People Aggregating Signals:
An Experiment and a Short Story

Michele Piccione
London School of Economics

Ariel Rubinstein
School of Economics, Tel Aviv University
and Department of Economics, New York University

April 19, 2023

Abstract: A decision maker observes multiple signals about an event. He has information about the frequency of the event given each individual signal and wishes to update his beliefs about the event. We examine the problem experimentally and identify some of the commonly used procedures for signal aggregation. These procedures are for the most part inconsistent with Bayesian updating. We apply some of these procedures to a well-known panel game that has previously been studied under the standard Bayesian assumptions and find that the properties of the equilibria differ significantly.

Keywords: signal aggregation, experimental economics, voting game
JEL classification: C9, D70, D80

We are grateful to Ayala Arad whose comment on the negative effects of a positive yet weaker additional signal triggered this project and to Aron Tobias for his insightful and generous comments. We also thank Sergiu Hart for his suggestions. We are indebted to the Centre for Behavioural and Experimental Social Science at The University of East Anglia, and especially to Theodore Turocy, and to the Center for Experimental Social Science at New York University, and especially to Anwar Ruff and Shayne Trotman, for their generosity.
1. Introduction

We begin by asking you to answer the following question:

Q1: A very small proportion of the newborns in a certain country have a specific genetic trait.
Two screening tests, A and B, have been introduced for all newborns to identify this trait. However, the tests are not precise. A study has found that:
70% of the newborns who are found to be positive according to test A have the genetic trait (and conversely 30% do not).
20% of the newborns who are found to be positive according to test B have the genetic trait (and conversely 80% do not).
The study has also found that when a newborn has the genetic trait, a positive result in one test does not affect the likelihood of a positive result in the other. Likewise, when a newborn does not have the genetic trait, a positive result in one test does not affect the likelihood of a positive result in the other.

Suppose that a newborn is found to be positive according to both tests. What is your estimate of the likelihood (in %) that this newborn has the genetic trait?

To answer the question you had to process distinct signals, a common occurrence in everyday life and also in many economic models. The usual story in economic theory is that an agent starts out with a prior belief about relevant unknowns (states of the world), understands the relationship between these unknowns and the signals or events that he has observed, and finally uses Bayes' rule to update his beliefs.

If a decision maker observes only one signal in Q1, then his inference is trivial: the likelihood of the trait is given by the description of the signal. Neither the base-rate information nor the distribution of the signal given the states of the world is of any use.

However, in Q1, there are two signals and Bayesian updating is not trivial, as discussed in Section 2. Note that if the base-rate probability were 20%, then the second signal could be ignored and the posterior probability would be 70%. Since the base rate is low, the correct probability of a trait should be strictly above 70%.

The Bayesian analysis yields some surprising conclusions. In particular, when the trait is rare, two positive results in conditionally independent tests indicate a very high
probability that the individual carries the trait, even when each individual test is unreliable in that a large proportion of those who test positive on a single test do not have the trait. This conclusion could be of significance for the legal sphere: two claims about an unlikely event, each unpersuasive on its own, can be very persuasive were considered jointly. For a discussion of aggregation in law see Porat and Posner (2012).

Introspection and a large psychological literature cast doubt on the hypothesis that Bayesian updating procedures resemble the procedures used in real life, even when the information needed to form Bayesian posteriors is clearly specified. In particular, much attention has been given to the “base-rate fallacy”, identified in Kahneman and Tversky (1973) and elucidated in Bar-Hillel (1980). For a survey of the literature on errors in probabilistic reasoning see Benjamin (2019) and for one on decision-making heuristics, see Gigerenzer and Gaissmaier (2011).

Our experimental question however is a different one. We are interested in the way that people aggregate signals. In the experimental part of the paper, we report the results of a series of experiments in which subjects answered questions like Q1. The vast majority of respondents aggregated the signals using a small number of procedures based on simple formulae that bear no similarity to Bayesian updating and yield one of the following answers to Q1: 14, 20, 45, 70 or 76. Only the last answer is qualitatively consistent with Bayesian updating.

Note an important feature of the information presented in Q1. In economic models, the information structure is usually described by specifying a prior distribution of the states of the world and the distribution of the signals given each state. In contrast, question Q1 specifies the distribution of the states of the world given an outcome of each signal, namely the likelihood of the genetic trait given a positive result in either test. Framing the information structure in this manner is representative of situations where decision makers have access to reliable statistics about the predictive success of individual signals but not of joint ones. This type of statistic is widely used in many real-life scenarios. In medicine, the results of a test are often assessed in terms of their positive predictive value (PPV) which is defined by the NIH as: “The likelihood that an individual with a positive test result truly has the particular gene and/or disease in question”\(^1\). Similarly, the effectiveness of an alarm system is often measured by the false alarm ratio.

\(^1\)https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/positive-predictive-value
(FAR), which is “the number of false alarms per the total number of warnings or alarms in a given study or situation”\(^2\). And just as importantly, this is the information we gather informally about a multitude of signals in daily life, such as the proportion of times an informer was right about the hiding place of a suspect, the chances of snow when predicted by a forecaster, or the likelihood that someone is at the door when the dog barks.

The second part of the paper provides a theoretical analysis of non-Bayesian procedures of signal aggregation in a conventional model analyzed in Duggan and Martinelli (2001): A panel is to determine whether a defendant is guilty or not. The panel consists of \(n\) referees each of whom receives a private signal about the guilt of the defendant and then votes whether or not to convict. A conviction requires unanimity. In the standard analysis, a referee votes to convict whenever he concludes that, conditional on his signal and on his vote being pivotal (namely, all the other referees voted to convict), the Bayesian posterior probability that the defendant is guilty is sufficiently large. The game has an informative equilibrium where each referee votes to convict if his signal is above some cutoff. Under some assumptions on the distributions of signals - which we also adopt - when the number of referees is large the voting mechanism yields the correct decision with a probability that approaches 1.

We will examine whether this model’s conclusions hold when the referees are not Bayesian but rather use one of the procedures observed in our experiments. More precisely, we assume that each referee aggregates two signals: his private signal and the event that he is pivotal. Each referee has access to information about the frequency of guilt given either signal separately, applies a signal-aggregation procedure in order to form a belief about the defendant’s guilt, and votes to convict if his belief is above some threshold.

Replacing the Bayesian approach with more realistic methods of signal aggregation yields dramatically different results. We focus on two procedures: the \textit{Avg} procedure which averages the probabilities of the event given each signal, and the \textit{Max} procedure which selects the maximum probability. We will show that, in equilibrium, as the number of referees increases:

(i) If all referees use the \textit{Avg} procedure, then the defendant is almost never convicted.

(ii) If all referees use the \textit{Max} procedure, then the probability of convicting a guilty de-

\(^2\)https://www.statisticshowto.com/false-alarm-ratio-definition/
fendant converges to 1 and that of convicting an innocent one converges to a positive probability.

(iii) The level of welfare converges to the same level in both procedures.

We chose this model mainly because it is simple and well known and particularly because it requires the agents to aggregate two signals: an explicit private signal and being pivotal given the equilibrium behavior of the other players. We agree with the criticism that the application of pivotal reasoning in voting is rather unrealistic (see Esporda and Vespa (2014) for experimental evidence). We consider this part of the paper a simple theoretical demonstration of the effects of more realistic procedures for signal aggregation.

Many other applications could have been considered. In mechanism design, auctions, bargaining, and pricing with rational expectations, agents aggregate multiple signals that are either observed directly or inferred from equilibrium, as in the panel model. Any claim in favor of the relevance of these models rests on the extent to which their results remain valid when agents aggregate signals in a realistic manner.

2. Bayesian analysis of signal aggregation

Suppose we are interested in ascertaining whether an individual has a particular genetic trait, and multiple conditionally independent screening tests, denoted by 1, ..., K, are available. Also suppose that eventually it will become known who has the trait and that a proportion \( \phi_k \in [s, 1) \) of the individuals who tested positive according to test \( k \) indeed have it where \( s \in (0, 1) \) is the frequency of the trait in the population. Denote by \( \pi \) the probability that an individual who is found positive in all \( K \) tests indeed has the trait. Define \( p_k \) and \( n_k \) to be the probabilities of a positive result in test \( k \) for an individual with and without the trait, respectively. By standard Bayesian updating, \( \frac{\phi_k}{1-\phi_k} = \frac{sp_k}{(1-s)n_k} \) and thus:

\[
\frac{\pi}{1-\pi} = \frac{s \prod_k p_k}{(1-s) \prod_k n_k} = \left( \frac{1-s}{s} \right)^{K-1} \prod_k \frac{\phi_k}{1-\phi_k}.
\]

Therefore, if there are at least two signals \( (K > 1) \), \( \pi \) depends on \( s \). Three qualitative conclusions follow from the above:

(i) A test \( j \) for which \( s = \phi_j \) can be ignored.
(ii) An additional positive result increases the probability that the individual has the trait.

(iii) Keeping all $\phi_k$ constant, $\pi$ converges to 1 as $s$ converges to 0 for any $K \geq 2$.

Thus, when $s$ is small, a positive result on two conditionally independent tests can significantly increase the probability of the individual having the trait. The conclusion is striking: multiple tests, each not particularly persuasive on its own, can be very persuasive together. It should be noted that the conclusion that two positive results increase the probability of the individual having the trait over and above that for each individual positive result does not depend on the assumption of conditional independence. For instance, a standard and natural condition such as affiliation of the trait and the tests yields the same conclusion. And yet, this conclusion often fails experimentally, as we shall see.

3. Common formulae for signal aggregation

The experimental results point to four formulae used by subjects to aggregate two positive signals:

- $M^c$: $1 - (1 - \phi_1)(1 - \phi_2)$
- $Avg$: $(\phi_1 + \phi_2)/2$
- $Max$: $\max\{\phi_1, \phi_2\}$
- $M$: $\phi_1 \phi_2$

These formulae depend only on $\phi_1$ and $\phi_2$, thus suggesting base-rate neglect. However, this is not an inevitable conclusion since the subjects’ selection of a formula may depend on the prior distribution. For instance, if $s = \phi_1 < \phi_2$ then the $Max$ formula coincides with Bayesian updating and is consistent with the understanding that the first positive result conveys no useful information. In contrast, if $s < \phi_1 < \phi_2$, Bayesian updating yields an answer strictly above $\max\{\phi_1, \phi_2\}$. In this case, $M^c$ is the only formula in the table that is consistent with the understanding that each positive result increases the posterior probability.

The $M^c$ formula is especially interesting. We conjecture that its rationale is as follows: Since for each test $k$ the proportion of those who have a positive result but do not have the trait is $1 - \phi_k$, by “conditional independence” the proportion of individuals

---

3See Milgrom and Weber (1982).
with two positive results who do not have the trait is \((1 - \phi_1)(1 - \phi_2)\) and the proportion of those who do is the residual. This logic would be correct if two independent tests were conducted on two different individuals and we wish to know the probability of at least one of them having the trait. In our context, this logic is incorrect and possibly reflects confusion about the assumption of conditional independence. The formula \(M\) is an alternative version of the above (incorrect) logic. However, unlike \(M^c\) it does not incorporate the understanding that the posterior is above both \(\phi_1\) and \(\phi_2\).

4. The experiments

We conducted a series of experiments to identify common approaches to signal aggregation in problems such as question Q1. Each question in the experiments contains:

I. A background story about a characteristic that is present in a population and either qualitative or quantitative information about its frequency (“a very small proportion” or a specific percentage, respectively).

II. Information about two “screening” tests - for each test, the percentage of individuals with the characteristic among those who test positive.

III. A non-technical explanation of conditional independence between the two tests.

IV. A question: What is your estimate of the likelihood that an individual who is found positive according to both tests has the characteristic?

The experimental research proceeded in two stages. First, we conducted the experiments on the platform arielrubinstein.org/gt, a website designed for carrying out pedagogical experiments in choice theory and game theory. Most of the subjects were current or past students in game theory courses. No monetary incentives were provided other than a few subjects being randomly chosen to receive $40 regardless of their answers. For us, this stage would have been sufficient. We nonetheless decided to carry out additional experiments based on more standard conventions in experimental economics even though the benefit in our case was likely to be negligible. The experiments carried out in the first stage are referred to as the pilot.

The final experiments were conducted through the Center for Experimental Social Science at New York University (NYU) and the Centre for Behavioural and Experimental Social Science at the University of East Anglia (UEA). Students registered with the labs
were invited to participate in a short online experiment. Each participant was assigned randomly to one of five questions: Q1, Q2, …, Q5. At UEA, one out of ten students received 20 pounds and at NYU each student received $10. Students were not incentivized in any other way for their answers. We do not report the results for each center separately since the differences were minor.

Q1 which was presented in the Introduction conveys the basic flavor of the experiments. Recall that the question concerns a genetic trait which affects a very small proportion of newborns. Two conditionally independent screening tests are available. The likelihood that a newborn has this trait when testing positive is 70% on one test and 20% on the other. The subject is asked to estimate the chances of a newborn having the trait if he tests positive on both tests.

In Q2 we changed the probabilities to 80% and 60%. In Q3, we changed the underlying story somewhat by replacing newborns with undergraduates who are either continuing on to graduate school or not. A positive result on a test is replaced by a student having taken a certain course. The proportion of students who continue to graduate school is 70% for those who took one of the courses and 20% for those who took the other. The results for Q1, Q2 and Q3 are presented in Table 1.

<table>
<thead>
<tr>
<th>Question</th>
<th>Type</th>
<th>Probability</th>
<th>Mean</th>
<th>Max</th>
<th>Avg</th>
<th>Maxc</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>genetic</td>
<td>70-20</td>
<td>51%</td>
<td>20%</td>
<td>14%</td>
<td>14%</td>
<td>20% (4%)</td>
</tr>
<tr>
<td>Q1 Pilot</td>
<td></td>
<td></td>
<td>51%</td>
<td>18%</td>
<td>11%</td>
<td>9%</td>
<td>31% (12%)</td>
</tr>
<tr>
<td>Q2</td>
<td>genetic</td>
<td>80-60</td>
<td>68%</td>
<td>21%</td>
<td>27%</td>
<td>15%</td>
<td>18% (11%)</td>
</tr>
<tr>
<td>Q3</td>
<td>students</td>
<td>70-20</td>
<td>55%</td>
<td>7%</td>
<td>27%</td>
<td>20%</td>
<td>21% (6%)</td>
</tr>
</tbody>
</table>

Table 1: Results for Q1, Q2 and Q3.

As is evident from the table, a clear majority (about 60%) of the subjects used one of the four formulae: $M$, $Avg$, $Max$ or $Maxc$. Around 20% chose an answer strictly above the larger between $\phi_1$ and $\phi_2$, which is qualitatively correct. There do not appear to be any significant differences in the results between the three questions. The median response time (MRT) was also quite similar in all three, ranging from 98 to 118 seconds.

It is worth comparing the above results with those obtained using the pedagogical site (the pilot). As is evident from Table 2, which only relates to Q1, the results are similar.
The differences were: (i) students in the pilot had a higher response time (the MRT was in the range of 120-200 seconds); (ii) a larger proportion of subjects in the pilot (31% vs. 20%) gave a qualitatively correct answer (a number exceeding both $\phi_1$ and $\phi_2$).

Q4 was used to determine the consistency of the approaches used by the subjects. We asked the subjects to simultaneously assess the likelihood of the trait given two different pairs of tests: Alice tested positive on two tests with accuracies of 70% and 20% respectively and Bob tested positive on two tests with accuracies of 50% and 40% respectively. We observe a considerable degree of consistency. Of the 92 subjects, about 60% were consistent in their use of one of the four formulae: $M$ (15%), $Avg$ (30%), $Max$ (13%) and $M^c$ (2%). Furthermore, 8% of the subjects consistently chose answers above $Max$ that differed from the answer according to $M^c$.

Q5 is identical to Q1 except that the base rate is 20% (rather than “a very small proportion”). As noted earlier, when the base rate is equal to $\phi_1$, the correct answer is $\phi_2$. Nonetheless, all four formulae are still used in this question. Notably, there is no significant change in the use of $Max$ - which is the correct answer - relative to Q1. About 13% of the subjects gave the base-rate probability (20%) as their answer. Note also that 16% of subjects chose an answer strictly above both $\phi_1$ and $\phi_2$, although in this question any answer different from 70% is qualitatively incorrect.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>$n$ (MRT)</th>
<th>$M$</th>
<th>20%</th>
<th>$Avg$</th>
<th>$Max$</th>
<th>$&gt; Max$ ($M^c$)</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q5</td>
<td>49%</td>
<td>94 [1214&amp; rbrack;</td>
<td>10%</td>
<td>13%</td>
<td>9%</td>
<td>17%</td>
<td>16% (3%)</td>
<td>36%</td>
</tr>
</tbody>
</table>

Table 2: Results for Q5.
Comments:
1. Notice the significant difference between the experimental results here and those attributed to the base-rate fallacy. When the prior probability of a trait is small, the base-rate fallacy literature reports that it is usually ignored and the experimental assessment of the trait given a positive signal is much higher than it should be with Bayesian updating. In contrast, we find that when the base rate of the trait is low, the experimental assessment of a trait given two positive signals is much lower than it should be.

2. Our choice of describing the prior in Q1 to Q4 qualitatively was deliberate. Since several of our pilot experiments show that quantitative descriptions of the prior do not alter the nature of the results, in the main experiment we opted for a qualitative framing that is closer to the way that data are described in real life.

3. Our objective is not to find the “real-life” distribution of updating procedures in inference problems such as Q1. Rather, our experimental results - and in our opinion any results in experimental economics - are mainly useful to confirm intuitions. In this paper, our intuition is that individuals often aggregate signals using a small number of simple and identifiable procedures that are incompatible with the Bayesian approach.

4. A version of Problem 6 in Bar-Hillel (1980) is closest in character to our questions. However, in that setup, information about the signals is presented as in standard economic models, namely by specifying their precision conditional on the true state. A modified, shorter version of her question is:

85% of the cabs in the city are Blue and 15% are Green. A cab was involved in a hit-and-run accident at night. There were two witnesses to the accident. Both claim that the errant cab was Green. The first witness is able to identify each color correctly about 80% of the time; the second witness identified each color correctly 70% of the time. What do you think are the chances that the errant cab was Green?

In Bar-Hillel (1980), 83% of the 29 subjects chose the average (75%) and thus “were disregarding the base rate”. We repeated the original experiment in our pilot, as in Bar-Hillel (1980), without monetary incentives. Our results differed from hers: only 16% of the 93 subjects answered 75% and the rest of the answers were widely dispersed: 8% chose 56%, 8% chose 15% and only 5% chose the correct answer (62%).
5. A panel story

In the rest of the paper, we will investigate the effects of using some of the non-Bayesian procedures observed in our experiments by means of a game analysed in Duggan and Martinelli (2001). As we wrote in the introduction, we include this analysis not because we think that it is a realistic framework for the study of voting, but as a demonstration of the sharp effects that replacing Bayesian signal aggregation with more realistic procedures can have in well-known economic models.

A panel of \( n \) referees is to decide whether a defendant is guilty. The prior probability of guilt is \( s \) and of innocence is \( 1 - s \). Each referee votes either Y (guilty) or N (not guilty) and the defendant is found guilty when the referees unanimously vote Y. The referees vote simultaneously. Prior to voting each referee receives a private signal in the form of a number in the interval \([0, 1]\). The signals are identically distributed and conditionally independent across the referees. The cdf of each signal conditional on the defendant being guilty is \( F \) and that conditional on being innocent is \( G \). The cdfs \( F \) and \( G \) have continuous density functions \( f \) and \( g \), respectively. We impose the following restrictions throughout:

(i) \( f(0) = 0, g(1) = 0, f(t) > 0 \) for all \( t > 0 \) and \( g(t) > 0 \) for all \( t < 1 \).

(ii) \( \frac{f(t)}{g(t)} \) is strictly increasing.

Note that (ii) implies that \( \frac{F(t)}{G(t)} < \frac{1 - F(t)}{1 - G(t)} \) for all \( t \in (0, 1) \), that \( \frac{F(t)}{G(t)} \) and \( \frac{1 - F(t)}{1 - G(t)} \) are strictly increasing, and together with (i), that \( \lim_{t \to 1} \frac{1 - F(t)}{1 - G(t)} = \infty \).

A referee prefers that the defendant is convicted if he believes that the probability of him being guilty is at least \( z \), a number in \((0, 1)\). This is consistent with each referee maximizing a vNM utility that is equal to 1 if the correct decision is made when the defendant is guilty; equals to \( \lambda = \frac{z}{1 - z} \) if the correct decision is made when the defendant is innocent; and equals to 0 if the incorrect decision is made. Therefore, the natural welfare function when all \( n \) referees use a common cutoff \( \alpha \), namely each referee votes Y if and only if his observed signal is at least \( \alpha \), is:

\[
W^n(\alpha) = s(1 - F(\alpha))^n + (1 - s)\lambda(1 - (1 - G(\alpha))^n).
\]

We focus on the case where \( z \geq 1/2 \geq s \). That is, the ex-ante belief that the defendant is
guilty is not stronger than the belief that he is innocent and the standards for conviction are higher than the standards for acquittal.

5.1 Signals and beliefs

The signals in our experiments are exogenous events that are communicated to the agents. In a strategic environment with incomplete information, the strategies of the players generate events that can be interpreted as “signals” and can be aggregated with other exogenous signals to formulate relevant posterior beliefs.

In the standard approach to binary voting models, a voter's best reply depends on his own information and the event in which his vote is pivotal. Following this approach, we assume that a referee forms his beliefs by aggregating two “signals”: (i) his own private signal; and (ii) the occurrence of the event that he is pivotal. Note that in this model the signal about a referee's vote being pivotal is conditionally independent of the referee's own vote. This feature allows the introduction of strategic considerations into non-Bayesian procedures in an uncomplicated way.

We refer to referee $i$'s belief that the defendant is guilty given his private signal $t$ and conditional on him being pivotal given a strategy profile $S_{-i}$ of the other players as his C-belief and denote it by $\mu_i(t, S_{-i})$. If all the players use the same cutoff $\sigma$, namely each votes Y if and only if his signal is above $\sigma$, then we denote the C-belief by $\mu_i(t, \sigma)$. Naturally, C-beliefs depend on the procedures - whether Bayesian or non-Bayesian - used by the referees to update their beliefs.

5.2 Equilibrium

To summarize, the model is the tuple $\langle n, s, z, F, G, (\mu_i) \rangle$ where $n$ is the number of referees, $s$ is the common prior probability that the defendant is guilty, $z$ is the common minimal belief as to the guilt of the defendant for which a conviction is optimal, $F$ and $G$ are the distributions of each private signal given that the defendant is respectively guilty or not guilty, and $\mu_i$ is referee $i$'s C-belief. We are left to define the equilibrium concept. A (symmetric) equilibrium is a cutoff $\sigma^* < 1$ such that for every referee $i$ we have $\mu_i(t, \sigma^*) \leq z$ for all $t < \sigma^*$ and $\mu_i(t, \sigma^*) \geq z$ for all $t > \sigma^*$. That is, in equilibrium whenever a referee votes N he believes that the probability of the defendant being guilty is sufficiently low (weakly below $z$) and whenever he votes Y he believes that it is sufficiently high (weakly above $z$). We require that $\sigma^* < 1$ in order to exclude the case in which all players vote N regardless of their signals and that no player is ever pivotal.
6. Bayesian equilibria

We first review the model under the standard approach where for each $i$ the C-beliefs are defined by Bayesian updating:

$$
\mu_i(t, \sigma) = \frac{s f(t)(1 - F(\sigma))^{n-1}}{s f(t)(1 - F(\sigma))^{n-1} + (1 - s) g(t)(1 - G(\sigma))^{n-1}}.
$$

Since $\mu_i$ is strictly increasing in $t$ for a fixed $\sigma < 1$, $\sigma^*$ is an equilibrium if and only if it satisfies:

$$
\mu_i(\sigma^*, \sigma^*) = z.
$$

Clearly, the equilibrium is unique.\(^4\) For later use, we rearrange the equilibrium condition. Let $r^k(\theta)$ be the probability that the defendant is guilty conditional on $k$ referees with a common cutoff $\theta$ voting Y, that is,

$$
r^k(\theta) = \frac{s(1 - F(\theta))^k}{s(1 - F(\theta))^k + (1 - s)(1 - G(\theta))^k},
$$

and we set $r^k(1) = 1$, which is the limit of $r^k(\theta)$ as $\theta \to 1$. The function $r^k$ is strictly increasing and $r^k(\theta) < r^{k+1}(\theta)$ for any $k \geq 1$ and $0 < \theta < 1$. Then, $\sigma^*$ is an equilibrium if and only if it satisfies:

$$
r^{n-1}(\sigma^*) = \frac{z g(\sigma^*)}{z g(\sigma^*) + (1 - z) f(\sigma^*)}.
$$

It follows from Duggan and Martinelli (2001) that the equilibrium condition also characterizes welfare maximization with a common cutoff. As the number of referees grows, the optimal cutoff decreases and converges to 0. Furthermore, the probability of an incorrect decision given the optimal cutoff converges to 0. That is, the equilibrium level of welfare in the standard Bayesian approach converges to ‘the “first-best”, namely $s + (1 - s)\lambda$.

\(^4\)Given our assumptions, we have that $\mu_i(0, 0) = 0$, and the function $\mu_i(t, t)$ is increasing, continuous and converges to 1 as $t \to 1$. Therefore, there is a unique $\alpha$ satisfying the equation $\mu_i(\alpha, \alpha) = z$. 
7. Non-Bayesian games: aggregating two signals

We now consider the case where all or some of the players are not Bayesian and use one of the procedures for signal aggregation observed in the experiments. All these procedures use a formula that is a function of two numbers: the probability of the relevant state conditional on each signal. In our case, the two numbers are:

(i) the probability $p(t)$ of the defendant being guilty given the referee’s own signal $t$, that is:

$$p(t) = \frac{s f(t)}{s f(t) + (1 - s) g(t)}$$

and (ii) the probability $r^{n-1}(\theta)$ that the defendant is guilty given that the referee is pivotal and the rest of the referees use the cutoff $\theta$. Note that the function $p$ is strictly increasing and satisfies $p(t) < r^k(t)$ for any $t \in (0, 1)$ and $k \geq 1$.

As mentioned above, in the standard analysis of the panel game the equilibrium cutoff maximizes welfare and when the number of referees is large the correct outcome is obtained almost with certainty. Our results show that these conclusions do not carry over to some of the non-Bayesian procedures observed in our experiments. In particular:

(1) If all the referees use the \textit{Avg} procedure, then the equilibrium cutoff is higher than the Bayesian one and when the number of referees is large almost all defendants are acquitted.

(2) If all the referees use the \textit{Max} procedure, then the equilibrium cutoff will typically be sub-optimal. When the number of referees is large, the equilibrium cutoff is below the Bayesian one and all guilty defendants are convicted, as well as some proportion of the innocent ones. Remarkably, the level of welfare converges to that achieved when the referees never convict.

(3) If some proportion of the referees - no matter how small - uses the \textit{Avg} procedure and the remaining referees are Bayesian then, when the number of referees goes to infinity, it is almost certain that the defendant will be acquitted.

While a “large” Bayesian panel approximates the first best, all of the other types of “large” panels we investigate achieve approximately the same level of welfare as a panel that never convicts, even though the members of the panel use different equilibrium
strategies.

7.1 The Avg game

Assume that all referees use the Avg procedure, that is referee $i$’s C-belief is:

$$\mu_i(t, \sigma) = \frac{1}{2}(p(t) + r^{n-1}(\sigma)).$$

Since C-beliefs are monotonic in $t$, equilibria are characterized by the solutions of the equation $\frac{1}{2}(p(\beta) + r^{n-1}(\beta)) = z$. The following claim states that the equilibrium cutoff is above the Bayesian one and as the number of referees grows the probability that any defendant is convicted declines to zero.

**Claim A** In the Avg game:

(i) There is a unique equilibrium denoted by $\beta^*$.

(ii) $\beta^* \geq \sigma^*$ with strict inequality unless $z = s = \frac{1}{2}$.

(iii) Denote by $\beta^*(n)$ the equilibrium in the game with $n$ referees. If $z > \frac{1}{2}$ then as $n \to \infty$

(a) $\beta^*(n) \to \tilde{\beta}$ where $p(\tilde{\beta}) + 1 = 2z$ and (b) the level of welfare converges to $\lambda(1 - s)$ (the defendant is almost never convicted).

**Proof.** (i) The function $\frac{1}{2}(p(\beta) + r^{n-1}(\beta))$ is increasing in $\beta$ and ranges from $\frac{s}{2}$ to 1 and hence there is a unique $\beta^*$ at which it equals $z$. Clearly this is the unique equilibrium.

(ii) Since $z \geq s$

$$p(\sigma^*) + r^{n-1}(\sigma^*) = \frac{sf(\sigma^*)}{sf(\sigma^*) + (1-s)g(\sigma^*)} + \frac{zg(\sigma^*)}{zg(\sigma^*) + (1-z)f(\sigma^*)} \leq \frac{zf(\sigma^*)}{zf(\sigma^*) + (1-z)g(\sigma^*)} + \frac{zg(\sigma^*)}{zg(\sigma^*) + (1-z)f(\sigma^*)} \leq 2z.$$ 

The first inequality holds strictly unless $s = z = \frac{1}{2}$ given that $z \geq \frac{1}{2} \geq s$. The second inequality follows from $\frac{x}{x + (1-z)y} + \frac{y}{y + (1-z)x} \leq 2$ which holds with equality only for $z = \frac{1}{2}$ or $x = y$ (or $z = 1$). Therefore, one of the inequalities will hold strictly unless $s = z = \frac{1}{2}$.

By the monotonicity of $p$ and $r^{n-1}$, it follows that $\beta^* \geq \sigma^*$ with strict inequality unless $z = s = \frac{1}{2}$. 

15
(iii) Since the function \( \frac{1}{2}(p(\beta) + r^{n-1}(\beta)) \) is increasing in \( \beta \) and \( n \), the sequence \( \beta^*(n) \) is decreasing. Since \( p(\beta^*(n)) \geq 2z - 1 > 0 \), \( \beta^*(n) \) is bounded away from 0 and the equilibrium probability that the defendant will be found guilty converges to 0 as \( n \to \infty \). Therefore, the sequence \( \beta^*(n) \) converges to the unique solution of \( p(\beta) + 1 = 2z \).

7.2 The Max game

Suppose that all referees use the Max procedure, that is, for all \( i \):

\[
\mu_i(t, \theta) = \max \{ p(t), r^{n-1}(\theta) \}.
\]

The following claim states that the equilibrium cutoff is always below the Avg cutoff and also below the Bayesian one when \( n \) is large. As the number of referees grows, the equilibrium level of welfare approaches that in the Avg game, namely the level of welfare in the case of a panel that never convicts, despite the fact that convictions occur with probability larger than \( s \).

**Claim B** In the Max game:

(i) There exists a unique equilibrium \( \gamma^* \) satisfying \( r^{n-1}(\gamma^*) = z \).

(ii) \( \gamma^* \leq \beta^* \).

(iii) Let \( \gamma^*(n) \) be the equilibrium in the game with \( n \) referees. There exists a sequence \( \eta(n) \) that converges to \( \infty \) such that for any \( n \), \( \sigma^*(n) > \gamma^*(n) \) if and only if \( \lambda < \eta(n) \).

(iv) As \( n \) increases, the probability of convicting a guilty defendant converges to 1 and that of convicting an innocent one converges to \( \frac{s}{(1-s)\lambda} \). The level of welfare converges to \( \lambda(1-s) \).

**Proof.**

(i) Let \( \gamma^* \) be the unique solution of \( r^{n-1}(\gamma) = z \). Then \( \gamma^* \) is an equilibrium: if a referee receives a signal below \( \gamma^* \) his C-belief is \( z \) and thus he is indifferent between voting Y and voting N; if he receives a signal above \( \gamma^* \), then his C-belief is at least \( z \) and thus voting Y is optimal.

There is no other equilibrium:

(a) A common cutoff \( \gamma < \gamma^* \) is not an equilibrium since a referee with a signal \( t \in (\gamma, \gamma^*) \) has a C-belief equal to \( \max \{ p(t), r^{n-1}(\gamma) \} \) which is less than \( z \) since \( p(t) < p(\gamma^*) < r^{n-1}(\gamma^*) = z \) and \( r^{n-1}(\gamma) < r^{n-1}(\gamma^*) = z \). Such a referee prefers to vote N.

16
(b) A common cutoff $\tilde{\gamma} > \gamma^*$ is not an equilibrium since in that case any referee’s C-belief is at least $r^{n-1}(\tilde{\gamma}) > z$ and thus voting N is not optimal given any signal.

(ii) The assertion follows from the Avg equilibrium condition $\frac{1}{2}(p(\beta^*) + r^{n-1}(\beta^*)) = z$ and the fact that $r^{n-1}(t) > p(t)$ for all $t \in (0, 1)$.

(iii) $\gamma^*(n)$ is the solution to $r^{n-1}(t) = \lambda$ and $\sigma^*(n)$ is the solution to $\frac{r^{n-1}(t)}{1-r^{n-1}(t)} f(t) = \lambda$. The two LHS functions are increasing and have the same value only at the point $\tilde{t}$ such that $f(\tilde{t}) = g(\tilde{t})$. Define $\eta(n) = \frac{r^{n-1}(\tilde{t})}{1-r^{n-1}(\tilde{t})}$. The sequence converges to infinity since $F(\tilde{t}) < G(\tilde{t})$.

(iv) Since $\gamma^*(n) < \sigma^*(n)$ for large $n$, the probability that a guilty defendant is found guilty goes to 1. The ratio between guilty and innocent defendants that are convicted is $\frac{\frac{k}{s(1-F(\gamma^*(n)))}}{\frac{k}{1-s}(1-G(\gamma^*(n)))}$ which converges to $\lambda$ by the equilibrium condition. Therefore, the probability of an innocent defendant being found guilty converges to $s \frac{1}{1-s} \frac{1-s}{z}$ and the level of welfare converges to $s + (1-s)\lambda \frac{1}{1-s} \frac{1-s}{z} = \lambda \frac{1}{1-s}$.

Remark (the Min game): By an argument similar to that in Claim B the only equilibrium for the game in which all referees follow the Min procedure is $\delta^*$ satisfying $p(\delta^*) = z$. Since the equilibrium is independent of $n$ and $\delta^* > 0$, the probability of conviction goes to zero as the number of referees grows, and the level of welfare converges to $(1-s)\lambda$, as in the Avg and Max games.

7.3 The Mixed Bayesian and Avg game

Suppose that $n\kappa$ referees are Bayesian and the rest are Avg. We assume that at least two referees are Avg. An extension of the equilibrium definition specifies that $\alpha^* < 1$ and $\beta^* < 1$ where $\alpha^*$ is the common cutoff of the Bayesian players and $\beta^*$ is the common cutoff of the Avg players.

The following claim shows that, as long as the proportion of Avg referees is positive, when the number of referees grows the probability that the defendant is convicted goes to zero.
Claim C  Suppose that $z > \frac{1}{2} > s$.\(^5\)

(i) An equilibrium exists.

(ii) Any sequence of equilibria $(\alpha^*(n), \beta^*(n))$ converges to $(0, \bar{\beta})$ where $p(\bar{\beta}) = 2z - 1$ and the equilibrium probability of conviction converges to zero as $n \to \infty$.

Proof. (i) Define $\psi(t) = \frac{1-G(t)}{1-F(t)}$. An equilibrium $(\alpha^*, \beta^*) \in (0,1) \times (0,1)$ is characterized as being a solution to the two equations:

\[
\begin{align*}
1 &= 1 + \frac{1-s}{f(\alpha)}(\psi(\alpha))^{kn-1}(\psi(\beta))^{n-kn} \\
\frac{1}{2}[p(\beta) + \frac{1}{1 + \frac{1-s}{s}(\psi(\alpha))^{kn}(\psi(\beta))^{n-kn-1}}] &= z
\end{align*}
\]

A solution in $(0,1) \times (0,1)$ exists.\(^6\)

(ii) Let $(\alpha^*(n), \beta^*(n))$ be a sequence of equilibria. The sequence $\alpha^*(n)$ converges to 0. If not there would be an $\epsilon > 0$ and a subsequence that is above $\epsilon$. Since $\psi(\epsilon) \in (0,1)$ and $\psi$ is increasing, the LHS of the first equation along the subsequence would converge to 1 > $z$.

Since $p(\beta^*(n)) > 2z - 1 > 0$ there is $\epsilon > 0$ such that $\beta^*(n) > \epsilon$ for all $n$. Since $\psi(\epsilon) \in (0,1)$, by the second equation $\beta^*(n)$ converges to $\bar{\beta}$, the unique solution of $p(\bar{\beta}) = 2z - 1$ and $\beta^*(n) > \bar{\beta}$. It follows that the probability that all $Avg$ referees will vote $Y$ is less than $s(1-F(\bar{\beta}))^{n(1-\kappa)} + (1-s)(1-G(\bar{\beta}))^{n(1-\kappa)}$ which converges to zero as $n \to \infty$. \(\square\)

---

\(^5\)When $s = z = \frac{1}{2}$ there exists an equilibrium which is identical to the one in which all $n$ players are Bayesian (or $Avg$). Let $\tau^*$ be the solution of the equation $\frac{G(\tau)}{1-G(\tau)}(\psi(\tau))^{n-1} = 1$. Then, $\tau^*$ also solves the equation $p(\tau) + 1/(1 + (\psi(\tau))^{n-1}) = 1$. Therefore, $\alpha^* = \beta^* = \tau^*$ is an equilibrium. Obviously, $\tau^*$ is also an equilibrium when all referees are either Bayesian or $Avg$.

\(^6\)For every $\beta < 1$ there is a unique $a \in (0,1)$ that solves the first equation since the LHS of the equation converges to 0 as $a \to 0$ and to 1 as $a \to 1$. Denote this solution by $a(\beta)$. For every $a \in [0,1]$ there is a unique $\beta \in (0,1)$ which solves the second equation since the LHS converges to 1 as $\beta \to 1$ and is equal to $\frac{1}{2(1+\psi(a))^{kn}} < \frac{1}{2}$ at $\beta = 0$. Denote this solution by $b(\alpha)$. The functions $a(\beta)$ and $b(\alpha)$ are continuous and decreasing. It can easily be verified using standard arguments that there is $(\alpha^*, \beta^*) \in (0,1) \times (0,1)$ which solves both equations.
**Comment:** An alternative modelling approach to belief formation with non-Bayesian procedures is that instead of aggregating two signals, each referee aggregates \( n \) signals: his own signal and one distinct signal for each Y vote cast by the other referees. More precisely, a referee \( i \) who receives the signal \( t \) and knows that all the other referees follow a common cutoff \( \theta \) forms his C-belief by aggregating:

(i) his own signal \( t \) (according to which the defendant is guilty with probability \( p(t) \));

and

(ii) \( n - 1 \) signals, one for each referee \( j \neq i \) who is voting Y, that is, referee \( j \)'s signal is at least \( \theta \) (for any one of these signals the defendant is guilty with probability \( r^1(\theta) \)).

Thus, under the \( \text{Avg} \) approach, \( \mu_i(t, \theta) = \frac{1}{n}(p(t) + (n - 1)r^1(\theta)) \) and under the \( \text{Max} \) approach, \( \mu_i(t, \theta) = max\{p(t), r^1(\theta)\} \).

The conclusions obtained in the previous section essentially remain valid under this approach to belief formation. In particular, in the \( \text{Avg} \) and \( \text{Max} \) games, the equilibrium probability of conviction goes to zero as the number of referees becomes large.
References


