 \times \dots \times \Theta_n$. There is a decision-maker who will choose an action after eliciting a type report from each agent. The environment is one of uncertainty: an action $a \in A$ is chosen, and subsequently nature realizes an outcome $o \in O$ according to function $\rho : A \rightarrow O$. So an action may be “choose box number 1” and an outcome may be “box number 1 containing \$50 was opened”. Each agent’s private type $\theta_i \in \Theta_i$ deter-

a priori probabilities, seems to me a meaningless question; and even if it had a meaning I do not see how it could be answered”.

⁵Another finding of the psychological literature is that people often overweight their own judgements (see, e.g., [McClelland and Bolger, 1994]); this is perhaps unsurprising since, as Yaniv and Klein [2000] discuss, “decision makers have privileged access to their internal reasons for holding their own opinion,” as opposed to that of others. Our model can naturally incorporate this fact with agents having inflated “self-weights” in their belief revision schemes.

⁶Related trust models appear in the multi-agent systems literature [Wang and Singh, 2007]. See [Ramchurn *et al.*, 2004] for an overview of trust in such settings, and also [Bhattacharya *et al.*, 1998] for a more economics-based perspective.

mines: a value for each outcome, with value function $v_i : \Theta_i \times O \rightarrow \mathfrak{R}$, and also beliefs about the outcome that would be generated by any chosen action. These beliefs can come in the form of a probability distribution over outcomes (so an agent who considers himself to have perfect knowledge of ρ would assign probability 1 to a particular outcome o given a), but for our purposes we need only assume that they determine an intrinsic expected social welfare value, given any type profile and action choice. We use notation:

$$\mathbb{E}_{\theta_i}[v(\theta', \rho(a))] \quad (1)$$

to represent the expectation, with respect to beliefs about ρ inherent in private type θ_i , of the social welfare that will result given that action a is chosen and the true type profile is $\theta' \in \Theta$. $v(\theta', o)$ is used as shorthand for $\sum_{j \in I} v_j(\theta'_j, o)$, i.e., the social welfare given types θ' and outcome o .

We have purposefully described a more or less completely general private values decision making setting with uncertainty which, in the form of assumptions below, we will have to back off of somewhat and will also enrich. We will limit our analysis to settings where private types are *just* about beliefs and not about distinguishing values:

Assumption 1 (outcome fully determines value). $\forall i \in I, \theta'_i, \theta''_i \in \Theta_i, o \in O, v_i(\theta'_i, o) = v_i(\theta''_i, o)$.

In words, the assumption says that once we know the outcome that was realized, we know the value obtained by each agent. So, for instance, an outcome may be an allocation of money to each agent: given the information inherent in the outcome description, if all agents value money the same way then there is no uncertainty about the social welfare. Note that this implies that, for any $\theta_i \in \Theta_i, \theta', \theta'' \in \Theta$ and $a \in A$, $\mathbb{E}_{\theta_i}[v(\theta', \rho(a))] = \mathbb{E}_{\theta_i}[v(\theta'', \rho(a))]$ since ρ is independent of agent types. Because we are making this assumption we can, and will, henceforth use simplified notation $v(o)$ for the social welfare of outcome o (with $v_i(o)$ and $v_{-i}(o)$ also defined accordingly). This is indeed a limiting assumption, excluding common scenarios such as allocation of things like art or music where preferences are private and highly variable; however other important domains—such as making group decisions regarding business plans where the goal is profit—fit the mold.

Each agent i has a *trust level* $t_{i,j} \in [0, 1]$ for every agent j representing the extent to which the beliefs of j are relevant to i 's formation of expectations regarding outcomes. We use notation $\mathfrak{t}_i = (t_{i,1}, \dots, t_{i,n})$ for agent i 's trust vector and we normalize such that $\sum_{j \in I} t_{i,j} = 1, \forall i \in I$. Agent trust levels determine the way updated expectations are formed via a trust-

weighted sum of the agent's intrinsic expectation plus that of the other agents.

Assumption 2 (trust-weighted linear expectation formation). $\forall i \in I, \mathfrak{t}_i \in [0, 1]^n, o \in O$, and $\theta = (\theta_1, \dots, \theta_n) \in \Theta$, agent i 's expectation regarding the social welfare of action a given type profile θ , written $\mathbb{E}_{\theta, \mathfrak{t}_i}[v(\rho(a))]$, equals:

$$\sum_{j \in I} t_{i,j} \cdot \mathbb{E}_{\theta_j}[v(\rho(a))] \quad (2)$$

Finally, we will assume that all agents' trust levels are known to the decision-maker independent of type reports:

Assumption 3 (trust levels are not private). $\forall i \in I, \mathfrak{t}_i$ is *a priori* known to the decision-maker.

Now considering the goal of the decision-maker (or “center,” c), we take a broad approach, allowing for him to also have a distinct trust level $t_{c,i}$ for every agent i in the group. From the decision-maker's perspective, the “efficient” action is the one that maximizes the sum of agent *a priori* expectations regarding social welfare, where each agent's component expectation is weighted according to the degree to which the decision-maker trusts the agent.⁷ Formally, for the decision-maker's trust vector $\mathfrak{t}_c = (t_{c,1}, \dots, t_{c,n})$, we define *trust-efficient choice function* f^* :

$$f^*(\theta) = \arg \max_{a \in A} \sum_{i \in I} t_{c,i} \cdot \mathbb{E}_{\theta_i}[v(\rho(a))] \quad (3)$$

The simplest, and arguably most fair and natural, example of a trust-efficient choice function is where the center trusts each agent equally (i.e., $\forall i \in I, t_{c,i} = \frac{1}{n}$). We call this a *fair-consensus efficient* choice function.⁸

The agents' role in the decision process is to make a report about their private information. Because agents are self-interested and may have conflicting beliefs, there is an incentives problem with respect to eliciting *truthful* reports. For instance, imagine that all agents trust only themselves ($\forall i \in I, t_{i,i} = 1$). Even if all agents have the same value for every outcome—say 0.8 for outcome o_1 and 0.2 for outcome o_2 —if one agent thinks action a_1 will likely yield outcome o_1 and a_2 will yield o_2 , while other agents think a_1 will likely yield o_2 and a_2 will yield o_1 , the first agent has incentive to

⁷Unlike in the case of the agents, though, the center's weights need not semantically map to trust levels. The weights could be defined arbitrarily according to any metric the center prefers; we would then no longer be accurate in calling the choice function “efficient” in the center's view, but all of this paper's results would still go through.

⁸Recalling the discussion from the introduction, one could say fair-consensus efficiency adopts the *equal weight view* in the aggregation of agent beliefs.

overstate his case for action a_1 in the hopes of overriding the other agents. The agent seeks—by lying—to “correct for” what, in his view, are mistaken beliefs of other agents. (A more detailed example follows soon.)

We will address this incentives problem with mechanism design. Formally a mechanism is a tuple (f, T) that defines a choice function $f : \Theta \rightarrow A$ and a vector of payment functions $T = (T_1, \dots, T_n)$, with $T_i : \Theta \times O \rightarrow \mathbb{R}$ for each $i \in I$. The payments are a function of the reported type profile *and* the outcome, and thus must be executed subsequent to realization of the outcome.⁹ Each agent i acts to maximize utility $u_i(\theta_i, \hat{\theta})$, where θ_i is i ’s private type and $\hat{\theta}$ is the vector of reported types. Agents have quasilinear utility (by assumption), so $u_i(\theta_i, \hat{\theta}) = v_i(\theta_i, f(\hat{\theta})) + T_i(\hat{\theta})$, which can be reduced to $u_i(\hat{\theta}) = v_i(\rho(f(\hat{\theta}))) + T_i(\hat{\theta})$ in light of Assumption 1. In the mechanisms we propose, expected utility will always be well-defined as a function of intrinsic beliefs; so given the context of such a mechanism, we use notation $\mathbb{E}_{\theta, \mathbb{t}_i}[u_i(\hat{\theta})]$ to denote expected utility to agent i with trust vector \mathbb{t}_i when the true type profile is θ and the reported type profile is $\hat{\theta}$, where again the uncertainty is with respect to ρ prior to outcome realization.

The role of the payments is to align incentives in order to elicit truthful type reporting which, in turn, allows for implementation of the desired choice function. We aim to accomplish this in *ex post Nash equilibrium*, which simply means that no agent can benefit from deviating as long as others don’t deviate, regardless of their types. There is a vast literature in mechanism design related to the solutions we employ (see [Jackson, 2000], Chapter 2 of [Parkes, 2001], or Chapter 2 of [Cavallo, 2008] for introductions), but no background knowledge of this is strictly necessary to understand the results that follow.

2.1 The straightforward special cases

A relatively easy first observation is that if all agents perform the same belief revision, then all agents and the center will have the same (updated) beliefs about outcomes and truthful reporting of types can easily be made the best strategy for everyone. (The proofs for all theorems except Theorem 4 are deferred to the Appendix.)

⁹It is more typical for payments to be executed based only on reported types, independent of the realized outcome. Our approach has parallels to [Mezzetti, 2004], who addresses interdependent values settings and proposes “two-stage” Groves mechanisms where agents report types, an outcome is chosen, agents then report their achieved value (the second reporting stage), and payments are made. We do not require a second reporting stage because we address settings where agent values are unambiguous given knowledge of the outcome that’s realized (Assumption 1).

Theorem 1. *Given Assumptions 1, 2, and 3, if $\forall i, j \in I$, $t_{i,j} = t_{c,j}$, and the decision-maker implements a trust-efficient choice function and pays each agent the value obtained by the other agents, truthful reporting is an *ex post Nash equilibrium*.*

If, instead, agent beliefs are completely impervious to the opinions of others in the group,¹⁰ recognizing a direct parallel between beliefs about social welfare and “preferences” (given an incentive-aligning “expected Groves” payment)¹¹ yields a solution.

Theorem 2. *Given Assumptions 1, 2, and 3, if $\forall i \in I$, $t_{i,i} = 1$ and the decision-maker implements a fair-consensus efficient choice function and pays each agent the values obtained by the other agents for the realized outcome plus the other agents’ intrinsic expected values for the chosen action, truthful reporting is a dominant strategy.*

2.2 An example

We have seen that “consensus trust” and “no trust” scenarios are relatively easily addressed. But in cases where there is a discrepancy in trust values, manipulation may indeed be beneficial without a substantially more sophisticated mechanism. Consider the example of Table 1, which represents a scenario with two agents and two alternatives. Agent 1 believes the first alternative would yield social value 0.7 and the second would yield 0.3. But agent 2 believes alternative 2 is more efficient, yielding 0.6 compared to 0.4 for alternative 1. Each agent has a trust level of 0.2 for the other agent (and 0.8 for himself).

	$\mathbb{E}_{\theta_1}[v(\rho(a))]$	$\mathbb{E}_{\theta_2}[v(\rho(a))]$
a_1	0.7	0.4
a_2	0.3	0.6

trust level $t_{1,2}$	trust level $t_{2,1}$
0.2	0.2

Table 1: 2-agent, 2-alternative decision problem. Agent 1 prefers alternative 1 and agent 2 prefers alternative 2. Each agent has trust level 0.2 for the other agent.

¹⁰In certain environments this is hardly an indefensible attitude; see, e.g., [van Inwagen, 1996], for a philosophical defense of sticking to a belief in the face of contradictory views held by people whose expertise you respect.

¹¹A so-called Groves mechanism pays each agent the aggregate value that *other* agents obtain, thus aligning incentives by making each agent’s expected utility equal to social welfare [Groves, 1973].

Consider a mechanism that chooses a fair-consensus efficient action (i.e., ascribes $t_{c,i} = \frac{1}{n}$ for each $i \in I$) and, as in Theorem 1, simply makes Groves payments (each agent is payed the value obtained by the other), aligning agent utilities with social welfare. If agents are truthful, then a_1 will be chosen (i.e., $f^*(\theta) = a_1$) and agent 1's expected utility is as follows. $\mathbb{E}_{\theta, \mathbb{t}_1}[u_1(\theta)] =$

$$t_{1,1} \cdot \mathbb{E}_{\theta_1}[v(\rho(f^*(\theta)))] + t_{1,2} \cdot \mathbb{E}_{\theta_2}[v(\rho(f^*(\theta)))] \quad (4)$$

$$= (1 - 0.2) \cdot \mathbb{E}_{\theta_1}[v(\rho(a_1))] + 0.2 \cdot \mathbb{E}_{\theta_2}[v(\rho(a_1))] \quad (5)$$

$$= 0.8 \cdot 0.7 + 0.2 \cdot 0.4 = 0.64 \quad (6)$$

Agent 2's is $\mathbb{E}_{\theta, \mathbb{t}_2}[u_2(\theta)] =$

$$t_{2,2} \cdot \mathbb{E}_{\theta_2}[v(\rho(f^*(\theta)))] + t_{2,1} \cdot \mathbb{E}_{\theta_1}[v(\rho(f^*(\theta)))] \quad (7)$$

$$= 0.8 \cdot 0.4 + 0.2 \cdot 0.7 = 0.46 \quad (8)$$

However, agent 2 expects to benefit from over-reporting his preference for alternative 2. If he reports 0.9 rather than 0.6, a_2 will be chosen and his expected utility will be $\mathbb{E}_{\theta, \mathbb{t}_2}[u_2(\theta_1, \theta'_2)] =$

$$t_{2,2} \cdot \mathbb{E}_{\theta_2}[v(\rho(f^*(\theta_1, \theta'_2)))] + t_{2,1} \cdot \mathbb{E}_{\theta_1}[v(\rho(f^*(\theta_1, \theta'_2)))] \quad (9)$$

$$= (1 - 0.2) \cdot \mathbb{E}_{\theta_2}[v(\rho(a_2))] + 0.2 \cdot \mathbb{E}_{\theta_1}[v(\rho(a_2))] \quad (10)$$

$$= 0.8 \cdot 0.6 + 0.2 \cdot 0.3 = 0.54 \quad (11)$$

The problem is not that agents want to achieve different things from the outcome—the Groves payment ensures that both are motivated only by social welfare—but rather that they have contradictory views about how best to achieve the goal, brought on as a result of conflicting trust levels (each thinks more highly of himself than the other).

3 The mechanism

We now present a general solution that will make truthful reporting in everyone's best interest, for all combinations of pairwise trust levels. Our proposed mechanism elicits type reports from the agents, chooses an outcome, observes the values obtained, and makes monetary payments. Intuitively, the mechanism makes a standard Groves *incentive-aligning* payment to orient all agents towards social-welfare maximization, and also makes a *trust-aligning* payment so that agents will be oriented towards weighing other agents' beliefs in the same way the center does. In actual fact payments will not change the “trust” one agent puts in another's beliefs, but they will change the incentives so that a rational agent with quasilinear utility will *act as though* trust were aligned in this way.

Mechanism 1. Given reports $\hat{\theta} \in \Theta$ and trust vectors $\mathbb{t}_c, \mathbb{t}_1, \dots, \mathbb{t}_n$, action $f^*(\hat{\theta}) = \arg \max_{a \in A} \sum_{j \in I} t_{c,j} \cdot \mathbb{E}_{\hat{\theta}_j}[v(\rho(a))]$ is implemented, an outcome $o^* = \rho(f^*(\hat{\theta}))$ is realized, and payments are executed for each $i \in I$:

$$T_i(\hat{\theta}, o^*) = v_{-i}(o^*) + \quad (12)$$

$$\sum_{j \in I \setminus \{i\}} \left(\frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} - t_{i,j} \right) \cdot \mathbb{E}_{\hat{\theta}_j}[v(\rho(f^*(\hat{\theta})))]$$

The incentive-aligning (Groves) payment is simply the value that others obtain from the realized outcome, and the trust-aligning payment is: summed over each other agent j , the ratio of the agent i 's self-trust to the center's trust in i , times the center's trust in j , minus i 's trust in j , all multiplied by j 's intrinsic expectation of the social value of the chosen action. In the special case where the center implements a fair-efficient choice function (equal trust for each agent), the payments reduce to:

$$T_i(\hat{\theta}, o^*) = v_{-i}(o^*) + \sum_{j \in I \setminus \{i\}} (t_{i,i} - t_{i,j}) \cdot \mathbb{E}_{\hat{\theta}_j}[v(\rho(f^*(\hat{\theta})))] \quad (13)$$

where the trust aligning payment is just the difference between the agent's self-trust and his trust in each other agent times that agent's expectation. So if self-trust is higher the agent is given a larger payment to boost his concern for the other's expectation; if it is lower he is charged to decrease his concern for the other agent's view. In the further specialization where agents trust all others equally (i.e., $t_{i,j} = \frac{1-t_{i,i}}{n-1}$ for each $j \in I \setminus \{i\}$),

$$T_i(\hat{\theta}, o^*) = v_{-i}(o^*) + \frac{nt_{i,i} - 1}{n-1} \cdot \sum_{j \in I \setminus \{i\}} \mathbb{E}_{\hat{\theta}_j}[v(\rho(f^*(\hat{\theta})))] \quad (14)$$

The mechanism is successful in that participants are always best-off reporting their true beliefs:

Theorem 3. Given Assumptions 1, 2, and 3, for any trust vectors $\mathbb{t}_c, \mathbb{t}_1, \dots, \mathbb{t}_n$, Mechanism 1 implements a trust-efficient choice function in a truthful ex post Nash equilibrium.

The equilibrium is ex post Nash rather than dominant strategy because truthtelling is optimal regardless of others' types *so long as they report truthfully*, since it is their true types that form the beliefs that are relevant to others' expectation formation. Each agent i 's expected utility from truthful participation given joint type θ (see the proof of Theorem 3 in the Appendix

for a derivation) equals:

$$\frac{t_{i,i}}{t_{c,i}} \cdot \sum_{j \in I} t_{c,j} \cdot \mathbb{E}_{\theta_j}[v(\rho(f^*(\theta)))] \quad (15)$$

This is the center’s expectation of social welfare times the ratio of i ’s self-trust to the center’s trust in i . Note that Theorems 1 and 2 are corollaries of Theorem 3. Theorem 1 covers the special case where each agent trusts all others and himself the same way the center does, and the second term of the payment defined in Mechanism 1 then reduces to 0. Theorem 2 covers the case where the center trusts all agents equally and each agent doesn’t trust others at all, and in that case Mechanism 1 will pay each agent the others’ realized values plus their intrinsic expected values (since $\frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} - t_{i,j} = 1$), as in Theorem 2.

Returning to the example of Table 1, again taking the center’s weights to be equal (i.e., 0.5) for each of the two agents, expected utility to the agents from truthful participation given types θ , drawing from Eq. (15), is as follows:

$$\begin{aligned} \mathbb{E}_{\theta, t_1}[u_1(\theta)] &= \frac{0.8}{0.5} \cdot \left(0.5 \cdot \mathbb{E}_{\theta_1}[v(\rho(a_1))] + 0.5 \cdot \mathbb{E}_{\theta_2}[v(\rho(a_1))]\right) \\ &= 0.8 \cdot (0.7 + 0.4) = 0.88 \end{aligned} \quad (16)$$

$$\begin{aligned} \mathbb{E}_{\theta, t_2}[u_2(\theta)] &= \frac{0.8}{0.5} \cdot \left(0.5 \cdot \mathbb{E}_{\theta_2}[v(\rho(a_1))] + 0.5 \cdot \mathbb{E}_{\theta_1}[v(\rho(a_1))]\right) \\ &= 0.8 \cdot (0.4 + 0.7) = 0.88 \end{aligned} \quad (17)$$

The manipulation that was beneficial for agent 2 with simple Groves payments goes away. If agent 2 were to report expected welfare 0.8 for action a_2 now, a_2 would be chosen and his expected utility would be lowered to:

$$\begin{aligned} \mathbb{E}_{\theta, t_2}[u_2(\theta_1, \theta'_2)] &= \frac{0.8}{0.5} \cdot \left(0.5 \cdot \mathbb{E}_{\theta_2}[v(\rho(a_2))] + 0.5 \cdot \mathbb{E}_{\theta_1}[v(\rho(a_2))]\right) \\ &= 0.8 \cdot (0.6 + 0.3) = 0.72 \end{aligned} \quad (18)$$

4 Balancing the budget

Mechanism 1 succeeds in achieving equilibrium implementation of the desired choice function, but it runs a large deficit. This is problematic because it implies that external subsidies are required for the mechanism’s execution. Funds can be recovered by charging each agent an extra quantity that is dependent only on the *others*’ beliefs, and this will not distort incentives. Ideally a payment scheme would charge agents enough to achieve *no-deficit* (aggregate payments to the agents are never positive) and yet maintain *interim individual rationality* (each agent’s expected utility is never

negative in a truthtelling equilibrium).¹²

However, in the current model there are significant obstacles to this goal. Specifically, the $v_{-i}(o^*)$ term paid to each agent i after realization of outcome o^* will in some cases be impossible to overcome with a payment based on others’ beliefs, because of the difficulty of computing a charge based on prior expectations that is universally (even in expectation) both high enough to overcome this payment and low enough to avoid IR violations *for all possible beliefs* i might have. Imagine that all agents have self-trust weight 0. Then, from Eq. (15) we can see that each agent’s expected utility from Mechanism 1 is 0, and so we cannot impose an additional charge on any agent. Yet, the mechanism may run a large deficit even if agents have 0 self-trust; specifically, this will occur when the obtained social value is larger than was expected.

Despite this challenge we can successfully proceed to a budget balancing technique if we make a further assumption that is rather natural given the type of setting we’re considering here:

Assumption 4 (common payoff). $\forall i, j \in I, \theta \in \Theta, o \in O, v_i(\theta_i, o) = v_j(\theta_j, o)$.

This assumption is valid in environments such as the toy example from the introduction where agents are facing a choice between money-containing boxes (with the money split evenly), or more realistic scenarios such as coordinating business decisions amongst managers who share the goal of company profit. We will also assume here that actions can’t bring negative value to agents (e.g., anything a chosen box contains can be costlessly discarded).¹³

Assumption 5 (no negative values). $\forall i \in I, \theta \in \Theta, o \in O, v_i(\theta_i, o) \geq 0$.

The budget-balanced mechanism we propose modifies Mechanism 1 by imposing a charge on each agent i that is a function of other agents reported types, specifically considering the action that would be chosen if i were ignored or not present. We let $f^*(\theta_{-i})$ denote an action choice that optimizes for the beliefs of agents other than i . $\forall \theta_{-i} \in \Theta_{-i}$,

$$f^*(\theta_{-i}) \in \arg \max_{a \in A} \sum_{j \in I \setminus \{i\}} t_{c,j} \cdot \mathbb{E}_{\theta_j}[v(\rho(a))] \quad (19)$$

Importantly, this action is completely independent of the reported type of agent i and thus can play a role in

¹²Note from Eq. (15) that Mechanism 1 satisfies ex post individual rationality (and by a wide margin) if social value for any outcome is known to be non-negative.

¹³A similar assumption is required to achieve ex post individual rationality in the case of the VCG mechanism for standard settings without uncertainty (see Theorem 2.8 of [Cavallo, 2008]).

a payment mechanism without influencing incentives. The mechanism:

Mechanism 2. Given reports $\hat{\theta} \in \Theta$ and trust vectors $\mathbb{t}_c, \mathbb{t}_1, \dots, \mathbb{t}_n$, action $f^*(\hat{\theta}) = \arg \max_{a \in A} \sum_{j \in I} t_{c,j} \cdot \mathbb{E}_{\hat{\theta}_j}[v(\rho(a))]$ is implemented, an outcome $o^* = \rho(f^*(\hat{\theta}))$ is realized, and payments are executed for each $i \in I$:

$$T_i(\hat{\theta}, o^*) = \sum_{j \in I \setminus \{i\}} \left(\frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} - t_{i,j} \right) \cdot \mathbb{E}_{\hat{\theta}_j}[v_i(\rho(f^*(\hat{\theta})))]$$
 (20)

$$- \sum_{j \in I \setminus \{i\}} \frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} \cdot \mathbb{E}_{\hat{\theta}_j}[v_i(\rho(f^*(\hat{\theta}_{-i})))]$$
 (21)

Thanks to Assumption 4, the incentive-aligning Groves payment is no longer required, so the mechanism provides only the trust-aligning payment (Eq. (20)) and then a budget-recovery charge (Eq. (21)) that is independent of the agent's report.

Theorem 4. Given Assumptions 1, 2, 3, and 4, for arbitrary trust vectors $\mathbb{t}_c, \mathbb{t}_1, \dots, \mathbb{t}_n$, Mechanism 2 implements a trust-efficient choice function in a truthful ex post Nash equilibrium.

Proof. Consider arbitrary joint type $\theta \in \Theta$ and agent $i \in I$. i 's expected utility, at the time of reporting, for reporting type $\hat{\theta}_i \in \Theta_i$ when others truthfully report θ_{-i} is:

$$t_{i,i} \cdot \mathbb{E}_{\theta_i}[v_i(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]$$
 (22)

$$+ \sum_{j \in I \setminus \{i\}} t_{i,j} \cdot \mathbb{E}_{\theta_j}[v_i(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]$$
 (23)

$$+ \sum_{j \in I \setminus \{i\}} \frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} \cdot \mathbb{E}_{\theta_j}[v_i(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]$$
 (24)

$$- \sum_{j \in I \setminus \{i\}} t_{i,j} \cdot \mathbb{E}_{\theta_j}[v_i(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]$$
 (25)

$$- \sum_{j \in I \setminus \{i\}} \frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} \cdot \mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta_{-i})))]$$
 (26)

$$= \frac{t_{i,i}}{t_{c,i}} \cdot \left(t_{c,i} \cdot \mathbb{E}_{\theta_i}[v_i(\rho(f^*(\hat{\theta}_i, \theta_{-i}))) \right]$$
 (27)

$$+ \sum_{j \in I \setminus \{i\}} t_{c,j} \cdot \mathbb{E}_{\theta_j}[v_i(\rho(f^*(\hat{\theta}_i, \theta_{-i}))) \right]$$
 (28)

$$- \sum_{j \in I \setminus \{i\}} \frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} \cdot \mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta_{-i})))]$$
 (29)

The last term (Eq. (29)) is independent of $\hat{\theta}_i$ and thus can be ignored from an incentives perspective. Ignor-

ing constant multiplier $\frac{t_{i,i}}{t_{c,i}}$, the remaining quantity—in light of Assumption 4—is precisely what f^* is defined to maximize, and so truthful reporting of $\hat{\theta}_i = \theta_i$ is a best-response. \square

Assumption 5 is not necessary for the above result demonstrating ex post efficiency, but it will be required to establish individual rationality and no-deficit.

Theorem 5. Given Assumptions 1, 2, 3, 4, and 5, for arbitrary trust vectors $\mathbb{t}_c, \mathbb{t}_1, \dots, \mathbb{t}_n$, Mechanism 2 is interim individual rational.

Theorem 6. Given Assumptions 1, 2, 3, 4, and 5, for arbitrary trust vectors $\mathbb{t}_c, \mathbb{t}_1, \dots, \mathbb{t}_n$, Mechanism 2 never runs a deficit.

For type profile θ , simplifying from Eqs. (27–29), any agent i 's expected utility from participating in the mechanism is:

$$\frac{t_{i,i}}{t_{c,i}} \cdot \left(t_{c,i} \cdot \mathbb{E}_{\theta_i}[v_i(\rho(f^*(\theta))) \right]$$
 (30)

$$\sum_{j \in I \setminus \{i\}} t_{c,j} \cdot \left(\mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta_{-i}))) \right]$$

– $\mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta)))] \right)$

In words this is the ratio of the agent i 's self-trust level to the center's trust level for i times: the center's trust level for i times i 's intrinsic expectation of the value of the outcome, minus the extent to which the other agents each intrinsically believe they will suffer from i 's influence on the action choice, weighted by the center's trust in each. This value is always non-negative and, interestingly, increases with the agent's self-trust level and also with the center's trust in the agent.¹⁴ In the case where the center seeks to implement a fair-consensus efficient action, payment to i reduces to:

$$T_i(\hat{\theta}, o^*) = \sum_{j \in I \setminus \{i\}} \left((t_{i,i} - t_{i,j}) \cdot \mathbb{E}_{\hat{\theta}_j}[v_i(\rho(f^*(\hat{\theta}))) \right]$$

$$- t_{i,i} \cdot \mathbb{E}_{\hat{\theta}_j}[v_i(\rho(f^*(\hat{\theta}_{-i}))) \right]$$
 (31)

and i 's expected utility reduces to:

$$t_{i,i} \cdot \left(\mathbb{E}_{\theta_i}[v_i(\rho(f^*(\theta)))] \right]$$
 (32)

$$\sum_{j \in I \setminus \{i\}} \left(\mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta_{-i}))) \right]$$

¹⁴Growth with $t_{i,i}$ can be seen directly from Eq. (30). To see growth with $t_{c,i}$ consider the algebraic rearrangement: $t_{i,i} \cdot \mathbb{E}_{\theta_i}[v_i(\rho(f^*(\theta)))] - \frac{t_{i,i}}{t_{c,i}} \sum_{j \in I \setminus \{i\}} t_{c,j} \cdot \left(\mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta_{-i}))) \right]$ – $\mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta)))]$. The subtracted term is non-negative and so as $t_{c,i}$ increases it gets smaller.

Returning again to the example of Table 1, under Mechanism 2 expected utility to the agents from truthful participation is as follows:

$$\begin{aligned} \mathbb{E}_{\theta, \mathfrak{t}_1}[u_1(\theta)] &= 0.8 \cdot \left(\mathbb{E}_{\theta_1}[v_1(\rho(a_1))] - \mathbb{E}_{\theta_2}[v_1(\rho(a_2))] + \mathbb{E}_{\theta_2}[v_1(\rho(a_1))] \right) \\ &= 0.8 \cdot \left(\frac{0.7}{2} - \frac{0.6}{2} + \frac{0.4}{2} \right) = 0.2 \end{aligned} \quad (33)$$

$$\begin{aligned} \mathbb{E}_{\theta, \mathfrak{t}_2}[u_2(\theta)] &= 0.8 \cdot \left(\mathbb{E}_{\theta_2}[v_2(\rho(a_1))] + -\mathbb{E}_{\theta_1}[v_2(\rho(a_1))] + \mathbb{E}_{\theta_1}[v_2(\rho(a_1))] \right) \\ &= 0.8 \cdot \left(\frac{0.4}{2} - \frac{0.7}{2} + \frac{0.7}{2} \right) = 0.16 \end{aligned} \quad (34)$$

And revenue equals:

$$\left(0.8 \cdot \frac{0.6}{2} - 0.6 \cdot \frac{0.4}{2} \right) + \left(0.8 \cdot \frac{0.7}{2} - 0.6 \cdot \frac{0.7}{2} \right) = 0.26 \quad (35)$$

5 Discussion

To motivate some closing reflections, consider this example: A group of several college friends, some Canadian and some American, have decided to start a company together and are debating whether to establish its headquarters in the United States or Canada. All the friends have identical expectations regarding all elements of uncertainty relevant to this decision except for one: will the current American president be re-elected or will his challenger win? They agree that the outcome of this event will have serious implications on the business climate in the USA, and they even agree on what the implications would be; but they do *not* agree on the likelihood of re-election. How can they reach a decision that incorporates all views and leaves no one with an incentive to misrepresent their true beliefs?

For settings like this proposals that start with “choose the Bayesian optimal action” border on the absurd, yet there are essentially no others in the previous mechanism design literature. Each individual’s opinion may be extremely well researched and thought-through, but for a *unique* event such as an election, the outcome of which depends on the free choices of human beings (the voters), there is no “objective” answer to the question of what probability distribution over outcomes is “correct”; in fact, it’s questionable whether discussion of probability even makes sense here at all. Instead, each individual—when confronted with the beliefs of his peers—will arrive at an updated belief about the situation through a process that reflects his trust in the reliability of the others’ opinions.

Exactly what that process is will surely vary from person to person. In this paper we’ve taken the most

prevalent model available in the psychological literature, made it broad and adaptable by incorporating individualized trust levels, and used it as the basis for an analysis of the incentives of group decision making problems. We provided a payment mechanism that solves the incentive problem without running a deficit, so long as certain assumptions are met. So, returning to our example, if one of the peers has been given decision-making authority (perhaps by virtue of his founding status), he can implement the payments we specified and: he will not require an external budget to do so, all agents will be provably best off reporting truthfully, and none will want to abandon the group due to fears about obtaining negative utility.

Because this work proposes a theory about how to engineer good decisions in the real world, the question of its weaknesses is intimately tied to the question of the validity of the assumptions inherent in it. Beyond the model of belief revision that we adopt (which can be adjusted as theories evolve based on further empirical studies), the most notable required assumption is that agent trust levels are a priori known to the decision-maker. In many circumstances that we’d like to be able to address, the way individuals view the reliability of others is private to them, and should thus be construed as part of their type. But, alas, the outlook for obtaining positive results without this assumption is not bright, particularly in the shade of general negative results in mechanism design for multidimensional, interdependent types.¹⁵ Yet there are settings where the assumption will have legitimacy. In some cases aspects of an individual’s past behavior unrelated to the decision problem may be indicative of his trust level of other individuals (although care must be taken with any such inferences not to create dangerous opportunities for gaming). In other cases the trustworthiness of each agent (or a good approximation thereof) may naturally be common-knowledge, perhaps based on a clear record of past prediction performance and/or pairwise interactions. Further exploration of practical ways to mitigate this issue is one important direction for future work.

References

- [Ariely *et al.*, 2000] Dan Ariely, Wing T. Au, Randall H. Bender, David V. Budescu, Christiane B. Dietz, Hongbin Gu, Thomas S. Wallsten, and Gal Zauberman. The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, 6(2):130–147, June 2000.
- [Bhattacharya *et al.*, 1998] R. Bhattacharya, T. M. Devinney, and M. M. Pillutla. A formal model of trust based

¹⁵[Jehiel *et al.*, 2006] shows that only constant functions are implementable in ex post equilibrium when types are multidimensional and interdependent.

- on outcomes. *The Academy of Management Review*, 23(3):459–472, July 1998.
- [Bonanno and Nehring, 1999] Giacomo Bonanno and Klaus Nehring. How to make sense of the common prior assumption under incomplete information. *International Journal of Game Theory*, 28(3):409–434, 1999.
- [Budescu and Yu, 2006] David V. Budescu and Hsiu-Ting Yu. To bayes or not to bayes? a comparison of two classes of models of information aggregation. *Decision Analysis*, 3(3):145–162, September 2006.
- [Budescu et al., 2003] David V. Budescu, Adrian K. Rantilla, Hsiu-Ting Yu, and Tzur M. Karelitz. The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, 90(1):178–194, January 2003.
- [Cavallo, 2008] Ruggiero Cavallo. *Social Welfare Maximization in Dynamic Strategic Decision Problems*. Ph.D. Thesis, Harvard University, 2008.
- [de Finetti, 1964] Bruno de Finetti. Foresight: its logical laws, its subjective sources. In H. E. Kyburg and H. E. Smokler, editors, *Studies in Subjective Probability*. Wiley, New York, 1964. translation of 1937 article originally in French.
- [Dekel and Gul, 1997] Eddie Dekel and Faruk Gul. Rationality and knowledge in game theory. In David M. Kreps and Kenneth F. Wallis, editors, *Advances in Economics and Econometrics*, volume 1, pages 87–172. Cambridge University Press, 1997.
- [Elga, 2007] Adam Elga. Reflection and disagreement. *Nous*, 41(3):478–502, 2007.
- [Fitelson and Jehle, 2009] Branden Fitelson and David Jehle. What is the “Equal Weight View”? *Episteme*, 6(3):280–293, 2009.
- [Genest and Zidek, 1986] Christian Genest and James V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, February 1986.
- [Groves, 1973] Theodore Groves. Incentives in teams. *Econometrica*, 41:617–631, 1973.
- [Hardwig, 1991] John Hardwig. The role of trust in knowledge. *The Journal of Philosophy*, 88(12):693–708, December 1991.
- [Harsanyi, 1967 68] J.C. Harsanyi. Games with incomplete information played by ‘Bayesian’ players. *Management Science*, 14:159–189, 320–334, 486–502, 1967–68.
- [Jackson, 2000] Matthew O. Jackson. Mechanism theory. In *The Encyclopedia of Life Support Systems*. EOLSS Publishers, 2000.
- [Jehiel et al., 2006] Philippe Jehiel, Moritz Meyer ter Vehn, Benny Moldovanu, and William R. Zame. The limits of ex post implementation. *Econometrica*, 74(3):585–610, 2006.
- [Lehrer and Wagner, 1981] Keith Lehrer and Carl Wagner. *Rational Consensus in Science and Society*. D. Reidel Publishing Co., Dordrecht, Holland, 1981.
- [Lehrer, 1983] Keith Lehrer. Rationality as weighted averaging. *Synthese*, 57(3):283–295, 1983.
- [McClelland and Bolger, 1994] A. G. R. McClelland and F. Bolger. The calibration of subjective probabilities: Theories and models 1980–94. In G Wright and P Ayton, editors, *Subjective Probability*, pages 453–482. Chichester: Wiley, 1994.
- [Mezzetti, 2004] Claudio Mezzetti. Mechanism design with interdependent valuations: Efficiency. *Econometrica*, 72(5):1617–1626, 2004.
- [Morris, 1995] Stephen Morris. The common prior assumption in economic theory. *Economics and Philosophy*, 11:227–253, 1995.
- [Parkes, 2001] David C. Parkes. *Iterative Combinatorial Auctions: Achieving Economic and Computational Efficiency*. PhD Thesis, Department of Computer and Information Science, University of Pennsylvania, 2001.
- [Ramchurn et al., 2004] Sarvapali D. Ramchurn, Huynh Dong, and Nicholas R. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(1):1–25, 2004.
- [Ramsey, 1931] Frank P. Ramsey. *Truth and probability*. Routledge, 1931. Chapter VII, reprinted in 2001.
- [van Inwagen, 1996] Peter van Inwagen. Is it wrong, always, everywhere, and for anyone, to believe anything, upon insufficient evidence? In Jeff Jordan and Daniel Howard-Snyder, editors, *Faith, Freedom, and Rationality*, pages 137–154. Rowman and Littlefield, 1996.
- [Wallsten et al., 1999] Thomas S. Wallsten, David V. Budescu, I. D. O. Erev, and Adele Diederich. Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, 10(3):243–268, April 1999.
- [Wang and Singh, 2007] Yonghong Wang and Munindar P. Singh. Formal trust model for multiagent systems. In *Proceedings of the 20th international joint conference on Artificial intelligence*, pages 1551–1556, 2007.
- [Yaniv and Kleinberger, 2000] Ilan Yaniv and Eli Kleinberger. Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2):260–281, November 2000.
- [Yaniv, 2004] Ilan Yaniv. Receiving other people’s advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93(1):1–13, January 2004.

Appendix

Proof of Theorem 1. For any type profile $\theta \in \Theta$ and any agent $i \in I$, if agents other than i are truthful, i 's expected utility $\mathbb{E}_{\theta, t_i}[u_i(\hat{\theta}_i, \theta_{-i})]$ from type report $\hat{\theta}_i$ will be expected social welfare and will equal:

$$t_{i,i} \cdot \mathbb{E}_{\theta_i}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) + \sum_{j \in I \setminus \{i\}} t_{i,j} \cdot \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) \quad (36)$$

$$= t_{c,i} \cdot \mathbb{E}_{\theta_i}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) + \sum_{j \in I \setminus \{i\}} t_{c,j} \cdot \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) \quad (37)$$

$$= \sum_{j \in I} t_{c,j} \cdot \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) \quad (38)$$

By definition of f^* , this is maximized by truthful report $\hat{\theta}_i = \theta_i$. \square

Proof of Theorem 2. The proposed mechanism defines, for all $i \in I$, $\hat{\theta} \in \Theta$, and realized outcome $o \in O$, $T_i(\hat{\theta}, o) = v_{-i}(o) + \sum_{j \in I \setminus \{i\}} \mathbb{E}_{\hat{\theta}_j}[v(\rho(f^*(\hat{\theta})))])$. Assume the true joint type is θ and agents other than i truthfully report θ_{-i} . If agent i reports $\hat{\theta}_i$, noting that $t_{i,i} = 1$ and $t_{i,j} = 0$ for all $j \neq i$, i 's expectation regarding social welfare will simply remain $\mathbb{E}_{\theta_i}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))])$, reduced from $t_{i,i} \cdot \mathbb{E}_{\theta_i}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) + \sum_{j \in I \setminus \{i\}} t_{i,j} \cdot \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))])$, and so his expected utility $\mathbb{E}_{\theta, t_i}[u_i(\hat{\theta}_i, \theta_{-i})]$ at the time of reporting will be:

$$\mathbb{E}_{\theta_i}[v_i(\rho(f^*(\hat{\theta}_i, \theta_{-i}))) + v_{-i}(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) + \sum_{j \in I \setminus \{i\}} \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) \quad (39)$$

$$= \mathbb{E}_{\theta_i}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) + \sum_{j \in I \setminus \{i\}} \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) \quad (40)$$

$$= n \sum_{j \in I} \frac{1}{n} \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) \quad (41)$$

By definition of fair-consensus efficiency, this is maximized by truthful report $\hat{\theta}_i = \theta_i$. This is true regardless of whether the other agents' reports $\hat{\theta}_{-i}$ are truthful or not, and so the result is dominant strategy rather than just ex post Nash equilibrium. \square

Proof of Theorem 3. We will demonstrate that truthful type reporting is an ex post Nash equilibrium, which will demonstrate efficiency (according to the center's trust weights) by definition of the mechanism. Consider arbitrary joint type $\theta \in \Theta$ and agent $i \in I$. i 's realized utility if he reports type $\hat{\theta}_i \in \Theta_i$, others truthfully report θ_{-i} , and outcome o is realized will be:

$$v(o) + \sum_{j \in I \setminus \{i\}} \left(\frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} - t_{i,j} \right) \cdot \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) \quad (42)$$

Recalling how expectations are formed from Assumption 2, at the time of type reporting agent i 's *expected* utility for announcing type $\hat{\theta}_i$ is then $\mathbb{E}_{\theta, t_i}[u_i(\hat{\theta}_i, \theta_{-i})] =$

$$t_{i,i} \cdot \mathbb{E}_{\theta_i}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) + \sum_{j \in I \setminus \{i\}} t_{i,j} \cdot \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) + \quad (43)$$

$$\sum_{j \in I \setminus \{i\}} \frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} \cdot \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) - \sum_{j \in I \setminus \{i\}} t_{i,j} \cdot \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) \quad (44)$$

$$= t_{i,i} \cdot \mathbb{E}_{\theta_i}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) + \frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} \cdot \sum_{j \in I \setminus \{i\}} \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) \quad (45)$$

$$= \frac{t_{i,i}}{t_{c,i}} \cdot \sum_{j \in I} t_{c,j} \cdot \mathbb{E}_{\theta_j}[v(\rho(f^*(\hat{\theta}_i, \theta_{-i})))]) \quad (46)$$

By definition of f^* , this is maximized by truthful report $\hat{\theta}_i = \theta_i$. \square

Proof of Theorem 5. Consider arbitrary type profile $\theta \in \Theta$ and agent $i \in I$. If θ is reported, i 's expected utility will be, simplifying from Eq. (29),

$$\frac{t_{i,i}}{t_{c,i}} \cdot \left(t_{c,i} \cdot \mathbb{E}_{\theta_i}[v_i(\rho(f^*(\theta)))] + \sum_{j \in I \setminus \{i\}} t_{c,j} \cdot \mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta)))] - \sum_{j \in I \setminus \{i\}} t_{c,j} \cdot \mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta_{-i})))] \right) \quad (47)$$

$$\geq \frac{t_{i,i}}{t_{c,i}} \cdot \left(t_{c,i} \cdot \mathbb{E}_{\theta_i}[v_i(\rho(f^*(\theta_{-i})))] + \sum_{j \in I \setminus \{i\}} t_{c,j} \cdot \mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta_{-i})))] - \sum_{j \in I \setminus \{i\}} t_{c,j} \cdot \mathbb{E}_{\theta_j}[v_i(\rho(f^*(\theta_{-i})))] \right) \quad (48)$$

$$= \frac{t_{i,i}}{t_{c,i}} \cdot \left(t_{c,i} \cdot \mathbb{E}_{\theta_i}[v_i(\rho(f^*(\theta_{-i})))] \right) \geq 0 \quad (49)$$

The first inequality follows from the definition of $f^*(\theta)$; if $f^*(\theta_{-i})$ were superior in expectation given all agents' beliefs, it would have been chosen instead as the trust-efficient action. The second inequality holds by Assumption 5. \square

Proof of Theorem 6. For arbitrary reported types $\hat{\theta}$ and arbitrary agent i , the payment made to i equals:

$$\sum_{j \in I \setminus \{i\}} \left(\frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} - t_{i,j} \right) \cdot \mathbb{E}_{\hat{\theta}_j}[v_i(\rho(f^*(\hat{\theta})))] - \sum_{j \in I \setminus \{i\}} \frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} \cdot \mathbb{E}_{\hat{\theta}_j}[v_i(\rho(f^*(\hat{\theta}_{-i})))] \quad (50)$$

$$\leq \sum_{j \in I \setminus \{i\}} \left(\frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} - t_{i,j} \right) \cdot \mathbb{E}_{\hat{\theta}_j}[v_i(\rho(f^*(\hat{\theta})))] - \sum_{j \in I \setminus \{i\}} \frac{t_{i,i} \cdot t_{c,j}}{t_{c,i}} \cdot \mathbb{E}_{\hat{\theta}_j}[v_i(\rho(f^*(\hat{\theta})))] \quad (51)$$

$$= - \sum_{j \in I \setminus \{i\}} t_{i,j} \cdot \mathbb{E}_{\hat{\theta}_j}[v_i(\rho(f^*(\hat{\theta})))] \leq 0 \quad (52)$$

The first inequality follows from the definition of $f^*(\hat{\theta}_{-i})$, which maximizes the sum of the (center's trust-weighted) expectations of agents other than i , and the second inequality follows from Assumption 5. Since the payment to each agent is never positive, the aggregate payment to all agents is never positive and thus a deficit never results. \square