Mechanisms for Making Accurate Decisions in Biased Crowds

Blake Riley
University of Illinois at Urbana-Champaign
briley2@illinois.edu

November 2015

Abstract

This paper studies procedures for identifying the true answer to a binary question using the opinions of potentially-biased individuals. It's common and natural to side with the majority opinion, but the majority may make the wrong choice when the agents are biased. Taking majority rule as a baseline, I study *peer-prediction decision rules*, which ask agents to predict the opinions of others in addition to providing their own. This extra information enables us to potentially recognize the correct answer even when the majority is wrong.

I first show that peer-prediction rules cannot be more accurate than the majority when we require them to satisfy the same symmetry conditions as majority rule and to be incentive-compatible for agents who intend to push the final decision towards their own opinion. Realistically though, not all agents distort their information strategically. I provide a simple decision rule based on the median agent's prediction that matches majority rule when all agents are strategic and makes more accurate decisions than majority rule when some agents are honest.

1 Introduction

In October 2015, a team of inmates from Eastern New York Correctional Facility debated a three-time national champion team from Harvard. After hearing arguments, the panel of judges decided by majority vote in favor of the inmates[1]. Majority rule is a natural way to make group decisions like this one for multiple reasons. First, it is simple and transparent. Second, the procedure does not favor one side over the other or make distinctions between judges. Finally, it is strategically robust, making honest revelation a dominant strategy if agents want the final decision to match their personal opinion.

[1] Leslie Brody, "Prison vs. Harvard in an Unlikely Debate," *The Wall Street Journal*, Oct. 8, 2015. <http://www.wsj.com/articles/an-unlikely-debate-prison-vs-harvard-1442616928>

Although judges might vote to favor their opinion, the point of the competition is to decide which team is most skilled, not to aggregate the judges' preferences. The choice of winner should accurately reflect which team did better (according to some criteria) for the debate to be legitimate. However, the only way to identify which team is most skilled is through the judges' subjective assessments, and a judge's opinion could be correct or mistaken relative to the underlying truth. Since individual judges can be mistaken, the goal when aggregating opinions is to maximize the probability of choosing the most deserving team.

Information aggregation through voting is a long-studied question initiated by the Marquis de Condorcet in 1785 in his essay on majority decisions. The standard model following Condorcet assumes agents have noisy signals about the true state that are correct more often than not. An example would be debate judges who are 70% likely to vote for Harvard when the Harvard team is in fact better and 30% likely to vote for Harvard when the inmates are actually better. Under Condorcet's model, majority rule is more accurate than any given judge, smoothing out noise in opinions to identify the "wisdom of crowds." However, aggregating noisy signals through majority rule can make matters worse when bias is present. Given the associations that come with Harvard undergraduates versus inmates convicted of violent crimes, a fair assessment is a lot to ask of a judge. Bias wouldn't be surprising and could go in either direction—discounting inmates because of their background or favoring them as underdogs.

Some degree of bias isn't fatal to the performance of majority rule. For instance, suppose each judge is 60% likely to favor Harvard in the state of the world where they are best and 90% likely to favor the inmates when they are best. This is a scenario where the judges are more impressed by a "good" team of inmates than a "good" team from Harvard. Nevertheless, when these opinions are aggregated, the group decision still favors the best team in each state of the world, with the only difference being Harvard wins by a smaller margin.

In contrast, suppose the bias is stronger and judges are 40% likely to correctly favor Harvard and 90% likely to correctly favor Eastern Correctional. The opinions are still correlated with the truth—comparatively more judges favor Harvard when they are best. Nonetheless, the Harvard supporters will be in the minority on average in each state. Majority rule will choose the inmates regardless of the truth.

Debate organizers concerned about potential bias could pick a decision procedure other than majority rule. However, doing so would require insight into the precise nature of the bias. For instance, a unanimity rule where the Harvard team wins only if all judges support them would counteract a bias towards Harvard, but would also exacerbate a bias towards the inmates. Debate organizers might not trust themselves to adjust the decision rule in the right direction, and the teams would be understandably upset at the asymmetric standard even if they did. A satifactory alternative to majority rule needs to be *neutral*, treating each option symmetrically.

Furthermore, groups like corporations, unions, or homeowners' associations need a single rule that can be applied consistently across different contexts. Achieving this

can be difficult given that the degree of bias may change how we should interpret a particular level of support for one team over another. For instance, if judges are biased towards Harvard, Harvard may deserve to lose even with a two-thirds majority. A single rule responsive to different circumstances has to collect additional information from agents beyond their opinions. In particular, a decision rule could ask agents to predict the opinions of other group members. By comparing the actual level of support with the predicted level of support, a "peer-prediction" decision rule can potentially make more accurate decisions than majority rule without knowing the likelihoods of opinions in each state.

Consider the following example: three judges are independently and identically 40% likely to correctly favor the Harvard team and 90% likely to correctly favor the inmates. Each judge puts equal prior probability on either team being best and updates their beliefs after observing their own opinion using Bayes' rule. Let the opinion of judge i be x_i and the best team be ω. Under majority rule, the Harvard team wins with probability

$$\begin{aligned}\Pr[\text{Majority for Harvard} \mid \omega = \text{Harvard}] &= \Pr[\text{Three Harvard supporters} \mid \omega = \text{Harvard}] \\ &\quad + \Pr[\text{Two Harvard supporters} \mid \omega = \text{Harvard}] \\ &= \left(\frac{4}{10}\right)^3 + 3\left(\frac{4}{10}\right)^2 \frac{6}{10} = 0.35\end{aligned}$$

when they're best, and the inmate team wins with probability 0.97 when they are best.

Rather than use majority rule, let's say Harvard wins if the percentage in support of Harvard is greater than the average predicted support for the team. For example, if the average predicted support for Harvard is 70%, then Harvard would win with 80% support and lose with 60%, despite still being favored by the majority. Framing the rule in terms of Eastern Correctional would produce identical decisions, so this rule is neutral. Conditional on their opinion, a Harvard supporter expects another judge to support Harvard with probability

$$\begin{aligned}\Pr[x_j = \text{Harvard} \mid x_i = \text{Harvard}] &= \Pr[x_j = \text{Harvard} \mid \omega = \text{Harvard}] \cdot \Pr[\omega = \text{Harvard} \mid x_i = \text{Harvard}] \\ &\quad + \Pr[x_j = \text{Harvard} \mid \omega = \text{ENYCF}] \cdot \Pr[\omega = \text{ENYCF} \mid x_i = \text{Harvard}] \\ &= \frac{4}{10}\frac{\frac{4}{10}\frac{1}{2}}{\frac{4}{10}\frac{1}{2} + \frac{1}{10}\frac{1}{2}} + \frac{1}{10}\frac{\frac{1}{10}\frac{1}{2}}{\frac{4}{10}\frac{1}{2} + \frac{1}{10}\frac{1}{2}} = 0.34\end{aligned}$$

Similarly, an Eastern Correctional supporter expects others to support Harvard with probability 0.22. Assume all judges reveal their beliefs honestly. Since the average predictions for Harvard are 34% with three supporters, 30% with two, 26% with one, and 22% with zero, Harvard wins unless the judges unanimously support the inmates. Under this rule, the probability of Harvard winning conditional on being best is

$$\begin{aligned}\Pr[\text{Harvard support} > \text{predicted} \mid \omega = \text{Harvard}] &= \Pr[100\% \text{ support} > 34\% \text{ predicted} \mid \omega = \text{Harvard}] \\ &\quad + \Pr[66\% \text{ support} > 30\% \text{ predicted} \mid \omega = \text{Harvard}] \\ &\quad + \Pr[33\% \text{ support} > 26\% \text{ predicted} \mid \omega = \text{Harvard}] \\ &= \left(\frac{4}{10}\right)^3 + 3\left(\frac{4}{10}\right)^2 \frac{6}{10} + 3\frac{4}{10}\left(\frac{6}{10}\right)^2 = 0.78\end{aligned}$$

and 0.73 for the inmates. The prior probability of this peer-prediction rule making the correct decision is then

$$\begin{aligned}
&\Pr[\omega = \text{Harvard}] \cdot \Pr[\text{Harvard support} > \text{predicted} \mid \omega = \text{Harvard}] \\
&\quad + \Pr[\omega = \text{ENYCF}] \cdot \Pr[\text{ENYCF support} > \text{predicted} \mid \omega = \text{ENYCF}] \\
&= \frac{1}{2}0.78 + \frac{1}{2}0.73 = 0.75
\end{aligned}$$

compared to 0.66 for majority rule, producing more accurate decisions on average in addition to being fairer between states.

While promising, a problem remains with this particular peer-prediction rule: it's not incentive compatible if judges want the decision to match their own opinion. Incentive compatibility guarantees participants will report honestly even if they're willing to misreport. A judge would benefit from strategically claiming all others will have the opposite opinion. If a Harvard supporter predicts no other judges will favor Harvard, the average prediction will be lower, making it possible to secure a win with fewer supporters. Analogously, an inmate supporter maximizes the chances of an inmate win by predicting unanimous support of Harvard from the others. Under this strategy, the averaged predictions of how many other judges favor Harvard are:

Harvard supporters	Inmate supporters	Average predictions	Support for Harvard
0	3	100%	0%
1	2	66%	33%
2	1	33%	66%
3	0	0%	100%

The percentage of Harvard supporters is greater than the average "prediction"(i.e. the condition for a Harvard win under this proposed peer-prediction rule) if and only if a majority favors Harvard. Because of strategic reporting, the outcome becomes identical to majority rule, and the potential benefits of using peer-prediction evaporate.

In this paper, I investigate whether incentive-compatible peer-prediction decision rules exist that are more accurate than majority rule. I require candidate peer-prediction rules to be *neutral*—symmetric between the two choices—and *anonymous*—symmetric between group members—like majority rule.

Different types of incentive compatibility constraints provide different guarantees for when a participant will truthfully reveal their information. Bayesian incentive compatibility is a standard requirement but is acknowledged to carry strong assumptions. A more robust alternative is dominant-strategy incentive compatibility. Majority rule, for instance, is dominant-strategy incentive compatible because voting for one team always makes it more likely they'll win. However, requring dominant-strategy incentive compatibility makes it impossible for the decision to incorporate predictions. Instead, I rely on an intermediate form of incentive compatibility based on iterated deletion of weakly interim-dominated strategies. A decision rule is *robustly implementable* if honest relevation survives this process of iterated deletion.

The paper proceeds as follows. Section 2 provides related literature. Section 3 describes the model and design objective. In section 4, I show predictions can play almost no role in deterministic, neutral, anonymous, and robustly implementable decision rules. As long as every agent thinks it's possible another agent holds the opposite opinion, the decision matches majority rule. Section 5 provides a characterization of randomized, neutral, anonymous, and robustly implementable decision rules in terms of a common functional form that varies only with the choice of two non-decreasing functions and two real numbers. In section 6, I numerically search for the optimal randomized mechanism using the analytical characterization of the previous section. Although randomized decision rules can non-trivially depend on agents' predictions, majority rule outperforms all rules that incorporate predictions. Despite the promise of peer-prediction rules for identifying the true state more frequently, these results show majority rule can't be beaten subject to incentive constraints.

However, since it is plausible some agents are willing to give sincere predictions, section 7 considers non-incentive-compatible rules that make more accurate decisions than majority rule when some agents are unconditionally honest and become equivalent to majority rule when all agents are strategic. For instance, one simple rule based on a weighted combination of the percentage in support and the median prediction makes 25-50% fewer mistakes than majority rule when half of the participants report honestly and half report strategically. Finally, section 8 concludes.

2 Related Literature

Extensive work has been done to answer when groups can make correct decisions through voting procedures and when information can be elicited from strategic agents. Research on the accuracy of collective decisions dates to the Marquis de Condorcet's essay on majority rule. Condorcet's jury theorem now has many different forms (Grofman et al. 1983). In its standard version, it says majority rule is almost certain to choose the correct state as the number of agents voting grows large. Furthermore, simple majority rule is the optimal decision rule when each state has equal prior probability and agent's opinions are distributed identically and independently conditional on the state (Nitzan and Paroush 1982).

Across various extensions, the critical assumptions of the Condorcet jury theorem are that the average voter is more likely to favor the correct state than not and preferences do not change conditional on being the pivotal voter. Austen-Smith and Banks (1996) reconsider the second assumption, showing that sincere voting is typically not equilibrium behavior when agents have aligned preferences for the decision to match the true state. Following work on strategic voting has primarily focused on comparing particular voting rules, often reaching the conclusion that requiring unanimity is worse than simple majority or any supermajority (Feddersen and Pesendorfer 1998; Gerardi 2000; Duggan and Martinelli 2001). In this paper, I take a mechanism design approach

to address violations of the first assumption while retaining the second.

A parallel line of research on eliciting information from strategic agents with differing preferences was initiated by Crawford and Sobel (1982). Many papers have considered elicitation from groups of experts, including Austen-Smith (1993); Feddersen and Pesendorfer (1997); Krishna and Morgan (2001); Battaglini (2004). Of particular relevance, Li et al. (2014); Wolinsky (2002); Glazer and Rubinstein (2004); Gerardi et al. (2009); Chwe (2010) take a mechanism design approach. Each of these considers implementation in Bayes-Nash equilibrium in contrast to my approach based on interim dominance-solvability and a lack of common knowledge about preferences or the information structure.

Peer-prediction mechanisms have been studied in the context of eliciting correlated private signals from groups of payment maximizers without preferences over the conclusions drawn from the collected information. Prelec (2004)'s *Bayesian truth serum* elicits signals in Bayes-Nash equilibrium even when the principal has no knowledge of the common prior or signal likelihoods, though the result holds only for a sufficiently large number of participants that depends on the unknown prior. Witkowski and Parkes (2012a) construct a variant of Prelec's mechanism that is incentive compatible for finite participants in the case of binary questions. Zhang and Chen (2014) and Riley (2014) provide detail-free mechanisms that are Bayesian incentive compatible for finite participants and any number of signals with arbitrary correlation structure. To my knowledge, this is the first paper to consider a peer-prediction mechanism without transfers.

The Bayesian truth serum scores also function as an anonymous and neutral decision rule that asymptotically chooses the correct state when agents are Bayesians with conditionally IID signals and a common prior, even in the presence of statistical bias (Prelec et al. 2014). However, this decision rule is not incentive compatible if agents have preferences over the result. Since the mechanism chooses the answer with the highest average score and scores can be unboundedly negative, a single agent can unilaterally force one answer off the table even if all others are honest. Although I consider non-incentive-compatible decision rules in this paper, my mechanisms dampen the influence of strategic behavior.

3 Model

A group of n agents face a decision between two choices A and B. The state $\omega \in \{A, B\}$ denotes the "correct" decision according to some standard, such as the most skilled of two competitors, the action that will maximize profits, or the true answer to a question. Where convenient, let the states have values $A = 1$ and $B = 0$. From the mechanism designer's perspective, the two states have equal prior probability.

Each individual i has an *opinion* $x_i \in \{a, b\}$ about the state and a *prediction* $p_i \in (0, 1)$ about the proportion of other agents who hold opinion a. In a slight abuse of notation, let x_i also be an indicator variable with values $x_i = 1$ if i holds the a opinion and $x_i = 0$

if i holds the b opinion. Let $n_a = \sum_i x_i$ be the number of participants stating opinion a, $n_b = n - n_a$ be the number of participants stating opinion b, and $\bar{x} = n_a/n$ be the proportion of respondents with opinion a. Let $\bar{x}_{-i} = \sum_{j \neq i} x_j/(n-1)$ be the proportion of agents other than i with opinion a.

Opinions are distributed independently conditional on the state with likelihoods $q_A = \Pr(x_i = a | \omega = A)$ and $q_B = \Pr(x_i = a | \omega = B)$. The likelihoods satisfy $q_A > q_B$, so opinions are positively correlated with the corresponding state but are otherwise unknown to the mechanism designer.

The prediction $p_i = \mathrm{E}_i[\bar{x}_{-i} \mid x_i]$ summarizes agent i's subjective beliefs about the opinions of others, and will be treated as a random variable distributed independently conditional on x_i from the perspective of the mechanism designer. Although I view agents symmetrically, the agents themselves can have arbitrary beliefs consistent with their predictions, seeing correlations between individuals or thinking particular agents are more likely to hold a position. For example, $p_i = 0.5$ is consistent with believing all other agents are equally and independently likely to hold either opinion, with others being perfectly correlated and equally likely to hold each opinion, or with half of the agents holding one opinion with certainty and half holding the other with certainty. I make no assumptions about higher-order beliefs.

Since agents can see correlations or distinctions between others, predictions aren't required to be consistent with Bayesian updating based on my specification. However, for predictions to retain some connection to the underlying state, I assume agents treat their opinions as IID signals on average, holding a "prior prediction" between the two likelihoods that is then updated upward upon observing $x_i = a$ or downward for $x_i = b$ plus some noise. In particular, I model predictions as normally distributed on a logistic scale:

$$\begin{aligned} &\ln\left(\tfrac{p_i}{1-p_i}\right) \sim \text{Normal}\left(\mu_{x_i}, \sigma^2\right) \quad \text{s.t.} \\ &\mu_a = \mu + \gamma, \quad \mu_b = \mu - \gamma \\ &\mu = \alpha \ln\left(\tfrac{q_A}{1-q_A}\right) + (1-\alpha)\ln\left(\tfrac{q_B}{1-q_B}\right) \end{aligned} \tag{1}$$

for some parameters $\alpha \in [0,1]$ and $\gamma \in \mathbb{R}_{++}$, which can be interpreted as the prior belief that $\omega = A$ and the amount of evidence participants consider their own opinion to be, respectively. The distribution of agent predictions comes into play when numerically evaluating the accuracy of mechanisms, so the choice of distribution can change the level of performance but doesn't substantively affect results.

3.1 Peer-prediction decision rules

A peer-prediction decision rule T for n agents takes opinions and predictions as inputs to produce a choice between the two states. Decision rules can be deterministic or randomized. A deterministic decision rule has output $T(x,p) \in \{A, B, \varnothing\}$, where $\varnothing$ is a "null

choice" that can be used in situations with exact ties. A randomized decision rule has output $T(x,p) \in [0,1]$ denoting the probability A is chosen.

I focus on neutral and anonymous decision rules, retaining the properties of majority rule that no bias is built in towards either state or the opinion of any individual:

Definition 1 (Neutrality)**.** *A mechanism $T(x,p)$ is neutral if relabeling states A and B results in the complement of T, i.e. $T(x,p) = 1 - T(1-x, 1-p)$ for all x and p.*

Definition 2 (Anonymity)**.** *A mechanism $T(x,p)$ is anonymous if relabeling agents does not change T, i.e. $T(x,p) = T(\sigma(x), \sigma(p))$ for all permutations σ.*

The mechanism designer's objective is to maximize the probability the decision matches the true state:

$$\max_{T} \Pr[T(x,p) = \omega] \tag{2}$$

or equivalently

$$\min_{T} \mathrm{E}[\,|\omega - T(x,p)|\,]. \tag{3}$$

Each agent prefers the decision to match their own opinion. In particular, an agent with $x_i = a$ chooses a report (x'_i, p'_i) to solve

$$\max_{(x'_i, p'_i)} \Pr[T((x'_i, x'_{-i}), (p'_i, p'_{-i})) = A] \tag{4}$$

based on their conjecture about the reports (x'_{-i}, p'_{-i}) of others. Agents with $x_i = b$ then minimize the above objective.

3.2 Robustly implementable mechanisms

Mechanism design involves finding a procedure for collecting messages from agents and aggregating the reports into the desired outcome for each type profile while respecting the incentives of each participant. In general, a mechanism $\mathscr{M} = (M, g)$ consists of a space of message profiles M and an outcome function $g : M \to A$, where A is the set of possible outcomes. A mechanism implements T when the outcome of the induced game under some solution concept matches T.

Peer-prediction mechanisms have a message space where agents report an opinion and a probability distribution over the opinions of others. A peer-prediction mechanism can be seen as a "semi-direct" mechanism, asking agents to report a portion of their type rather than their full type, including a hierarchy of higher-order beliefs. The Bayesian truth serum (Prelec 2004) is a leading example of a peer-prediction mechanism. This mechanism has truth-telling as a Bayes-Nash equilibrium for sufficiently large groups of payment maximizers with an unknown common prior. The average difference in group scores can distinguish the true answer asymptotically, even with in the presence of false

consensus (Prelec and Seung 2007). However, existing peer-prediction mechanisms assume agents care only about payments, not about influence. If agents have preferences over the aggregate score used to estimate the state, the Bayesian truth serum becomes highly manipulable.

Additionally, existing peer-prediction mechanisms depend on agents sharing a common prior [2]. While common priors are often singled-out as unrealistic, a possibly more concerning feature is that agents receive a single signal with agreed-upon conditional likelihoods. Realistically, each expert has seen evidence of various levels of strength that he may or may not have updated on properly, which points toward some form of robust implementation beyond Bayes-Nash equilibrium.

Standard notions of robust implementation include implementation in dominant-strategy or ex-post Nash equilibrium. However, any mechanism that makes one strategy a best response regardless of the types of others can't be sensitive to predictions in equilibrium. Two agents with the same opinion and different predictions have the same preferences ex-post, so the same outcome will be assigned when the two behave identically. Independence from higher-order beliefs is usually seen as a benefit, but comes at the cost of ruling out peer-prediction mechanisms before we even begin. Instead, I'll consider a peer-prediction mechanism to be robustly implementable if honest reporting is the dominance solvable outcome of the mechanism, surviving iterated deletion of weakly interim dominated strategies. Throughout the paper, I will rely on only two stages of strategy deletion.

Definition 3 (Weak interim dominance)**.** *A strategy m_i weakly interim dominates m_i' for an agent of type (x_i, p_i) if*

$$\int\int u_i(x_i, g(m_i, m_{-i}))\, d\phi(m_{-i} \mid x_{-i}, p_{-i})\, d\pi(x_{-i}, p_{-i}) \geq \tag{5}$$

$$\int\int u_i(x_i, g(m_i', m_{-i}))\, d\phi(m_{-i} \mid x_{-i}, p_{-i})\, d\pi(x_{-i}, p_{-i}) \tag{6}$$

for all beliefs π (a distribution over type profiles of others) and ϕ (a distribution over strategy profiles conditional on type profiles) such that $\mathrm{E}_\pi[\bar{x}_{-i}] = p_i$ to be consistent with i's type, with strict inequality for some beliefs.

Definition 4 (Dominance solvability)**.** *Given a mechanism $\mathcal{M} = (M, g)$, let $D_i^{\mathcal{M}}(x_i, p_i)$ be the set of strategies m_i that survive iterated deletion of all weakly interim dominated strategies at each stage for agent i of type (x_i, p_i). A mechanism is interim dominance solvable if $g(m) = g(m')$ for all profiles with $m_i, m_i' \in D_i^{\mathcal{M}}(x_i, p_i)$.*

Definition 5 (Robust implementation)**.** *A mechanism $\mathcal{M} = (M, g)$ robustly implements a peer-prediction mechanism T if the unique dominance solvable outcome when agents have types (x, p) is $T(x, p)$.*

[2] Unless priors and posteriors can be elicited separately, as in Witkowski and Parkes (2012b).

Definition 6 (Robust incentive compatibility)**.** *A peer-prediction mechanism $T(x,p)$ is robustly incentive compatible if honesty is an interim best response for all conjectures about others' types consistent with the agent's prediction:*

$$\begin{aligned}\int T((a,x_{-i}),(p_i,p_{-i}))\,d\pi(x_{-i},p_{-i}) &\geq \int T((x_i',x_{-i}),(p_i',p_{-i}))\,d\pi(x_{-i},p_{-i}) \\ &\geq \int T((b,x_{-i}),(p_i,p_{-i}))\,d\pi(x_{-i},p_{-i})\end{aligned} \tag{7}$$

for all x_i', p_i, p_i', and beliefs π such that

$$\mathrm{E}_\pi[\bar{x}_{-i}] = \int \frac{\#\{x_j = a \mid j \neq i\}}{n-1}\,d\pi(x_{-i},p_{-i}) = p_i.$$

The following proposition provides a version of the revelation principle for this setting:

Proposition 1. *A mechanism $\mathscr{M} = (M,g)$ can robustly implement T only if T is robustly incentive compatible.*

Proof of Proposition 1 (Robust incentive compatibility). Suppose mechanism $\mathscr{M} = (M,g)$ robustly implements T, assigning outcome $g(m) = T(x,p)$ for each strategy profile $m \in \prod_i D_i^{\mathscr{M}}(x_i,p_i)$. Hence, given any $m_i \in D_i^{\mathscr{M}}(a,p_i)$ and $m_i' \in D_j^{\mathscr{M}}(x_j',p_j')$, we must have

$$\begin{aligned}\int T((a,x_{-i}),(p_i,p_{-i}))\,d\pi(x_{-i},p_{-i}) &= \int\int g(m_i,m_{-i})\,d\phi(m_{-i}\mid x_{-i},p_{-i})\,d\pi(x_{-i},p_{-i}) \\ &\geq \int\int g(m_i',m_{-i})\,d\phi(m_{-i}\mid x_{-i},p_{-i})\,d\pi(x_{-i},p_{-i}) \\ &= \int T((x_i',x_{-i}),(p_i',p_{-i}))\,d\pi(x_{-i},p_{-i})\end{aligned}$$

for all beliefs π (a distribution over type profiles of others) and ϕ (a distribution over strategy profiles conditional on type profiles) such that

$$\mathrm{E}_\pi[\bar{x}_{-i}] = p_i \quad \text{and}$$
$$\Pr_\phi[m_{-i}\mid x_{-i},p_{-i}] > 0 \implies m_{-i} \in \prod_{j\neq -i} D^{\mathscr{M}}(x_j,p_j)$$

since m_i either weakly dominates m_i' or is equivalent to it when agent i is type (a,p_i) and other agents play their dominance solvable strategies. This follows similarly for types (x_i',p_i') and (b,p_i), yielding the condition of robust incentive compatibility in line 7. □

Although the revelation principle provides some justification for restricting attention to incentive-compatible mechanisms, I will explore non-incentive-compatible decision rules that implement majority rule when all agents are strategic and outperform majority rule when some agents are honest later in the paper.

4 Deterministic decision rules

Consider decision rules which deterministically output a single state for any given profile. Even if first-order beliefs are included in reports, these turn out to play no functional role in the mechanism since the output is too coarse to respond to predictions. A robustly implementable, neutral, and anonymous decision rule can deviate from majority rule only when some agent mistakenly claims the realized profile was impossible:

Proposition 2. *If $T : \{a,b\}^n \times [0,1]^n \to \{A,B,\varnothing\}$ is a neutral, anonymous, and robustly implementable decision rule with $T(x) = \varnothing$ only if $\bar{x} = \frac{1}{2}$, then it agrees with majority rule on all profiles with interior predictions $p \in (0,1)^n$.*

The proof proceeds by showing profiles where agents correctly predict a bare majority must agree with majority rule and then expanding the set of profiles in agreement via incentive compatibility.

For an example of a deterministic decision rule where predictions do matter, consider a rule for three agents that maps all type profiles to the majority opinion except for

$$\begin{aligned}
&T((a,0),(a,0),(a,0)) = B,\\
&T((a,0),(a,0),(b,p_3)) = B \quad \forall\, p_3 \in (0,1],\\
&T((b,1),(b,1),(b,1)) = A, \text{ and}\\
&T((b,1),(b,1),(a,p_3)) = A \quad \forall\, p_3 \in [0,1),
\end{aligned}$$

as well as similar profiles for anonymity. This rule is neutral, anonymous, and robustly incentive compatible, so agreement with majority rule isn't required to extend to all profiles with extreme beliefs.

If decisions between the two states are randomized, the probability of choosing the A state in a neutral, anonymous, and robustly implementable mechanism is characterized in the next section.

5 Randomized decision rules

Randomized decision rules map report profiles into probabilities. Unlike deterministic rules, this set of decision rules can non-trivially incorporate predictions since there is more fine-grained control over the output.

As shown in the following theorem, all neutral, anonymous, and robustly implementable randomized rules for given n have a specific functional form that differ only by reference types $\phi_1, \phi_2 \in [\frac{1}{2}, 1]$ and nondecreasing functions τ and ξ. In this characterization, T can be decomposed into a *base score* (line 8) that depends solely on the proportion of agents endorsing a. The base score is adjusted by the mean *prediction scores* (line 9) of each agent, signed according to their opinion. The base score provides

sufficient incentive for reports with a false opinion to be interim dominated. Conditioning on each player always wanting to honestly reveal their true opinions, agents will want to give their true prediction as long as their marginal influence is a proper scoring rule for the proportion of a endorsements. The parameters ϕ_1 and ϕ_2 describe prediction types where incentive constraints bind exactly. The function ξ weights prediction scores, controlling the magnitudes of rewards and punishments for prediction accuracy in each region of the unit interval.

This representation embeds the design constraints into the functional form, reducing the optimal mechanism design problem to a mildly-constrained search across ϕ_1, ϕ_2, τ, and ξ.

Proposition 3. *A neutral and anonymous peer-prediction randomized decision rule T is robustly implementable for n participants only if T can be represented as*

$$T(x,p) = \frac{1}{2} + \operatorname{sign}\left(\tfrac{n_a}{n} - \tfrac{1}{2}\right)\left(\tau\left(\left|\tfrac{n_a}{n} - \tfrac{1}{2}\right|\right) + \mathbb{1}(n \text{ odd})\frac{\delta\left(\frac{n-1}{2}\right)}{2n} + \frac{1}{n}\sum_{m=\lceil n/2\rceil}^{\max\{n_a,n_b\}-1}\delta(m)\right) \tag{8}$$

$$+ \frac{1}{n}\sum_{i:\,x_i=a} R_\xi\left(p_i, \tfrac{n_a-1}{n-1}\right) - \frac{1}{n}\sum_{i\,x_i=b} R_\xi\left(1-p_i, 1-\tfrac{n_a}{n-1}\right) \tag{9}$$

$$\text{s.t.}\quad \delta(m) = \max\left\{-R_\xi\left(\phi_1, \tfrac{m}{n-1}\right) - R_\xi\left(1, 1-\tfrac{m}{n-1}\right),\ -R_\xi\left(1-\phi_2, 1-\tfrac{m}{n-1}\right)\right\}$$

$$R_\xi(p_i, \bar{x}) = \int_0^{p_i} (\bar{x} - t)\, d\xi(t)$$

for $\phi_1, \phi_2 \in [\frac{1}{2}, 1]$ and non-decreasing functions $\xi : [0,1] \to \mathbb{R}_+$ and $\tau : [0, \frac{1}{2}] \to \mathbb{R}_+$. This representation is sufficient for robust implementation if τ is strictly increasing and the maximal output satisfies $T((a,1),\ldots,(a,1)) \leq 1$.

The requirement for sufficiency that τ be strictly increasing ensures incentives are strict, while the requirement that $T((a,1),\ldots,(a,1)) \leq 1$ ensures the output of T is always a proper probability contained in the unit interval.

6 Determining the optimal randomized mechanism

Using the representation stated in the previous section, I now investigate the optimal randomized decision rule. The mechanism design problem is to solve

$$\min_T \mathrm{E}[\,|\omega - T(x,p)|\,]$$
$$\text{s.t. } T \text{ is neutral, anonymous, and robustly incentive compatible,}$$

which is equivalent to

$$\min_{\phi_1,\phi_2,\tau,\xi} \mathrm{E}[\,|\omega - T(x,p)|\,] =$$
$$\Pr(\omega = A)\,\mathrm{E}[1 - T(x,p)\,|\,\omega = A] + \Pr(\omega = B)\,\mathrm{E}[T(x,p)\,|\,\omega = B]$$
$$\text{s.t. } \phi_1, \phi_2 \in [\tfrac{1}{2}, 1] \text{ and } \xi : [0,1] \to \mathbb{R}_+, \tau : [0, \tfrac{1}{2}] \to \mathbb{R}_+ \text{ are non-decreasing.}$$

Unfortunately, this problem isn't amenable to typical first-order solution methods. Corner solutions are likely since the objective function is linear in T and T is affine in τ and possibly ξ. When the objective is locally affine, first-order conditions at the boundary become trivial.

Since the conditional opinion likelihoods aren't known to the mechanism operator, a prior distribution over likelihoods must be specified. Some natural distributions of likelihoods include:

1. Uniform over all likelihood pairs (q_A, q_B) with positive correlation, satisfying $q_A > q_B$
2. Those concentrated around the diagonal or in a band offset from the diagonal
3. Uniform over all unbiased likelihood pairs, satisfying $q_B \leq 0.5 \leq q_A$
4. Uniform over all biased likelihood pairs, satisfying $q_B < q_A \leq 0.5$ or $0.5 \leq q_B < q_A$

The first and second possibility can be interpreted as each agent independently knowing the true state with probability λ and otherwise having opinion a with probability $(1-\lambda)\gamma$ and opinion b with probability $(1-\lambda)(1-\gamma)$, with both probabilities unknown. The first corresponds to a uniform prior over both λ and γ. The second corresponds to a normal distribution (restricted to the unit interval) over λ and a uniform distribution over γ, allowing for more precise information about the expertise of participants.

As noted earlier, I model predictions as normally distributed on a logistic scale for parameters $\alpha \in [0,1]$ and $\gamma \in \mathbb{R}_{++}$ corresponding to a prior prediction and an degree of adjustment, respectively. In particular, I assume $\alpha, \gamma \sim \text{Unif}[0,1]$. Then, taking expectations across parameters $\theta = (q_A, q_B, \alpha, \gamma) \in [0,1]^4$, the likelihood of types in a given state is

$$\begin{aligned} \mathrm{E}_\theta[\Pr(x,p \mid \omega, \theta)] &= \int_\theta q_\omega^{n_a}(1-q_\omega)^{n-n_a}\left(\prod_{i=1}^{n} f_{x_i}(p_i \mid \theta)\right) g(q_A, q_B)\, d\theta \qquad (10) \\ \text{s.t.} \quad f_{x_i}(p_i \mid \theta) &= \frac{1}{p_i(1-p_i)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{\sigma^2}\left(\mu_{x_i}(\theta) - \ln\left(\frac{p_i}{1-p_i}\right)\right)^2\right) \\ \mu_{x_i}(\theta) &= \alpha \ln\left(\frac{q_A}{1-q_A}\right) + (1-\alpha)\ln\left(\frac{q_B}{1-q_B}\right) + (2\mathbb{1}(x_i = a) - 1)\gamma. \end{aligned}$$

I set $\sigma^2 = 1$ to produce a realistic amount of dispersion without the distribution bunching around 0 and 1, which tends to occur when the variance grows larger. Figure 1 shows typical prediction distributions.

6.1 Representing decision rules numerically

As shown in Proposition 3, optimization over the class of robustly implementable mechanisms involves a search over three components: the scoring rule weighting function $\xi(t)$, the reference types ϕ_1, ϕ_2, and the extra base score $\tau(t)$. For a numerical solution,

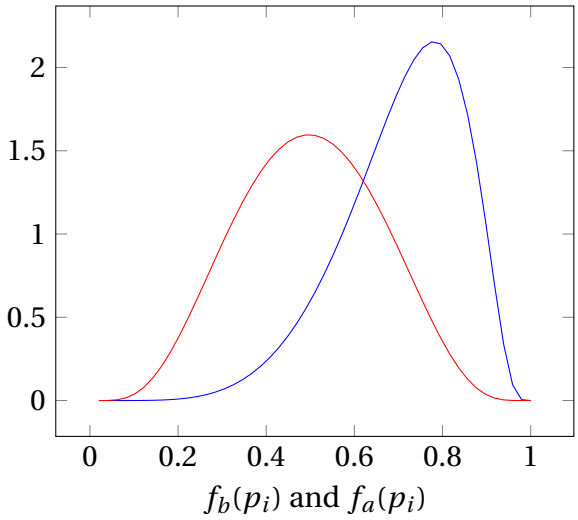

Figure 1: Prediction densities for agents with opinions a and b when $q_A = 0.8$, $q_B = 0.4$, $\sigma^2 = 1$, $\alpha = 0.5$, and $\gamma = 0.5$.

I approximate this infinite-dimensional problem with a finite-dimensional representation.

In full generality, the weighting function ξ used to parameterize the scoring rule $R_\xi(p_i, \bar{x}_{-i})$ can be any non-decreasing function with a domain of $[0,1]$. I approximate a general ξ by decomposing it into a continuously differentiable function and a step function, producing a scoring rule

$$R_\xi(p_i, \bar{x}_i) = \int_0^{p_i} (\bar{x}_{-i} - t)\xi'(t)\,dt + \sum_{k=1}^{K_\xi} \lambda_k \mathbb{1}(p_i \geq t_k)(\bar{x}_{-i} - t_k). \tag{11}$$

On the discrete portion, ξ has K_ξ atoms at points $t_k \in [0,1]$ with weights $\lambda_k \in \mathbb{R}_+$. On the continuously differentiable portion, I assume ξ' is piecewise linear with $H_\xi - 1$ segments at regular intervals, giving the integral a manageable closed form. The contribution to the total score on an interval $[t_1, t_2]$ where $\xi'(t)$ is linear is

$$\begin{aligned} &\int_{t_1}^{t_2} (\bar{x}_{-i} - t)\left(\tfrac{\xi'(t_2)-\xi'(t_1)}{t_2-t_1}(t - t_1) + \xi'(t_1)\right) dt = \\ &\qquad \frac{t_2 - t_1}{6}\big(3(\bar{x}_{-i} - t_1 - t_2)(\xi'(t_1) + \xi'(t_2)) - t_1\xi'(t_1) - t_2\xi'(t_2)\big). \end{aligned} \tag{12}$$

The score $R_\xi(p_i, \bar{x}_{-i})$ is the sum of this amount on each linear segment inside $[0, p_i]$, so ξ' can be parameterized by H_ξ values $\xi'_h \in \mathbb{R}_+$ at $0, 1/(H_\xi - 1), 2/(H_\xi - 1), \ldots, (H_\xi - 2)/(H_\xi - 1), 1$.

I also represent $\tau\left(\left|\frac{n_a}{n} - \frac{1}{2}\right|\right)$ using a continuous τ' with $H_\tau - 1$ linear segments and K_τ weighted atoms. Between densities parameters, atom locations, and atom weights for ξ and τ and the two reference types $\phi_1, \phi_2 \in [\frac{1}{2}, 1]$, the total parameter space is $H_\xi + 2K_\xi + H_\tau + 2K_\tau + 2$ dimensional.

6.2 Optimization methods

Using this finite-dimensional approximation, the scoring rule R_ξ is linear in the vectors of density values and atoms weights. The estimator $T(x,p)$ is then convex in these parameters when $\bar{x} > \frac{1}{2}$ and concave when $\bar{x} < \frac{1}{2}$ due to the changing sign on the maximum taken in $\delta(m)$. Since the estimator is neutral, we are always free to reassign labels to make a the majority opinion and $\bar{x} \geq 1/2$ so that the overall objective is convex in these parameters. The domain for each of these parameters is the entire positive real line, but since the objective diverges as any parameter diverges, the minimizer will be in some bounded interval.

The estimator is less well-behaved in terms of the reference types and atom position. A scoring rule is quasiconcave with ϕ as the prediction, and an atomic scoring rule $R(p_i, \bar{x}_{-i}) = \lambda_k \mathbb{1}(p_i \geq t_k)(\bar{x}_{-i} - t_k)$ is quasiconvex in t_k, but this won't necessarily aggregate up into quasiconvexity of the estimator or objective. The estimator is also discontinuous in these parameters, though the discontinuities will be smoothed out in expectation in the objective. Consequently, a global optimization procedure may be necessary to thoroughly search the parameter space.

The optimization problem is unconstrained aside from bounds on each parameter and possibly a constraint that the output is contained in the unit interval. For the output to be inside the unit interval, it is sufficient that the outcome when agents are unanimous and know they are unanimous satisfies $T((a,\ldots,a),(1,\ldots,1)) \leq 1$. Values outside the unit interval are nonsensical for randomized decision rules. Values outside the unit interval are still undesirable for an estimator but might be acceptable if they occur only for nearly unanimous inputs, which we expect to be rare. After all, there is little reason to conduct a survey if an answer is obvious and everyone thinks it's obvious.

Optimization is done through the *Multi-level Single-linkage* global optimization algorithm, a multistart method that uses a clustering heuristic to avoid repeatedly returning to the same local minima on each local optimization. For local optimizations, I used Rowan (1990)'s *Subplex* algorithm, a variant of the Nelder-Mead simplex method done through a sequence of subspaces.

6.3 Optimal randomized decision rules don't use predictions

Unlike deterministic decision rules, randomized decision rules are able to incorporate predictions while remaining robustly implementable. However, randomized output typically hurts when maximizing the probability of a correct decision or minimizing the absolute deviation, so it's unclear whether the potential benefit is worth the cost.

Optimization over the class of robustly implementable peer-prediction decision rules returns a mechanism that depends only on opinions, using a τ with a single step and zeroing out ξ. This finding holds varying n and the prior on opinion likelihoods. Note the optimal randomized mechanism isn't necessarily majority rule. For some priors on opinion likelihoods (such as a uniform prior over all biased likelihood pairs), the optimal mechanism chooses A when $\bar{x}$ is sufficiently high, B when $\bar{x}$ is sufficiently low, and

randomizes between them with equal probability when $\bar{x}$ is in an interval around $\frac{1}{2}$. If majority rule is the optimal randomized mechanism that uses only opinions, then it is also optimal in the class of peer-prediction mechanisms.

7 Simple peer-prediction rules with some sincere agents

The preceding results show majority rule is either the only robustly implementable decision rule or the only one worth considering, modified at most by randomizing in some interval around $\bar{x} = \frac{1}{2}$. Nevertheless, like most incentive-compatible direct mechanisms, the direct mechanism for majority rule takes strategic behavior for granted. Unlike an allocation setting, it is plausible that some agents are willing to unconditionally tell the truth and don't have preferences over the outcome. A non-incentive-compatible decision rule could implement majority rule when all agents are strategic and outperform it whenever some agents are sincere.

If all agents are Bayesians who think opinions are IID based on underlying likelihoods, all predictions will be inside the interval $[q_B, q_A]$. Without knowing the likelihoods themselves, a third party could easily conclude the state is likely to be A if the proportion of a opinions is higher than most predictions.

Although I allowed agent predictions as more dispersed, a similar identification of the state is possible in this setting. I model predictions as satisfying

$$\ln\left(\tfrac{q_B}{1-q_B}\right) < \mathrm{E}\left[\ln\left(\tfrac{p_i}{1-p_i}\right) \mid x_i = a\right] \quad \text{and} \quad \mathrm{E}\left[\ln\left(\tfrac{p_i}{1-p_i}\right) \mid x_i = a\right] < \ln\left(\tfrac{q_A}{1-q_A}\right)$$

which implies

$$q_B < \text{median}(p_i \mid x_i = a) \quad \text{and} \quad \text{median}(p_i \mid x_i = b) < q_A.$$

All else equal, we expect the state is more likely to be A when $\bar{x}$ is higher and when the proportion of a opinions is higher than the median group predictions, so one simple decision rule takes a linear combination of these magnitudes [3]:

$$T(x,p) = \mathbb{1}\left(\lambda_1(\bar{x} - \tfrac{1}{2}) + \lambda_2(\bar{x} - \text{median}(p_i | x_i = a)) + \lambda_3(\bar{x} - \text{median}(p_i | x_i = b)) > 0\right).$$

For neutrality, we must have $\lambda_2 = \lambda_3$. For some partial incentive compatibility, the expression should have $\lambda_1 + \lambda_2 + \lambda_3 > 0$ to be increasing in $\bar{x}$. Under these constraints, the decision rule above is equivalent to

$$T(x,p) = \mathbb{1}\left(\bar{x} + \tfrac{\lambda}{2}(1 - \text{median}(p_i | x_i = a) - \text{median}(p_i | x_i = b)) > \tfrac{1}{2}\right).$$

This decision rule has majority rule as the unique dominance solvable outcome. Assuming $\lambda > 0$, all reports for an agent with $x_i = a$ are weakly dominated by either $(a, 0)$ or

[3] To avoid taking the median of an empty group, assume the output matches the unanimous opinion if all agents agree.

$(b, 0)$, depending on which group median the agent has the most influence over. Once all reports with interior predictions are eliminated, reporting one's true opinion becomes the unique weakly dominant strategy for each agent. The group medians cancel out, leaving only a comparison of $\bar{x}$ to $\frac{1}{2}$.

When all agents are sincere, this decision rule does quite well. For instance, when $n = 50$ and there is a uniform prior over opinion likelihoods, this rule for $\lambda = 0.9$ misclassifies the state approximately 13.5% of the time compared to 25% of time for majority rule.

The median is well-known as a robust location estimator, able to withstand up to 50% of the inputs being adversarially altered before becoming invalid. Suppose each agent is strategic with identical probability ρ and sincere otherwise. Since there are two weakly dominant strategies, it's not obvious what an agent will do when it expects only some agent to be strategic. If all strategic agents report their true opinion and a prediction $p_i \in \{0, 1\}$, then the group medians quickly degrade to the extremes when $\rho > \frac{1}{2}$, reducing the decision to majority rule.

In contrast, consider the following even simpler decision rule:

$$T(x, p) = \mathbb{1}\left(\bar{x} + \lambda\left(\tfrac{1}{2} - \text{median}(p)\right) > \tfrac{1}{2}\right)$$

Call this the *median prediction rule*. Notice the decision is simply majority rule when $\lambda = 0$. When $\lambda > 0$, the unique weakly dominant strategy for the median prediction rule is for an agent with $x_i = a$ to report $(a, 0)$ and an agent with $x_i = b$ to report $(b, 1)$. Again assuming that agents have an IID chance of being strategic, the median of all predictions isn't influenced by strategic behavior until $\rho > 1 - \frac{\bar{x}}{2}$ since the inputs are being manipulated by two opposing groups of agents rather than a single-minded adversary.

Figure 2 depicts how the percentage of misclassified states for $n = 15$ and $n = 100$ varies for different weights λ in the decision rule. This is shown for varying percentages of strategic agents. The optimal weight λ depends on the number of strategic agents, starting around $\lambda = 0.7$–0.8 for completely honest agents and increasing as agents become more strategic. The plots show the median prediction rule being more accurate on average for every choice of λ (except $\lambda < 1/n$, which is too small to change the decision from majority rule).

Figure 3 depicts the percentage of misclassified states as the percentage of strategic agents varies, with $\lambda \simeq 0.8$ optimized for $\rho = 0$ and $\lambda \simeq 0.95$ optimized for $\rho = 0.5$. As agents become more strategic and ρ increases to one, the median prediction rule agrees with majority rule more and more frequently.

While there is little reason to think the median prediction rule is optimal, it is simple and robust. Adding predictions to the group decision reduces the errors of majority rule due to bias and, at worst, becomes equivalent to majority rule when agents act strategically. Majority rule is still a useful means of aggregating preferences, but whenever the underlying goal is to aggregate information and it's conceivable that the majority can make the wrong choice, the median prediction rule is a strong alternative.

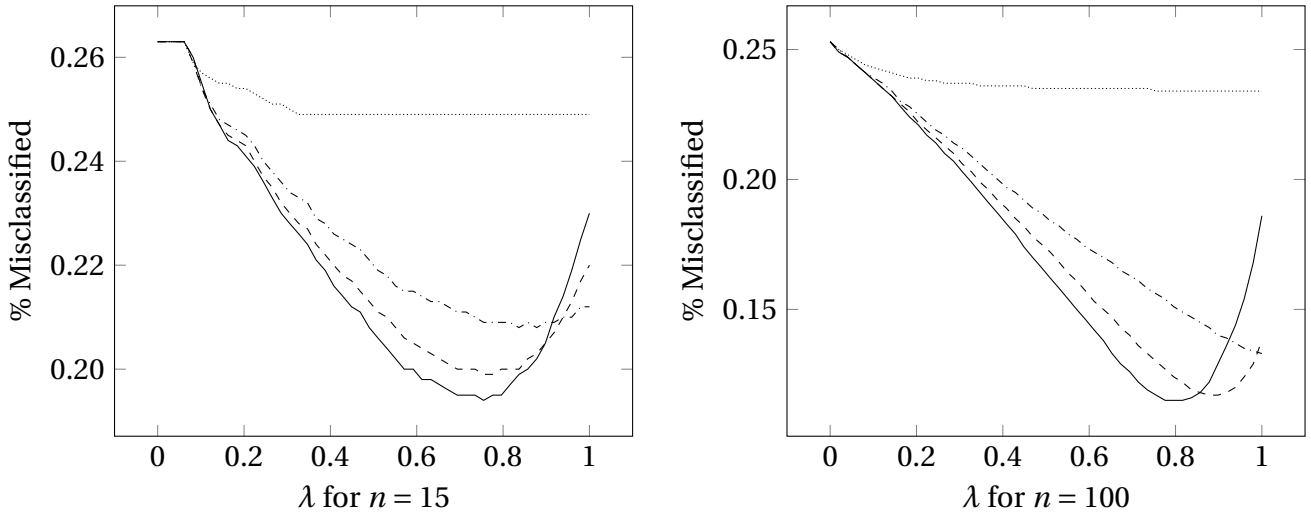

Figure 2: Effect of varying weight λ in the median prediction rule for percentage of strategic agents ρ in $\{0.0, 0.25, 0.5, 0.9\}$ in solid to dotted lines, respectively.

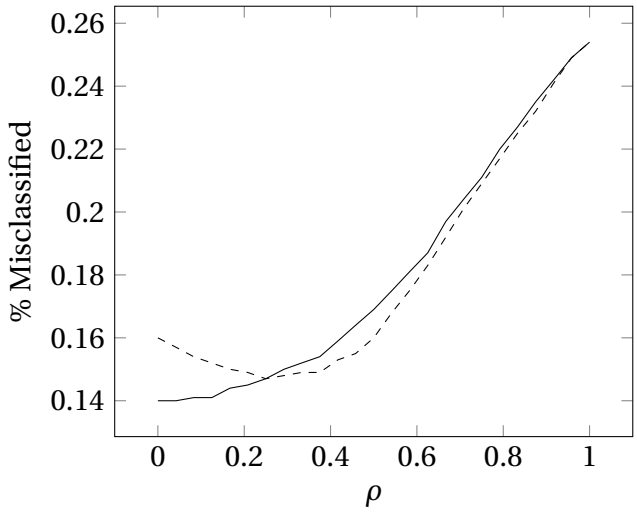

Figure 3: Percentage of incorrect decisions by the median prediction rule for $n = 50$ as the percentage of strategic agents ρ varies with $\lambda = 0.8$ solid and $\lambda = 0.95$ dashed.

8 Conclusion

My model takes a broad view of potential sources of bias, capturing two sources usually considered in isolation: preference-based bias and statistical bias. Agents with a preference-based bias have some stake in the conclusions drawn from their information. Agents are willing to distort or garble information themselves in order to influence the final results. This form of bias has been thoroughly investigated in the literature under the assumption of commonly known information and structure in order to facilitate analysis under Bayes-Nash equilibrium.

Alternatively, statistical bias is a systematic tendency for participants to hold a particular opinion besides the true state of the world, even if agents are sincere or have a common interest. Under this notion, opinions are biased in the sense that their likelihoods aren't symmetric across states of the world. In a mild form, 80% of the population might hold one opinion when it's correct, while only 60% of the population hold the opposite opinion when it's correct. In an extreme form, an opinion might be held by 90% of the population when it's correct and 75% of the population when it's incorrect, putting it in the majority regardless of the true state.

These two categories of bias are logically separate but are not easily separated in practice. Psychological notions of cognitive bias—defined as systematic deviations from some standard of judgment—can often be interpreted as a preference, and common usage conflates the two. The success of peer-prediction mechanisms can be seen as exploiting the false consensus effect identified in social and cognitive psychology (Marks and Miller 1987). Debate exists whether the false consensus effect is a cognitive bias or the rational consequence of updating beliefs about others conditional on one's own attributes (Dawes 1989), but either is compatible with my model.

I assume agents' preferences over conclusions do not change conditional on the reports of others, in contrast to the strategic voting literature where preferences can change dramatically after updating on others' information. Reality is somewhere in between, with people updating on the information of others at a discount relative to their own information (Yaniv and Kleinberger 2000). In the canonical example of a jury voting whether to convict a defendant, it's very plausible an agent would revise their opinion upon learning others are unanimous since a juror (ideally) doesn't have a personal connection to the question of guilt. Experimental work by Guarnaschelli et al. (2000) roughly supports the strategic voting model of Feddersen and Pesendorfer (1998), though the experiment sensibly asked participants to make the bland decision of which jar they drew colored balls from. In a more emotionally-charged situation like a committee decision wrapped up in office politics, I expect agents to stick to their opinions regardless of how others might report.

Possible future directions include experimental work testing the accuracy of these mechanisms, expanding the scope of the model beyond binary questions, and characterizing the equivalence class of peer-prediction decision rules that implement majority rule in a way amenable to optimizing accuracy in partially-strategic "equilibrium."

9 Computational Details

The numerical results were computed in *Julia, v0.4.0* using the MLSL and SBPLX optimization algorithms of *NLopt.jl* and the h-adaptive integration algorithm of *Cubature.jl*, both implemented by Steven G. Johnson.

Appendix

Proof of Proposition 2. Let T by any deterministic, neutral, anonymous, and robustly implementable decision rule. At least one such decision rule exists since majority rule satisfies these properties.

Suppose n is odd. I will establish the following facts about T in turn:

1. $T\big(\underbrace{(a,\tfrac{1}{2}),\ldots,(a,\tfrac{1}{2})}_{\frac{n+1}{2}},(b,p_{\frac{n+3}{2}}),\ldots,(b,p_n)\big)=A,\quad \forall p_{\frac{n+3}{2}},\ldots,p_n\in[0,1]$
2. $T\big((a,1),\ldots,(a,1)\big)=A$
3. $T\big((a,p_1),(b,\frac{1}{n-1}),\ldots,(b,\frac{1}{n-1})\big)=B,\quad \forall p_1\in[0,1]$
4. $T\big((a,p_1),(a,1),\ldots,(a,1)\big)=A,\quad \forall p_1\in(0,1]$
5. $T\big(\underbrace{(a,p_1),\ldots,(a,p_{\frac{n+1}{2}})}_{\frac{n+1}{2}},(b,p_{\frac{n+3}{2}}),\ldots,(b,p_n)\big)=A,\quad \forall p_1,\ldots,p_n\in(0,1)$
6. $T\big((a,p_1),\ldots,(a,p_m),(b,p_{m+1}),\ldots,(b,p_n)\big)=A,\quad \forall p_1,\ldots,p_n\in(0,1),\forall m\geq\frac{n+1}{2}$

The first three facts say that majorities of various sizes map to the majority opinion when those supporters have correct beliefs. The fourth says that one member of a full majority can have an arbitrary prediction without disturbing the outcome. The fifth says that all members of a bare majority can have arbitrary interior beliefs without changing the outcome. Finally, the sixth is the conclusion of the theorem.

I prove the first fact by contradiction. Suppose there are some predictions $p'_{\frac{n+3}{2}},\ldots,p'_n$ such that

$$T\big(\underbrace{(a,\tfrac{1}{2}),\ldots,(a,\tfrac{1}{2})}_{\frac{n+1}{2}},(b,p'_{\frac{n+3}{2}}),\ldots,(b,p'_n)\big)=B.$$

We must then also have

$$T\big((a,p_1),\underbrace{(a,\tfrac{1}{2}),\ldots,(a,\tfrac{1}{2})}_{\frac{n-1}{2}},(b,p'_{\frac{n+3}{2}}),\ldots,(b,p'_n)\big)=B,\quad \forall p_1\in[0,1]$$

for agent one with type $(a,\frac{1}{2})$ to report truthfully, since the agent could be certain this profile will occur (consistent with the prediction of $p_1=\frac{1}{2}$ that half of the other agents have opinion a) and thus can't expect to switch the outcome to A by reporting some other prediction. In particular,

$$T\big((a,\tfrac{n-3}{2(n-1)}),\underbrace{(a,\tfrac{1}{2}),\ldots,(a,\tfrac{1}{2})}_{\frac{n-1}{2}},(b,p'_{\frac{n+3}{2}}),\ldots,(b,p'_n)\big)=B.$$

Successively applying the same reasoning to all agents with opinion a yields

$$T\Big(\underbrace{(a,\tfrac{n-3}{2(n-1)}),\ldots,(a,\tfrac{n-3}{2(n-1))})}_{\frac{n+1}{2}},(b,p'_{\frac{n+3}{2}}),\ldots,(b,p'_n)\Big)=B.$$

For an agent with type $(b,\frac{1}{2})$ to report truthfully, we must have

$$T\Big(\underbrace{(a,\tfrac{n-3}{2(n-1)}),\ldots,(a,\tfrac{n-3}{2(n-1))})}_{\frac{n-1}{2}},(b,\tfrac{1}{2}),(b,p'_{\frac{n+3}{2}}),\ldots,(b,p'_n)\Big)=B$$

and then

$$T\Big(\underbrace{(a,\tfrac{n-3}{2(n-1)}),\ldots,(a,\tfrac{n-3}{2(n-1)})}_{\frac{n-1}{2}},\underbrace{(b,\tfrac{1}{2}),\ldots,(b,\tfrac{1}{2})}_{\frac{n+1}{2}}\Big)=B$$

by successively applying incentive compatibility for the remaining agents with opinion b. Applying neutrality and anonymity yields

$$T\Big(\underbrace{(a,\tfrac{1}{2}),\ldots,(a,\tfrac{1}{2})}_{\frac{n+1}{2}},\underbrace{(b,\tfrac{n+1}{2(n-1)}),\ldots,(b,\tfrac{n+1}{2n-2})}_{\frac{n-1}{2}}\Big)=A.$$

For the agents with type $(b,\frac{n+1}{2(n-1)})$ who think the previous profile is certain (consistent with their prediction) to report truthfully, the outcome cannot switch to B for any other prediction report. Changing the predictions of agents with opinion b in turn yields

$$T\Big(\underbrace{(a,\tfrac{1}{2}),\ldots,(a,\tfrac{1}{2})}_{\frac{n+1}{2}},(b,p'_{\frac{n+3}{2}}),\ldots,(b,p'_n)\Big)=A,$$

which is the original profile assumed to map to B, resulting in a contradiction.

The second fact follows from the first. Changing an agent from opinion b and an arbitrary prediction to opinion a and an accurate prediction for that profile must leave the outcome unchanged at A for incentive compatibility. This can be repeated until all agents have opinion a. Notice that as the types of other agents change, what was once an accurate prediction might become inaccurate. Updating the prediction of an agent with opinion a to be accurate for the profile must also leave the outcome unchanged, so the prediction for each can be changed to $p_i=1$, resulting in $T\big((a,1),\ldots,(a,1)\big)=A$.

The third fact follows from the first similarly to the second. All but one agent with opinion b can be replaced by an agent with opinion a and an accurate opinion for that profile. By anonymity, this yields

$$T\big((b,p_1),(a,\tfrac{n-2}{n-1}),\ldots,(a,\tfrac{n-2}{n-1})\big)=A,\quad\forall p_1\in[0,1]$$

and finally by neutrality,

$$T\big((a,p_1),(b,\tfrac{1}{n-1}),\ldots,(b,\tfrac{1}{n-1})\big)=B,\quad\forall p_1\in[0,1].$$

To establish the fourth fact, suppose agent one has type (a, p_1) with $p_1 \in (0,1]$ based on a belief that all other agents share type $(a,1)$ with probability p_1 and type $(b, \frac{1}{n-1})$ with probability $1-p_1$. If $T((a,p_1),(a,1),\ldots,(a,1)) = B$, then agent one expects the outcome from reporting truthfully to always be B by fact 3. If the agent misreported as type $(a,1)$, then the outcome would occasionally be A, producing a strictly better deviation. Therefore, we must have

$$T\big((a,p_1),(a,1),\ldots,(a,1)\big) = A, \quad \forall p_1 \in (0,1]$$

The fifth fact follows similarly to the fourth. Suppose agent one has type (a,p_1). If $p_1 \in (0,\frac{1}{2})$, consider an agent who is sure either fact 1 or 3 would apply if he reported $(a,\frac{1}{2})$. Since the choice of prediction won't change the outcome when fact 3 applies, the outcome for being honest must match the outcome when fact 1 would apply. Alternatively, if $p \in (\frac{1}{2},1)$, consider an agent who is certain either fact 1 or 4 would apply when reporting $(a,\frac{1}{2})$. Since this report always results in outcome A, the honest report must also result in A for the same profile of others. Taking these observations together with fact 1, we have

$$T\big((a,p_1),\underbrace{(a,\tfrac{1}{2})\ldots,(a,\tfrac{1}{2})}_{\frac{n-1}{2}},(b,p_{\frac{n+3}{2}}),\ldots,(b,p_n)\big) = A, \quad \forall p_1, p_{\frac{n+3}{2}}\ldots,p_n \in (0,1)$$

Repeating this reasoning for the remaining agents with opinion a yields

$$T\big(\underbrace{(a,p_1),\ldots,(a,p_{\frac{n+1}{2}})}_{\frac{n+1}{2}},(b,p_{\frac{n+3}{2}}),\ldots,(b,p_n)\big) = A, \quad \forall p_1,\ldots,p_n \in (0,1)$$

For the sixth fact, notice that replacing an agent with opinion b with an agent type $(a,\frac{n+1}{2(n-1)})$ in fact 5 must preserve the outcome of A. Applying the same argument as in the proof of fact 5 says any prediction $p_i \in (0,1)$, not just $\frac{n+1}{2(n-1)}$, must produce an outcome of A. This process can be repeated, adding further a supporters inductively. Therefore, any number of agents with opinion a and interior beliefs can be added, resulting in

$$T\big((a,p_1),\ldots,(a,p_m),(b,p_{m+1}),\ldots,(b,p_n)\big) = A, \quad \forall p_1,\ldots,p_n \in (0,1), \forall m \geq \frac{n+1}{2},$$

which concludes the proof that any neutral, anonymous, and robustly incentive compatible decision rule must be equivalent to majority rule when agents have interior predictions for odd n.

For even n, the first step is to establish that

$$T\big(\underbrace{(a,\tfrac{n}{2n-2}),\ldots,(a,\tfrac{n}{2n-2})}_{\frac{n}{2}+1},(b,p_{\frac{n}{2}+2}),\ldots,(b,p_n)\big) = A, \quad \forall p_{\frac{n}{2}+2},\ldots,p_n \in [0,1]$$

analogously to the first fact when n is odd. From this, the remaining facts follow, concluding with agreement with majority rule for all interior predictions when a majority exists. Furthermore,

$$T\big(\underbrace{(a,\tfrac{n-2}{2n-2}),\ldots,(a,\tfrac{n-2}{2n-2})}_{\frac{n}{2}},\underbrace{(b,\tfrac{n}{2n-2}),\ldots,(b,\tfrac{n}{2n-2})}_{\frac{n}{2}}\big) = \varnothing$$

for neutrality because this profile with correct predictions is complementary to itself. Changing the prediction of an agent with opinion a can't switch the outcome to A without giving an agent in this profile an incentive to misreport. The outcome also cannot switch to B without making a report of $(a, \frac{n-2}{2n-2})$ dominate an honest report of (a, p_i) for an agent that puts positive probability on this profile since the prediction doesn't matter for any non-balanced profile. By induction, all profiles with $\bar{x} = \frac{1}{2}$ and interior predictions must have $T(x, p) = \varnothing$ in agreement with majority rule.

□

Necessity of Proposition 3. Suppose agent i believes p_{-i} is fixed conditional on x_{-i}, reducing beliefs over the types of others to $\pi(x_{-i})$. Incentive compatibility implies

$$\sum_{x_{-i}} \pi(x_{-i})\, T((a, x_{-i}), (p_i, p_{-i})) \geq \sum_{x_{-i}} \pi(x_{-i})\, T((a, x_{-i}), (p_i', p_{-i}))$$

for all p_i, p_i', p_{-i}, and π such that $\mathrm{E}_\pi[\bar{x}_{-i}] = p_i$, so that agent i does not want to misreport her prediction p_i. Hence, T is a proper scoring rule for the mean of x_{-i} from the perspective of agent i holding $x_i = a$ fixed. By the McCarthy-Savage representation of proper scoring rules, T must be representable from the perspective of agent i as

$$T((a, x_{-i}), p) = \kappa_i(x, p_{-i}) + G_i(p_i; p_{-i}) + (\bar{x}_{-i} - p_i) G_i'(p_i; p_{-i}) \tag{13}$$

using some G_i convex in p_i, where G_i' is a subderivative in p_i. Without loss of generality, we can suppose $G_i(0; p_{-i}) = 0$ and $G_i'(0; p_{-i}) = 0$ by folding $G_i(0; p_{-i})$ and $\bar{x}_{-i} G_i'(0; p_{-i})$ into κ_i if necessary. Since $G_i'(p_i; p_{-i})$ must be non-decreasing as a subderivative of a convex function, it has bounded variation on $[0, 1]$ and we are free to write it as a Lebesgue-Stieltjes integral:

$$G_i'(p_i; p_{-i}) = \int_0^{p_i} d\xi_i(t; p_{-i}). \tag{14}$$

Then, we have

$$G_i(p_i; p_{-i}) = \int_0^{p_i} \int_0^t d\xi_i(s; p_{-i})\, dt = \int_0^{p_i} (p_i - t)\, d\xi_i(t; p_{-i}) \tag{15}$$

after a change of variables. Plugging the last two lines into line 13 yields

$$T((a, x_{-i}), p) = \kappa_i(x, p_{-i}) + \int_0^{p_i} (\bar{x}_{-i} - t)\, d\xi_i(t; p_{-i}), \tag{16}$$

which is closely related to the Schervish (1989) representation (see also Lambert (2011)). This representation prescribes the specific way that p_i and the proportion $\bar{x}_{-i}$ must interact for incentive compatibility, up to a weighting by ξ_i. For T to be neutral between A and B, we must have

$$T((b, x_{-i}), p) = \kappa_i(x, p_{-i}) - \int_0^{1-p_i} (1 - \bar{x}_{-i} - t)\, d\xi_i(t; 1 - p_{-i})$$

so $T(x,p)+T(1-x,1-p)=1$. With this form for each agent, it follows by anonymity that

$$T(x,p)=\kappa(\bar{x})+\sum_{i:\,x_i=a}\int_0^{p_i}(\bar{x}_{-i}-t)\,d\xi(t)-\sum_{i:\,x_i=b}\int_0^{1-p_i}(1-\bar{x}_{-i}-t)\,d\xi(t),$$

since $\bar{x}$ contains all information preserved under permutations of x and ξ can't depend on the identity of the agent. Although ξ_i could have depended on the predictions of other agents to be a proper scoring rule for agent i, those predictions can only appear in their respective integrals to be proper for the remaining agents.

Again taking p_{-i} to be known conditional on x_{-i}, incentive compatibility implies T is higher in expectation when agent i reports her true type (a,p_i) than when reporting any (b,p_i'). Since the mechanism is anonymous, an agent's beliefs can be reduced to a distribution over the number of other agents $m=\sum_{j\neq -i}x_j$ with the a opinion rather than on x_{-i} directly, even if the underlying belief treats other agents asymmetrically. We have

$$\begin{aligned}
&\sum_{m=0}^{n-1}\pi(m)\Bigg(\kappa\left(\tfrac{m+1}{n}\right)+\int_0^{p_i}\left(\tfrac{m}{n-1}-t\right)d\xi(t)\\
&\qquad+\sum_{j:\,x_j=a}\int_0^{p_j}\left(\tfrac{m}{n-1}-t\right)d\xi(t)-\sum_{j:\,x_j=b}\int_0^{1-p_j}\left(1-\tfrac{m+1}{n-1}-t\right)d\xi(t)\Bigg)\\
&\geq\sum_{m=0}^{n-1}\pi(m)\Bigg(\kappa\left(\tfrac{m}{n}\right)-\int_0^{1-p'}\left(1-\tfrac{m}{n-1}-t\right)d\xi(t)\\
&\qquad+\sum_{j:\,x_j=a}\int_0^{p_j}\left(\tfrac{n_a-1}{n-1}-t\right)d\xi(t)-\sum_{j:\,x_j=b}\int_0^{1-p_j}\left(1-\tfrac{m}{n-1}-t\right)d\xi(t)\Bigg)
\end{aligned}\tag{17}$$

$$\begin{aligned}
\iff&\sum_{m=0}^{n-1}\pi(m)\left(\kappa\left(\tfrac{m+1}{n}\right)-\kappa\left(\tfrac{m}{n}\right)+\sum_{j:\,x_j=a}\int_0^{p_j}\tfrac{1}{n-1}\,d\xi(t)+\sum_{j:\,x_j=b}\int_0^{1-p_j}\tfrac{1}{n-1}\,d\xi(t)\right)\\
&\geq-\int_0^{p_i}\left(p_i-t\right)d\xi(t)-\int_0^{1-p'}\left(1-p_i-t\right)d\xi(t)
\end{aligned}\tag{18}$$

for all $p_i,p_i',p_j(x_{-i})$, and beliefs π such that $\mathrm{E}_\pi[m/(n-1)]=p_i$. The last statement is true only if

$$\sum_{n_a=0}^{n-1}\pi(m)\left(\kappa\left(\tfrac{m+1}{n}\right)-\kappa\left(\tfrac{m}{n}\right)\right)\geq-\int_0^{p_i}\left(p_i-t\right)d\xi(t)-\int_0^{1-p'}\left(1-p_i-t\right)d\xi(t),\tag{19}$$

taking $p_j(a)=0$ and $p_j(b)=1$. This inequality says that the expectation of κ's first differences must be greater than a function of the mean of the distribution. Following a similar argument for agents with opinion b yields the differences in κ having the lower bound

$$\sum_{m=0}^{n-1}\pi(m)\left(\kappa\left(\tfrac{m+1}{n}\right)-\kappa\left(\tfrac{m}{n}\right)\right)\geq-\int_0^{p'}\left(p_i-t\right)d\xi(t)-\int_0^{1-p_i}\left(1-p_i-t\right)d\xi(t)\tag{20}$$

Since the right-hand side of each lower bound is quasi-convex in p' (non-increasing at $p' < p_i$ and non-decreasing at $p' > p_i$), each inequality is satisfied for all p' if and only if it holds for $p' \in \{0,1\}$. Combined, these yields

$$\begin{aligned}\sum_{m=0}^{n-1} \pi(m)\left(\kappa\left(\tfrac{m+1}{n}\right)-\kappa\left(\tfrac{m}{n}\right)\right) \geq \max\Bigg\{ &-\int_0^{p_i}\left(p_i-t\right)d\xi(t),\\ &-\int_0^{p_i}\left(p_i-t\right)d\xi(t)-\int_0^{1}\left(1-p_i-t\right)d\xi(t),\\ &-\int_0^{1-p_i}\left(1-p_i-t\right)d\xi(t),\\ &-\int_0^{1}\left(p_i-t\right)d\xi(t)-\int_0^{1-p_i}\left(1-p_i-t\right)d\xi(t)\Bigg\}\end{aligned} \tag{21}$$

for all $p_i \in [0,1]$ and all π such that $\mathrm{E}_\pi[m/(n-1)] = p_i$.

A lower bound on the expectations of κ's differences for all distributions is equivalent to the differences being separated from the right-hand side by some convex function of p_i. The four quantities in the lower bound are each concave in p_i. The first and fourth are maximized at zero while the second and third are maximized at one, as can be seen by taking first-order conditions via Leibniz's rule. Since the first and fourth are symmetric around $\frac{1}{2}$ with the third and second respectively, attention can be restricted to the second and third terms when considering $p_i \geq \frac{1}{2}$.

Since the terms of the lower bound are concave in p_i, the least restrictive convex upper bound for each term is a supporting line at some point in $[\frac{1}{2},1]$. The supporting line of the second term at ϕ_1 is

$$\begin{aligned}&(p_i-\phi_1)\left(-\int_0^{\phi_1} d\xi(t)+\int_0^{1} d\xi(t)\right)-\int_0^{\phi_1}\left(\phi_1-t\right)d\xi(t)-\int_0^{1}\left(1-\phi_1-t\right)d\xi(t)=\\ &-\int_0^{\phi_1}\left(p_i-t\right)d\xi(t)-\int_0^{1}\left(1-p_i-t\right)d\xi(t)\end{aligned} \tag{22}$$

and the supporting line of the third term at ϕ_2 is

$$(p_i-\phi_2)\left(\int_0^{1-\phi_2} d\xi(t)\right)-\int_0^{1-\phi_2}\left(1-\phi_2-t\right)d\xi(t)=-\int_0^{1-\phi_2}\left(1-p_i-t\right)d\xi(t) \tag{23}$$

The pointwise maximum of the supporting lines is convex and increasing, so this provides a minimal bound of the differences in κ above $\frac{1}{2}$. Define $\delta(m)$ for $m+1 \geq \lceil \frac{n}{2} \rceil$ as

$$\begin{aligned}\delta(m) = \max\Bigg\{ &-\int_0^{\phi_1}\left(\tfrac{m}{n-1}-t\right)d\xi(t)-\int_0^{1}\left(1-\tfrac{m}{n-1}-t\right)d\xi(t),\\ &-\int_0^{1-\phi_2}\left(1-\tfrac{m}{n-1}-t\right)d\xi(t)\Bigg\}\end{aligned} \tag{24}$$

Then, we have $\kappa\left(\frac{m+1}{n}\right) \geq \kappa\left(\frac{m}{n}\right) + \delta(m)$ for $m+1 \geq \lceil \frac{n}{2} \rceil$ when the expectation in line (21) is evaluated at degenerate distributions. Neutrality implies $\kappa\left(\frac{1}{2}\right) = \frac{1}{2}$ and $\kappa(\bar{x}) + \kappa(1-\bar{x}) = 1$, so without loss of generality

$$\kappa\left(\tfrac{n_a}{n}\right) = \frac{1}{2} + \tau\left(\tfrac{n_a}{n} - \tfrac{1}{2}\right) + \mathbb{1}(n \text{ odd}) \frac{\delta\left(\frac{n-1}{2}\right)}{2} + \sum_{m=\lceil n/2 \rceil}^{n_a - 1} \delta(m) \tag{25}$$

for $n_a \geq \lceil \frac{n}{2} \rceil$ with non-decreasing $\tau : [0, \frac{1}{2}] \to \mathbb{R}_+$ to account for excess differences in κ above $\delta(m)$. Since the base score κ must be negatively symmetric around $\frac{1}{2}$, we then have

$$\kappa\left(\tfrac{n_a}{n}\right) = \frac{1}{2} + \text{sign}\left(\tfrac{n_a}{n} - \tfrac{1}{2}\right) \left(\tau\left(\left|\tfrac{n_a}{n} - \tfrac{1}{2}\right|\right) + \mathbb{1}(n \text{ odd}) \frac{\delta\left(\frac{n-1}{2}\right)}{2} + \sum_{m=\lceil n/2 \rceil}^{\max\{n_a, n_b\} - 1} \delta(m) \right) \tag{26}$$

for all n_a. Without loss of generality, a scaling factor of $\frac{1}{n}$ could have been applied to each scoring rule originally and carried through, resulting in the statement of the theorem. □

Sufficiency of Proposition 3. The sufficiency of this representation follows from iterated deletion of interim dominated strategies in the direct mechanism. Consider an agent of type (a, p_i) who conjectures the average proportion of reported opinions is $\hat{p}_i$. By the conditions on the base score, a report of $(a, \hat{p}_i)$ weakly prefers good as all reports (b, p'). A comparison of lines (17) and (18) above shows the agent will strictly prefer $(a, \hat{p}_i)$ to (b, p') as long as the agent thinks there is some chance that p_j and $1 - p_j$ (when $x_j = a$ and $x_j = b$, respectively) are outside a neighborhood of zero where $\xi(t)$ is uniformly zero. Otherwise, a strict incentive from a strictly increasing τ or partial honesty is necessary to guarantee dominance. An analogous argument for agents of type (b, p_i) rules out all (a, p'). Since each agent prefers submitting their true opinion, it follows that each agent weakly prefers submitting their true prediction of the opinions of other agents since T is a proper scoring rule for each agent. Consequently, honest reporting always survives iterated deletion of weakly interim dominated strategies. Other strategies might also survive if agents are indifferent between these reports and honesty, but all will result in the same outcome as honest reporting indifference occurs only when T is constant, with ξ uniformly zero in some interval containing those reports. Therefore, the unique dominance solvable outcome for type profile (x, p) conincides with $T(x, p)$. □

References

AUSTEN-SMITH, D. 1993. Interested experts and policy advice: multiple referrals under open rule. *Games and Economic Behavior 5*, 3–43.

AUSTEN-SMITH, D. AND BANKS, J. S. 1996. Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review 90,* 1, 34–45.

BATTAGLINI, M. 2004. Policy advice with imperfectly informed experts. *Advances in Theoretical Economics 4,* 1.

CHWE, M. S.-Y. 2010. Anonymous Procedures for Condorcet's Model: Robustness, Nonmonotonicity, and Optimality. *Quarterly Journal of Political Science 5,* 1, 45–70.

CRAWFORD, V. P. AND SOBEL, J. 1982. Strategic information transmission. *Econometrica 50,* 6, 1431–1451.

DAWES, R. M. 1989. Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology 25,* 1, 1–17.

DUGGAN, J. AND MARTINELLI, C. 2001. A Bayesian Model of Voting in Juries. *Games and Economic Behavior 37,* 2, 259–294.

FEDDERSEN, T. AND PESENDORFER, W. 1997. Voting behavior and information aggregation in elections with private information. *Econometrica 65,* 5, 1029–1058.

FEDDERSEN, T. AND PESENDORFER, W. 1998. Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting. *American Political Science Review 92,* 1, 23–35.

GERARDI, D. 2000. Jury verdicts and preference diversity. *American Po 94,* 2, 395–406.

GERARDI, D., MCLEAN, R., AND POSTLEWAITE, A. 2009. Aggregation of expert opinions. *Games and Economic Behavior 65,* 2, 339–371.

GLAZER, J. AND RUBINSTEIN, A. 2004. On optimal rules of persuasion. *Econometrica 72,* 6, 1715–1736.

GROFMAN, B., OWEN, G., AND FELD, S. L. 1983. Thirteen theorems in search of the truth. *Theory and Decision 15,* 261–278.

GUARNASCHELLI, S., MCKELVEY, R. D., AND PALFREY, T. R. 2000. An Experimental Study of Jury Decision Rules. *American Political Science Review 94,* 2, 407–423.

KRISHNA, V. AND MORGAN, J. 2001. A model of expertise. *Quarterly Journal of Economics 116,* 2, 747–775.

LAMBERT, N. S. 2011. Elicitation and evaluation of statistical forecasts.

LI, H., ROSEN, S., AND SUEN, W. 2014. Conflicts and Common Interests in Committees. *American Economic Review 91,* 5, 1478–1497.

MARKS, G. AND MILLER, N. 1987. Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin 102,* 1, 72–90.

NITZAN, S. AND PAROUSH, J. 1982. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review 23,* 2, 289—-297.

PRELEC, D. 2004. A Bayesian truth serum for subjective data. *Science 306,* 5695, 462–466.

PRELEC, D. AND SEUNG, H. S. 2007. An algorithm that finds truth even if most people are wrong.

PRELEC, D., SEUNG, H. S., AND MCCOY, J. 2014. Finding truth even if the crowd is wrong.

RILEY, B. 2014. Minimum Truth Serums with Optional Predictions.

ROWAN, T. 1990. Functional Stability Analysis of Numerical Algorithms. Ph.D. thesis, University of Texas at Austin.

SCHERVISH, M. J. 1989. A general method for comparing probability assessors. *The Annals of Statistics 17,* 4, 1856–1879.

WITKOWSKI, J. AND PARKES, D. 2012a. A robust Bayesian truth serum for small populations. In *Proc. of the 26th AAAI Conf. on Artificial Intelligence (AAAI 2012)*.

WITKOWSKI, J. AND PARKES, D. 2012b. Peer prediction without a common prior. In *Proc. of the 13th ACM Conference on Electronic Commerce (EC 2012)*.

WOLINSKY, A. 2002. Eliciting information from multiple experts. *Games and Economic Behavior 41,* 1, 141–160.

YANIV, I. AND KLEINBERGER, E. 2000. Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational behavior and human decision processes 83,* 2, 260–281.

ZHANG, P. AND CHEN, Y. 2014. Elicitability and knowledge-free elicitation with peer prediction. In *Proc. of the 2014 Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2014)*. 245–252.