An Epistemic Free-Riding Problem?

Christian List and Philip Pettit[1]

1 August 2003

Karl Popper noted that, when social scientists are members of the society they study, they may affect that society. If the individuals to whom a theory initially applies come to understand that theory, then this understanding may affect their behaviour in such a way that the theory ceases to be applicable. This may be called the problem of reflexivity. In this paper, we identify such a problem in an apparently unlikely area: in the area of Condorcet's famous jury theorem. Suppose that each individual member of some decision-making body has a greater than 0.5 chance of making a correct judgment, and suppose further that all individuals' judgments are independent from each other. Then the jury theorem states that the majority will make a correct judgment with a probability approaching 1 as the number of individuals increases. We argue that, if the individuals come to understand the jury theorem, then they may cease to make independent judgments, thereby undermining one of the conditions for the application of the theorem. Specifically, we suggest that the individuals may be faced with a temptation to free-ride on the epistemic efforts of others. We first develop the problem in some detail and then ask whether there are any escape routes that can protect the jury theorem against the effect of reflexivity.

1. Introduction

One of the hallmark themes of Karl Popper's approach to the social sciences was the insistence that when social scientists are members of the society they study, then they are liable to affect that society. In particular, they are liable to affect it in such a way that the claims they make lose their validity. "The interaction between the scientist's pronouncements and social life almost invariably creates situations in which we have not only to consider the truth of such pronouncements, but also their actual influence on future developments. The social scientist may be striving to find the truth; but, at the same time, he must always be exerting a definite influence upon society. The very fact that his pronouncements do exert an influence destroys their objectivity." (Popper 1963, 16).

Suppose that someone propounds a novel theory of social behaviour, and that the theory is well confirmed by how people actually behave. Imagine that the success of the theory leads to its coming to be widely understood and accepted. And now suppose that the very fact of such popularization has an effect on how people behave, leading some or all of them to act contrary to the theory's predictions. Where this happens, the theory is undermined in the general manner that Popper describes. The people to whom it is supposed to apply become reflectively aware of it, and this reflection causes a reaction that impacts negatively on the truth of the theory, or at least on the range of its

application. Our paper is concerned with this sort of influence, one that we describe as reflexivity.[2]

We focus, more particularly, on how reflexivity may operate in a rather unlikely area. The theory we take as our target, if indeed it can be called a theory, is the jury theorem that was first identified by the eighteenth century French thinker, the Marquis de Condorcet (1976), and that has recently attracted renewed attention (Grofman, Owen and Feld 1983, Lahda 1992, Estlund 1994, Austen-Smith and Banks 1996, Feddersen and Pesendorfer 1998, List and Goodin 2001, List 2003). We argue that Condorcet's result is subject to a reflexivity effect of broadly the kind that Popper had in mind. If people come to be aware of the theorem, and to accept it, then in some circumstances they may act in a way that undermines some of the preconditions for the application of the theorem, so that the theorem ceases to apply to them. More specifically, they may do this as a result of a temptation to free-ride on the epistemic efforts of others.

The material in this paper is divided into three sections. We introduce the jury theorem in the first section. We look at the core problem in the second. And we consider in the third section whether there are any escape routes that can protect the jury theorem against the effect of reflexivity. We present the arguments of the opening two sections, first in an informal way, and then in a more formal, technically exact manner. Although the technical presentation is essential for formulating the arguments properly, readers who wish to skip it can get a sense of our argument from the informal presentation.

2. The Condorcetian background

2.1 An informal presentation

Suppose that there is an issue of fact on which a number of individuals are each to judge. Let the issue be whether something is the case or not the case – assuming that it must be one or the other. The fact that it is the case will be represented as '$X = 1$', the fact that it is not the case as '$X = 0$'. The issue may be observational: say, whether the car in an accident drove through the lights on the red or not. Or it may be empirical but more speculative: say, whether the population size of the United States is greater than 300 million or not. Or it may be adjudicative, as in whether a defendant is guilty or not guilty.

Or it may be theoretical, as with the issue of whether or not quarks can exist separately from one another. Or whatever: there is no limit on the possibilities in play.

Suppose that the members of the relevant group each have a better-than-random, i.e. better than 0.5, chance of making a correct judgment. Suppose, more demandingly, that they each have the same chance of making a correct judgment; say, a chance of 0.6. Suppose, further, that their individual judgments are independent, given the relevant fact of the matter. This means, roughly, that any individual's chance of making a correct judgment is not increased or decreased by any facts about how any of the other individuals have judged the issue. It is not the case, for example, that anyone defers to another, always copying the judgment of the other.[3] And suppose, finally, that they each reveal their own judgment truthfully when the group takes a vote on the issue in question. These independent judgments we describe as the *private judgments* of the individuals.

What does it mean to say that each individual has a chance, or *competence level*, of 0.6 of making a correct judgment? It means that if $X = 1$, then the individual has a 0.6 probability of judging that $X = 1$; and that if $X = 0$, the individual has a 0.6 probability of judging that $X = 0$. The individual has that competence level in tracking the fact that the proposition under judgment is true, if it is true; and has the same competence level in tracking the fact that the proposition is false, if it is false (Nozick 1981).

Condorcet showed that if these conditions are met, and every group member casts a vote on the issue, then two remarkable consequences follow. The first is that the probability of the majority making a correct judgment on the issue (assuming there is no tie) will be greater than that of any one individual's making a correct judgment; in our example, it will be greater than 0.6. And the second is that as the group size increases, the probability of the majority making a correct judgment approaches 1. The lesson is that under the conditions outlined, there is epistemic safety in numbers. No one individual has a better chance of making a correct judgment than the majority, and going along with the majority looks like a better and better epistemic strategy as the group size increases.

We do not present the formal argument for the jury theorem here but it is not difficult to get a sense of why it holds. Consider a biased coin that has a 0.6 probability of coming up heads and a 0.4 probability of coming up tails. Now think about the chance of

the coin coming up heads on a single throw; this is 0.6. And then ask yourself whether there is a greater chance of its coming up heads more often than not, as you keep throwing it. Intuitively, there is a greater chance of this happening than there is of the coin coming up heads on a single throw. Let each coin toss represent the judgment of an individual, and let heads represent a correct judgment, and tails an incorrect one. Then our reasoning about the coin should make it plausible that there is a better chance of the majority making a correct judgment than there is of any single individual making a correct judgment.

So much for the first part of Condorcet's result. The second – that the probability of the majority making a correct judgment approaches 1 as the group size increases – can also be made plausible by analogy with the coin case. Again imagine a coin that has a 0.6 probability of coming up heads and a 0.4 probability of coming up tails. Think about how often you expect the coin to come up heads and how often tails when it is thrown repeatedly. Statistically, you expect it to come up heads 60% of the time and tails 40% of the time. First consider the case of ten throws. The expected heads-tails pattern is 6-4. But the actual pattern may still deviate from this expected one: it might be 7-3, or 5-5, or occasionally even 4-6. But now consider the case of a hundred throws. Here the expected heads-tails pattern is 60-40. Again, the actual pattern may deviate from this; it might be 58-42, or 63-37, or 55-45. But intuitively it will be less likely than in the case of ten throws that we get heads less than 50% of the time. Finally, consider the case of a thousand, ten thousand, or a million throws. Given the expected frequency of 60%, it will be less and less likely that the coin comes up heads less than 50% of the time. By the law of large numbers, the actual frequency of heads will approximate the expected one increasingly closely as the number of throws increases. This implies that the probability of getting a majority on the 'heads' side rather than the 'tails' side approaches 1 as the number of throws increases.

Condorcet described the result as a jury theorem and it is easy to see why it might apply to multi-member juries making a judgment on whether or not a defendant is guilty. The use of simple majority voting in that case is somewhat problematic, though, if we are primarily concerned, not with the chance of the jury getting the right answer, but rather

with the chance of the jury not convicting an innocent person. But we leave aside that complication here. There are many different cases where that complication does not arise.

Some of those cases, like the jury example, involve a collective body of people who act together with a view to making a judgment that is definitive in some way. They may involve a committee deciding on who should get a prize, raising the question with every candidate of whether he or she has made the best contribution or not. Or they may involve an appointments or promotions or examinations committee of a parallel kind. Or any of a variety of similar bodies: say, the commission that is required to judge on whether a certain proposal should be accepted or not, or the board of an organisation that has to judge on what course of action should be recommended to members, and so on.

But Condorcet's issue arises even in cases where there is no collective body making a joint decision. Take as an example the community of scientists in a given subject-area. While these scientists will collaborate and compete with one another, thereby establishing a community, they will not constitute a collective body that has a task to discharge that parallels the task of a jury or committee. But the jury theorem still has a lesson for such a community. If we suppose, perhaps fancifully, that each member has the same greater than 0.5 chance of making a correct judgment on a given issue, then the majority has a greater chance still of making a correct judgment and this chance approaches 1 as the size of the community increases. Here too there is safety in numbers.[4]

2.2 A formal presentation

We assume that there are two possible states of the world:

$X = 1$ (e.g. 'guilty');
$X = 0$ (e.g. 'not guilty').

For simplicity, we assume that each of these two states has an equal prior probability of 0.5, but this assumption plays only a minor role for most of our argument. Consider a jury with n members, labelled 1, 2, …, n, where $n > 1$. Each juror i makes a *private judgment* J_i about the state of the world. The private judgment J_i takes the value 0 or 1. The judgment J_i is said to be *correct* if and only if it coincides with the state of the world X. Following Condorcet's original framework, we assume:

Competence. For each x (0 or 1), each juror i has a probability $p > 0.5$ of making the private judgment $J_i = x$, given that the state of the world is $X = x$.

The parameter p is interpreted as the *individual competence level* of juror i. By our competence assumption, all jurors have the same individual competence level.

Independence of private judgments. The private judgments of the n jurors, J_1, …, J_n, are independent from each other, given the state of the world.

Private judgment voting. Each juror i submits the private judgment J_i as his or her vote, denoted V_i, i.e. $V_i = J_i$.

The last two conditions jointly entail that the votes of the n jurors, V_1, …, V_n, like the private judgments, are independent from each other, given the state of the world.

Suppose the votes V_1, …, V_n are aggregated by simple majority voting. Specifically, let $V = V_1 + \ldots + V_n$ be the total number of votes for '$X = 1$'. Then $V > n/2$ means that there is a majority of votes for '$X = 1$' (e.g. for 'guilty'), and $V < n/2$ means that there is a majority of votes for '$X = 0$' (e.g. for 'not guilty').

Let P_n be the probability that there will be a majority of votes for '$X = 1$' among the n jurors, given that the state of the world is $X = 1$. Then P_n is also equal to the probability that there will be a majority of votes for '$X = 0$' among the n jurors, given that the state of the world is $X = 0$. The parameter P_n can be interpreted as the *collective competence level* of the n-member jury. We can now state the jury theorem formally:

Condorcet jury theorem.

- If n is odd (ruling out majority ties), the n-member jury is collectively more competent than each individual juror, i.e. $P_n > p$.
- The jury's collective competence level P_n converges to 1 as the number of jurors n tends to infinity.

3. The core problem

3.1 An informal treatment

The jury theorem is good epistemic news, identifying a prospect whereby a group of individuals might be collectively better at tracking the truth on a given issue than any

of the group members. But the theorem will apply only if every group member does his or her bit. All members must form a judgment independently of the judgments of others, letting that judgment reflect their individual competence level. They must be willing to go with their own private judgment when they cast their vote. They may have deliberated together and listened to the evidence and argument produced by others but, in the end, they must go their own epistemic way.

To put this otherwise, the good results predicted by the jury theorem are an aggregate effect of individual efforts to be epistemically independent. When the group members are independent in this way, then they will each individually have a 0.6 chance of making a correct judgment, even when the votes of the other group members are given. That is, the conditional probability of each making a correct judgment, given how others vote, will still be 0.6. This is analogous to our coin example. No matter how the other coin throws turn out, the chance of the coin coming up heads on any one throw will always be 0.6. Even if, against the odds, the coin has come up tails ten times in a row, the chance of its coming up tails on the next throw will still be only 0.4. Similarly, regardless of how others vote, the probability, conditional or unconditional, of any one individual's making a correct judgment on the issue under consideration will remain fixed at 0.6.

We can express these points more formally by referring to the degree of support that the votes of the individuals and the group give to the hypothesis that things are, or are not, how the individuals or the group judge them to be. Call this hypothesis *H*. The unconditional degree of support that an individual vote gives to *H* can be interpreted as the warrant that that individual's voting in a given way provides for believing in *H*. Likewise, the unconditional degree of support that the group's vote gives to *H* can be interpreted as the warrant that a given voting pattern across that group as a whole provides for believing in *H*.

Like the concept of probability, the concept of degree of support comes in an unconditional and a conditional form. We can consider not only the unconditional degree of support that an individual's vote gives to *H*, but also the conditional degree of support that the vote gives to *H*, given the votes of the other individuals. The conditional degree

of support can be interpreted as the additional warrant that the individual's vote provides for believing in *H*, given that we have already observed the votes of the other individuals.

When the votes of different individuals are independent from each other, the conditional and unconditional degrees of support take the same value: the conditional degree of support associated with each individual's vote, given the votes of the others, is equal to the unconditional degree of support that the vote provides. This implies that, when the group members' judgments are independent from each other, then the unconditional degree of support that the group's vote as a whole gives to *H* is simply the sum of the unconditional degrees of support that the individual votes each give to *H*. So the degree of support provided by each individual vote contributes in full measure to the degree of support given to *H* by how the group as a whole votes. And the degree of support given to *H* by the voting pattern across the group is an increasing function of the number of individuals voting in support of *H*. But, as we will see, when independence is violated, the conditional and unconditional degrees of support can come apart.

Now we are in a position to see how popularization of the Condorcet jury theorem may undermine the theorem's conditions and in some cases put the result in jeopardy.[5] Suppose that the group members become aware of the theorem, recognizing that the chance of the majority making a correct judgment is greater than the chance of any one individual making a correct judgment. There are several kinds of conditions under which this awareness can feed back onto the individuals' behaviour, impacting negatively on the conditions of the Condorcet jury theorem. We focus on three such sets of conditions, addressing the most benign one first and turning to the less benign ones next.

3.1.1 Updating in truthful deliberation

Suppose that prior to taking a vote on the issue in question, the group engages in collective deliberation and the individuals all reveal their private judgments to each other. As before, we assume that each individual has a greater than 0.5 chance of making a correct private judgment, and that the private judgments of different individuals are independent from each other, given the relevant fact of the matter. For the moment, we also assume that the individuals all reveal their private judgments truthfully. This is a crucial assumption; below we discuss the case where it is relaxed.[6]

In the scenario presented the individuals can each observe the pattern of private judgments across the group as whole, and they each have the opportunity to revise their own judgment when a vote is subsequently taken. So there are two stages. At the first (deliberation) stage, the individuals reveal their private judgments to each other; at the second (voting) stage, they each submit a vote. We further assume that, at the voting stage, each individual cares about submitting a correct judgment as his or her vote, even if this may involve a change from his or her original private judgment.

Suppose I am one of the individuals and have observed the pattern of private judgments across the group. I believe that these judgments satisfy the conditions of the Condorcet jury theorem, as assumed above. Should I stick to my own private judgment when submitting my vote, or should I do something else?

If the majority of private judgments conincides with my own private judgment, this should reinforce my belief that I have made a correct judgment, and so I should stick to it. But if the majority of private judgments is different from my own one, then my understanding of the Condorcet jury theorem should lead me to reason as follows. There is a much better chance that the majority will have made a correct judgment than that I will have made a correct judgment. In terms of degree of support: if the majority judges that *H* is true while I privately judge that *H* is false, then the majority gives a much greater degree of support to *H* than my own private judgment gives to the negation of *H*. So I should update my judgment, following the majority opinion. So in either case – whether or not the majority opinion coincides with my own private judgment – I should vote in accordance with the majority opinion. My own private judgment disappears from the scene at this point. Notice, however, that in the present scenario my own private judgment will have made its epistemic contribution, as it will have made a contribution to the majority opinion at the stage of group deliberation.

This reasoning is symmetrical across the group. Under the conditions outlined, all individuals will engage in the same reasoning and, at the voting stage, they will each submit a vote according to the majority opinion that was established at the deliberation stage. The effect of this is unanimity at the voting stage. Whichever opinion commands a majority at the deliberation stage will become the unanimity opinion at the voting stage.[7]

Although the votes, unlike the private judgments, do not satisfy the independence requirement of the jury theorem, the theorem's main prediction is not undermined here. The probability that the group unanimously makes a correct judgment at the voting stage will simply equal the probability that the majority makes a correct judgment at the deliberation stage. By the Condorcet jury theorem, the latter probability is greater than the probability that any one individual makes a correct private judgment, and converges to one as the group size increases. So, likewise, the former probability – that the group unanimously makes a correct judgment at the voting stage – will have these properties.

Here reflexivity will undermine Condorcet's assumptions that individuals vote their private judgments and that different individuals' votes, as opposed to private judgments, are independent from each other. But, interestingly, it will not undermine Condorcet's conclusion: the group's collective competence will still be greater than the competence of each individual member, and will still approach one as the group size increases. This is because the individuals will each have applied the majority calculus in their own reasoning, making their voting decision by identifying the majority opinion among all individuals' private judgments, including, crucially, their own one.

Reflexivity will, however, undermine another prediction of Condorcet's framework. In Condorcet's original case, where all individuals vote their own private judgments, the degree of support given to the hypothesis that things are as the majority says they are is an increasing function of the majority size. This means that information about the majority size also gives us information about the degree of support this majority gives to the hypothesis in question. But in the present case, where the majority opinion at the deliberation stage is turned into a unanimous opinion at the voting stage, this fine-grained information gets lost. Regardless of whether the original majority supporting the hypothesis was narrow or broad, we will always obtain a unanimous vote across the group (see also Goodin 2002).[8]

3.1.2 Epistemic free-riding

The case of reflexivity we have just identified is a relatively benign one. While reflexivity undermines some of Condorcet's assumptions, it does not undermine

Condorcet's main conclusion. We will now see that, if the conditions of the previous case are subtly modified, reflexivity can have much less benign consequences.

As before, each individual has a greater than 0.5 chance of making a correct private judgment, and the private judgments of different individuals are independent, given the relevant truth of the matter. Again, prior to taking a vote, the individuals can observe the judgments expressed by (some of) the other individuals. In the previous case, we described this as a deliberation stage. But the assumption can be met in this case without requiring a distinct stage of that sort. The individuals may for example observe some of the others' judgments if voting takes the form of an open ballot. The judgments are expressed by a show of hands, and an individual can take a moment to observe how many others raise their hands before deciding whether or not to raise his or her own hand too. The most important difference from the previous case is that we do not assume that the individuals will always reveal their private judgments truthfully. They may of course do so, but we will not take this for granted.

Suppose that one or more individuals satisfy the following conditions:

- however much they care for the group's making a correct judgment, they each have a motive for wanting to be correct in the judgments they individually express;
- they each believe that others express their private judgments truthfully.[9]

Imagine that these conditions are fulfilled in my case. I have observed the judgments expressed by the others and have to decide what judgment to express myself. As I assume that the others' judgments are all independent as required by the jury theorem, I come to believe that there is a better chance that the majority will have made a correct judgment on the issue in hand than that I will have individually made a correct judgment. Thus I stand a better chance of expressing a correct judgment by expressing the same judgment as the majority does than by expressing my own private judgment. So I should withhold my private judgment and go along with the majority; and when I submit my vote I should vote as I expect the majority to vote, rather than as my private judgment would lead me to vote.

Can conditions like these be fulfilled? Given the idealizing assumption that the members of a group have an equal, greater than 0.5 chance of making a correct private judgment on a particular issue, we think they can. There are many possible reasons why I might want to express a correct judgment myself, ranging from the case where I want for selfish reasons not to appear to be wrong on some issue – say, the case where my reputation or perhaps even my remuneration as a group member depends on my track record at making correct judgments – to the case where I have commendable motives for wanting to get the issue right: say, the case where as a member of a scientific community I want to communicate the truth to my students or colleagues, or I want to organize my research around sound assumptions. In many such cases I may have good reason to believe that others are unlikely to think as I do, even if they are aware of Condorcet's result; and that they are likely to come down in majority support – a majority that forms independently of my own private judgment – for one or another position.

So the popularization of the Condorcet theorem may have the consequence that one or more group members will be led to free-ride on the efforts of others. When only one individual free-rides like this, always withholding their own private judgment and expressing the judgment of the majority, then they will raise their individual chance of expressing a correct judgment on the issue, by Condorcet's result. Thus the *unconditional* degree of support that their revised judgment provides for the hypothesis that things are as their judgment suggests will equal the degree of support provided by the majority – which is strictly greater than the unconditional degree of support that their independent private judgment would have provided. But the *conditional* degree of support their updated judgment gives to that hypothesis, given the judgments of the other individuals – i.e. the contribution their revised judgment makes to the overall degree of support provided for that hypothesis by the group's set of judgments – will fall to zero. That marks the free-riding character of their behaviour. Their "judging" as they do will contribute nothing to the warrant that the majority judgment provides for the hypothesis. Moreover, their "judging" as they do might mislead the group into thinking that the warrant is greater than it actually is. If their free-riding is not transparent – that is, if the group assumes that all individuals expressed independent judgments when in fact some were free-riding – then their judgment will suggest, mistakenly, that the warrant is that

corresponding to a majority of *k* out of *n* individuals, when really it is that corresponding to a majority of *k*-1 out of *n*-1 individuals.

But again the reasoning is – at least potentially – symmetrical across the group. If more than one individual free-rides in the manner described, withholding their own private judgment and deferring to the judgment of whatever majority they expect to form, then the consequences can be quite dramatic. Suppose, for example, that votes are taken sequentially; that is, the individuals submit their votes one by one in a particular sequence, and each individual can observe the votes submitted earlier in the sequence. Then the first individual will vote his or her private judgment, since there will not be any majority for this individual to defer to. The second individual will presumably still vote his or her private judgment, since there is no reason to think that the first individual was more likely to have made a correct judgment. But now, beginning with either the third or the fourth individual in the sequence, the temptation to free-ride may arise. After the first three votes have been cast – or, in case the first two individuals agree, after the first two votes have been cast – there will typically be a majority in one or the other direction.[10] Thus any individual whose vote is taken subsequently will be led to think that the preceding majority is more likely to have made a correct judgment than they are. To be precise, any individual will be led to think this if they believe that the individuals earlier in the sequence expressed their private judgments truthfully.[11] So the individual who cares about expressing a correct judgment will withhold their own private judgment and go with the majority of individuals preceding them in the sequence.

In this fashion, any majority that accidentally emerges among the first two or three jurors may grow further and further, suggesting, mistakenly, an increasing degree of support for the hypothesis that things are as this growing majority says they are. But in fact the degree of support for that hypothesis will not increase at all beyond the first two or three individuals. All subsequent votes will simply be the result of epistemic free-riding and not the result of the expression of independent private judgments, and therefore the conditional degree of support these subsequent votes give to the hypothesis in question is zero. It is easy to see that, depending on the particular sequence in which the votes are taken, any voting outcome could in principle emerge like this, and the good news of the jury theorem is undermined quite dramatically. The phenomenon we have

described is sometimes referred to as an informational cascade (Bikhchandani, Hirshleifer and Welch 1992).

3.1.3 A classical free-riding problem?

Under the conditions given, free-riding may occur, but there is nonetheless a disanalogy with the classical free-riding problem of the *n*-person prisoners' dilemma. A central condition in our case was that each individual acts on the assumption that others express their private judgments truthfully. By contrast, a classical free-riding problem requires that it is in an individual's interest to free-ride not only under one specific assumption about how others behave, but under every such assumption – or at least under most of them (Pettit 1986). In short, a classical such problem requires that free-riding is the dominant strategy. But in the present case, while individuals may be tempted to free-ride if they believe that others express independent private judgments, that temptation may decline if they believe that others are free-riding too.[12] Were I to think that everyone in the group was simply expressing the judgment they expected the majority to express – for instance, were I to think that the majority is the result of an informational cascade rather than the result of the agreement among independent private judgments – then I would have no reason to think that going along with the majority would increase the chance of being correct. And so I would cease to free-ride – I would begin to express my independent private judgment – under that assumption.

There is a further, slightly different set of conditions, however, under which something closer to a classical free-riding problem might arise. The individuals each value the group's making a correct judgment on the issue in question, but they do not care too much about making a correct private judgment themselves. Nonetheless, each individual has the capacity to make an independent private judgment satisfying Condorcet's assumptions. If an individual chooses to exercise this capacity, then he or she will have a greater than 0.5 chance of making a correct private judgment, and the private judgments of different individuals will be independent in the relevant manner. However, exercising that capacity – making an independent private judgment – is costly; it requires time and effort.

So, instead of the previous conditions, the following is true of the individuals. Apart from the case where their own vote is pivotal for the group vote, they each have a preference ranking as in a prisoners' dilemma. They each prefer that they should all vote independently than that no one should do so, since they value the group's being right about the issue; and for the same reason they prefer that more others, rather than fewer others, vote independently. But, as they find it burdensome to vote independently themselves, each most wants to be a lone defector – a lone non-independent voter – and least wants to be a lone conformer – that is, the only person to vote independently.

Under these conditions, and absent the prospect of being pivotal, we can see how the members of the group, having learned about Condorcet's result, might each be tempted to free-ride. Each will think that to the extent that others vote independently, there is lesser reason for them to bother doing so: except for the case where they are pivotal, the contribution that their independent voting would make is not going to be large enough to compensate for the effort. And each will think equally that to the extent that others do not vote independently, the same is true: except for the case where they are pivotal, the contribution that their independent voting would make is going to represent an effort wasted.

While this set of conditions would give rise to something closer to a classical free-riding problem, it is probably less likely to be fulfilled than the previous set. Less likely, but still possible. Consider the case where a fairly large committee has to make some collective judgment, for instance on the merit of different candidates for a job or different submissions for a prize. Voting independently in such a case might be burdensome, involving a lot of research and consideration, and yet everyone on the committee might prefer that it get the result right. If everyone puts aside the chance of being the pivotal voter – the larger the group the lower that chance – then each will be tempted to think that if others vote independently, the effort of ensuring independence will not be required; and that if others do not vote independently, the effort of ensuring independence will be wasted. The present free-riding problem is similar to the rational ignorance problems discussed in the literature: it may sometimes be rational to remain ignorant rather than to engage in the costly acquisition of information (Brennan and Lomasky 1993, ch. 7).

3.2 A formal treatment

We begin by introducing the concept of degree of support in general terms, namely in terms of the support some item of evidence gives to some hypothesis. We then apply that concept to Condorcet's framework. First, we discuss the unconditional degree of support an individual juror's vote – for or against '$X = 1$' – gives to the hypothesis that $X = 1$. Secondly, we discuss the unconditional degree of support the vote of the entire jury gives to that hypothesis. Thirdly, we discuss the conditional degree of support an individual juror's vote gives to the hypothesis, given the votes of the other jurors. Finally, we sketch the reflexivity problem.

3.2.1 The concept of degree of support

Let H be some hypothesis and E some item of evidence. We use the notation $Pr(A)$ to denote the unconditional probability of A, and $Pr(A|B)$ to denote the conditional probability of A, given B. For any item of evidence E, the *unconditional degree of support E* gives to H is defined as

$$l(H, E) := log\left(\frac{Pr(E|H)}{Pr(E|\neg H)}\right)$$

(Fitelson 2001). Note some properties of this measure of the degree of support:

- The [degree of support E gives to H] is related to the [probability that H is true, given E] as follows. Let $r = Pr(H)$ be the prior probability that H is true. Then

  $$Pr(H|E) = \frac{r}{r + (1\text{-}r)\ exp(\text{-}l(H, E))}.$$

- The measure also has the following property:

  $$l(H, E) = \begin{cases} > 0 & \text{if } Pr(H|E) > Pr(H) \\ = 0 & \text{if } Pr(H|E) = Pr(H) \\ < 0 & \text{if } Pr(H|E) < Pr(H). \end{cases}$$

  This means: If the measure is positive (respectively negative), then observing the evidence E increases (respectively decreases) our degree of belief in H.

The degree of support can also be defined conditionally. For any two items of evidence E_1 and E_2, the *conditional degree of support E_2 gives to H conditional on E_1* is defined as

$$l(H, E_2 \mid E_1) := log\left(\frac{Pr(E_2|H \wedge E_1)}{Pr(E_2|\neg H \wedge E_1)} \right).$$

The *conditional* degree of support can be interpreted as the additional warrant E_2 gives to H, given that we have already observed E_1.

It can easily be seen that, for any two items of evidence E_1 and E_2, the degree of support the *conjunction* of E_1 and E_2 gives to H is

$$l(H, E_1 \wedge E_2) = l(H, E_1) + l(H, E_2 \mid E_1).$$

Now consider the special case where E_1 and E_2 are independent from each other conditional on H (and on $\neg H$). Such independence implies $Pr(E_2|H \wedge E_1) = Pr(E_2|H)$ and $Pr(E_2|\neg H \wedge E_1) = Pr(E_2|\neg H)$. Therefore we have

$$l(H, E_2 \mid E_1) = l(H, E_2),$$

i.e. [the conditional degree of support E_2 gives to H, given E_1] equals [the unconditional degree of support E_2 gives to H]. Therefore

$$l(H, E_1 \wedge E_2) = l(H, E_1) + l(H, E_2),$$

i.e. the degree of support is additive. The distinction between unconditional and conditional degree of support is crucial for the discussion below.

3.2.2 The unconditional degree of support an individual juror's vote gives to H

Let us restate Condorcet's framework in terms of hypothesis testing (List 2003). We make Condorcet's original assumptions of competence, independence of private judgments, and private judgment voting. Let H denote the hypothesis that $X = 1$ (e.g. that the defendant is guilty). Suppose we want to test H. Each juror's vote V_i is a potential item of evidence relevant to H. What degree of support does each such vote give to H?

Suppose our evidence is that juror i has voted for '$X = 1$'. We have

$$l(X = 1, V_i = 1) = \log(p / (1\text{-}p)). \qquad (*)$$

Thus $l(X = 1, V_i = 1)$ is greater than 0, so long as $p > 0.5$. This makes sense from an intuitive perspective: observing that a juror with competence $p > 0.5$ has voted for '$X = 1$' makes $X = 1$ more likely to be true. Moreover, by our assumption that the prior

probability of each state of the world is 0.5 (i.e. $r = 0.5$), the posterior probability that $X = 1$, given an individual juror's vote for '$X = 1$' is p.[13]

Next, suppose our evidence is that juror i has voted for '$X = 0$'. We have

$$l(X = 1, V_i = 0) = \log((1-p) / p). \qquad (**)$$

Then $l(X = 1, V_i = 0)$ is less than 0, so long as $p > 0.5$. Again, this makes intuitive sense: observing that a juror with competence $p > 0.5$ has voted for '$X = 0$' makes $X = 1$ less likely to be true. Here, by our assumption that the prior probability of each state of the world is 0.5, the posterior probability that $X = 0$, given an individual juror's vote for '$X = 0$' is p.[14]

3.2.3 The degree of support the vote of the jury collectively gives to H

What about the degree of support that the voting pattern across all jurors gives to H? Again consider the situation where all of Condorcet's original conditions are satisfied. Suppose our evidence E is that precisely h out of n jurors have voted for '$X = 1$'. For simplicity, suppose the first h jurors have voted for '$X = 1$', while the remaining n-h jurors have voted for '$X = 0$'; that is, we consider the evidence

$$E = (V_1 = 1 \wedge V_2 = 1 \wedge \ldots \wedge V_h = 1) \wedge (V_{h+1} = 0 \wedge V_{h+2} = 0 \wedge \ldots \wedge V_n = 0).$$

By Condorcet's conditions of independence of private judgments and private judgment voting, the votes of different jurors are independent from each other conditional on the state of the world X. Thus the degree of support is additive, and we have:

$$l(X = 1, E) = m \log(p / (1-p)),$$[15]

where $m = h-(n-h)$ is the absolute margin between the number of votes for '$X = 1$' and the number of votes against '$X = 0$'.[16] The same result holds for any other set of h out of n jurors; the assumption that precisely the first h jurors voted for '$X = 1$' is no loss of generality here. Note the following:

- If there are more votes for '$X = 1$' than for '$X = 0$', then the jury's voting pattern supports $X = 1$.
- If there are more votes for '$X = 0$' than for '$X = 1$', then the jury's voting pattern supports $X = 0$.

- In both cases, the greater the difference between the number of 'guilty' and 'not guilty' votes, the greater the relevant degree of support.

By our assumption that the two possible states of the world have the same prior probability, the posterior probability that $X = 1$ (respectively, that $X = 0$), given the evidence that precisely h out of n jurors have voted for '$X = 1$' (respectively, for '$X = 0$'), is $p^m/(p^m+(1-p)^m)$, where m is as defined above. If the given evidence is only that a majority among the n jurors have voted for '$X = 1$' (respectively for '$X = 0$'), but the evidence does not include any information about the precise size of that majority, then the posterior probability that $X = 1$ (respectively, that $X = 0$) is P_n, i.e. equal to the collective competence level (List 2003).

3.2.4 The conditional degree of support an individual juror's vote gives to H, given the votes of the other jurors

Suppose that some, but not all, of the jurors – up to n-1 of them – have cast their votes. Let E denote the evidence constituted by the particular voting pattern across these jurors. Suppose juror i is not among those jurors who have already cast their votes. What is the *conditional* degree of support juror i's vote gives to H, given E? Suppose juror i votes for '$X = 1$'. Then the conditional degree of support juror i's vote gives to H, given E, is $l(X = 1, V_i = 1 \mid E)$. If Condorcet's conditions of independence of private judgments and private judgment voting are met, we have:

$$l(X = 1, V_i = 1 \mid E) = l(X = 1, V_i = 1) = \log(p / (1-p)).$$

An analogous result holds when juror i votes for '$X = 0$'.

$$l(X = 1, V_i = 0 \mid E) = l(X = 1, V_i = 0) = \log((1-p) / p).$$

In short, under independence of private judgments and private judgment voting, [the conditional degree of support juror i's vote gives to H, given the other jurors' votes] is equal to [the unconditional degree of support that juror's vote gives to H].

This fact is responsible for the power of the original jury theorem: under independence of private judgments and private judgment voting, the support different jurors' votes give to H is additive. There are no diminishing marginal returns on adding further jurors. Supposing that the competence condition is also satisfied, adding more and

more jurors typically increases the degree of support the jury's voting pattern gives to the hypothesis that things are as the majority says they are, and there is in principle no upper bound to this increase.

3.2.5 Updating in truthful deliberation

As before, suppose the n jurors satisfy the conditions of competence and independence of private judgments. But, unlike before, we now suppose that there are two stages. In the first (deliberation) stage, the jurors reveal their private judgments to each other. We assume that they all reveal these judgments truthfully.[17] In the second (voting) stage they each submit a vote. A juror's vote may be either his or her own private judgment or a revised judgment, where the revision might be based on the juror's observation of the pattern of private judgments across the jury in the first stage. We assume that each juror cares about submitting a correct judgment as his or her vote, even if this requires deviating from his or her own private judgment.

Suppose juror i observes the private judgments across the entire jury, as revealed in the first stage, including juror i's own private judgment. This allows juror i to determine whether or not a majority of private judgments supports H, i.e. whether or not $J > n/2$, where $J = J_1 + \ldots + J_n$. For simplicity, assume that n is odd; so there is never a majority tie.

If juror i understands the Condorcet jury theorem, he or she will realize that the probability that the majority of private judgments across the jurors coincides with the state of the world, P_n, exceeds his or her own individual competence, p. The degree of support the majority judgment gives to H (or $\neg H$, depending on how the majority goes) exceeds the degree of support juror i's private judgment individually gives to H (or $\neg H$).

As we have assumed, at the second stage, juror i cares about submitting a correct judgment as his or her vote. Consider the following two voting strategies:

The private judgment strategy. Vote for '$X = 1$' (i.e. $V_i := 1$) *if and only if* $J_i = 1$.

The updating strategy. Vote for '$X = 1$' (i.e. $V_i := 1$) *if and only if* $J > n/2$.

The private judgment strategy is the one assumed in Condorcet's original result. What are the differences between the two strategies?

- On the private judgment strategy, juror i's probability of submitting a correct judgment as a vote is
  $Pr(V_i = 1 \mid X = 1) = Pr(V_i = 0 \mid X = 0) = p$,
  as in Condorcet's original framework.
- On the updating strategy, juror i's probability of submitting a correct judgment as a vote is
  $Pr(V_i = 1 \mid X = 1) = Pr(V_i = 0 \mid X = 0) = P_n$,
  which, by the jury theorem, is greater than p and can be arbitrarily close to 1 when the jury size n is large.

If juror i wants submit a correct judgment as a vote with a high probability, he or she will adopt the updating strategy. Moreover, a rational juror should be moved by the epistemic force of the majority judgment. While a single private judgment in support of H (respectively $\neg H$) warrants only a degree of belief of p in H (respectively $\neg H$), a majority judgment in support of H (respectively $\neg H$) warrants a degree of belief of P_n in H (respectively $\neg H$),[18] which by the jury theorem is greater than p.

This reasoning is symmetrical across the n jurors, and they will therefore all adopt the updating strategy. We obtain the following result:

- Whenever a majority of private judgments supports the hypothesis that $X = 1$ at the first stage, all jurors will unanimously vote for '$X = 1$' at the second stage, i.e. if $J > n/2$ then $V = n$.
- Whenever a majority of private judgments supports the hypothesis that $X = 0$ at the first stage, all jurors will unanimously vote for '$X = 0$' at the second stage, i.e. if $J < n/2$ then $V = 0$.

So the original jury theorem, as stated above, continues to hold. This is because the majoritarian aggregation has been performed in the individual reasoning of every juror. Any single vote that results from a juror's application of the updating strategy will carry the same epistemic weight that the pattern of private judgments collectively carried at the first stage. The jurors will all speak with the same voice at the second stage. The unconditional degree of support a single juror's vote gives to H (or $\neg H$) at the second

stage equals the degree of support the jurors' private judgments collectively give to that hypothesis at the first stage, i.e. for each juror i,

$$l(X = 1, V_i = 1) = l(X = 1, J > n/2) = \log(P_n/(1\text{-}P_n))$$
$$= l(X = 0, V_i = 0) = l(X = 0, J < n/2) = \log(P_n/(1\text{-}P_n)),$$

where $P_n = Pr(J > n/2 \mid X = 1) = Pr(J < n/2 \mid X = 0)$, as in Condorcet's original framework. So additional votes are redundant once all jurors have updated their judgments. As soon as the vote of a single juror is given, the conditional degree of support given to H (or $\neg H$) by every additional vote is zero. Formally, for any i and j (where $i \neq j$), we have

$$l(X = 1, V_j = 1 \mid V_i = 1) = l(X = 0, V_j = 0 \mid V_i = 0) = 0,$$

since, after every juror's updating of their judgment, $Pr(V_j = 1 \mid V_i = 1) = Pr(V_j = 0 \mid V_i = 0) = 1$.

In the present scenario, while Condorcet's conditions of competence and independence of private judgments are satisfied, the condition of private judgment voting is undermined by reflexivity, as is Condorcet's prediction that the votes themselves, and not only the private judgments, are independent, given the state of the world. However, the effect of reflexivity is benign here, in that the central conclusion of the jury theorem continues to hold, as we have seen, albeit via a somewhat different mechanism.

However, what is lost at the second stage, after the updating, is the fine-grained information about the size of the majority for $X = 1$ (or for $X = 0$, as the case may be) among the private judgments. In Condorcet's original framework this information allows us to determine that the degree of support given to H by a majority of h out of n votes is precisely $m\log(p / (1\text{-}p))$, where $m = h\text{-}(n\text{-}h)$. In the present case, on the other hand, we can only determine, in a more coarse-grained way, that the degree of support given to H by a unanimous vote for H is $\log(P_n/(1\text{-}P_n))$.

3.2.6 The free-riding problem

We now subtly modify the conditions of the previous case. We still suppose that the jurors satisfy competence and independence of private judgments. But we change our assumption on what information each juror has about the other jurors' judgments. We

now suppose that, prior to taking a vote, the jurors may observe the judgments expressed by some, but not necessarily all, of the other jurors. A case of particular interest here is the one of sequential voting: the jurors cast their votes by a show of hands one by one in a sequence, and each juror can observe the votes cast by the jurors preceding him or her in that sequence. Unlike before, however, we do not assume that the jurors will always express their private judgments truthfully; they might do so, but we do not presuppose this. But we continue to assume that each juror cares about submitting a correct judgment as his or her vote.

Our question is whether or not jurors will have an incentive to express their private judgments truthfully. Consider the sequential case just described. The jurors reveal their judgments – cast their votes – in a sequence, say 1, 2, 3 up to n. Suppose that jurors 1 up to i-1 have cast their votes, $V_1, \ldots, V_{i-1}$, and suppose, for the moment, that they have each voted their private judgments truthfully, i.e. $V_1 = J_1, \ldots, V_{i-1} = J_{i-1}$. Suppose it is now juror i's turn to cast a vote, V_i, and juror i acts on the assumption that the votes cast by jurors 1 up to i-1 are these jurors' truthful private judgments. To make things simple, we assume that i is an odd number.

So juror i knows the judgments (by assumption, the true private judgments) of jurors 1 up to i-1; and, moreover, juror i knows his or her own private judgment. This allows juror i to determine whether or not a majority among the judgments of jurors 1 up to i (including his or her own private judgment) supports $X = 1$, i.e. whether or not $V_1+\ldots+V_{i-1}+J_i > i/2$.

Unless juror i is the first or second juror in the sequence, the understanding of the Condorcet jury theorem will lead him or her to reason as follows. Assuming that the other jurors have revealed their private judgments truthfully, the probability, P_i, that the majority among the i private judgments up to juror i coincides with the state of the world exceeds juror i's own individual competence, p. In degree of support terms, the degree of support the majority judgment among jurors 1 up to i gives to H (or $\neg H$, depending on how that majority goes) exceeds the degree of support juror i's private judgment individually gives to H (or $\neg H$).

So, as before, juror i may be tempted to adopt an updating strategy:

The updating strategy. Vote for '$X = 1$' (i.e. $V_i := 1$) *if and only if* $V_1+\ldots+V_{i-1}+J_i > i/2$.

Again, contrast this with the private judgment strategy, introduced above:

The private judgment strategy. Vote for '$X = 1$' (i.e. $V_i := 1$) *if and only if* $J_i = 1$.

What are the differences between the two strategies in the present context? Note that the assumption that jurors 1 up to i-1 have revealed their private judgments truthfully is crucial here.

A: The implications for juror *i*'s probability of submitting a correct judgment as a vote

- On the private judgment strategy, juror i's probability of submitting a correct judgment as a vote is
  $Pr(V_i = 1 \mid X = 1) = Pr(V_i = 0 \mid X = 0) = p$,
  as in Condorcet's original framework.
- On the updating strategy, juror i's probability of submitting a correct judgment as a vote is
  $Pr(V_i = 1 \mid X = 1) = Pr(V_i = 0 \mid X = 0)$
  $= Pr(V_1+\ldots+V_{i-1}+J_i > i/2 \mid X = 1) = Pr(V_1+\ldots+V_{i-1}+J_i < i/2 \mid X = 0) = P_i$,
  which, by the jury theorem, is greater than p and can be arbitrarily close to 1 when i is large.

So, if juror i gets certain private benefits for exhibiting (what appears to be) a high competence level in his or her voting pattern (e.g. a reward for his or her track record at voting correctly), then the juror has an incentive to adopt the updating strategy rather than the private judgment strategy. Moreover, as before, if juror i is rational, he or she should be moved by the epistemic force of the majority judgment among jurors 1 up to i: while juror i's private judgment in support of H (respectively $\neg H$) warrants only a degree of belief of p in H (respectively $\neg H$), the majority judgment among jurors 1 up to i in support of H (respectively $\neg H$) warrants a degree of belief of P_i in H (respectively $\neg H$),[19] which by the jury theorem is greater than p – assuming that jurors 1 up to i-1 have indeed truthfully revealed their independent private judgments.

B: The implications for the unconditional degree of support juror *i*'s vote gives to *H*

- On the private judgment strategy, as shown in section 3.2.2, the unconditional degree of support juror *i*'s vote gives to *H* or ¬*H* is
  $l(X = 1, V_i = 1) = l(X = 0, V_i = 0) = \log(p / (1\text{-}p))$.
- On the updating strategy, the unconditional degree of support juror *i*'s vote gives to *H* or ¬*H* is
  $l(X = 1, V_i = 1) = l(X = 0, V_i = 0) = \log(P_i/(1\text{-}P_i))$,
  which, by the jury theorem, is greater than $\log(p/(1\text{-}p))$ because $P_i > p$.

C: The implications for the conditional degree of support juror *i*'s vote gives to *H*, given the votes of jurors 1 up to *i*-1

- On the private judgment strategy, as shown in section 3.2.4, the conditional degree of support juror *i*'s vote gives to *H* or ¬*H*, given the votes of jurors 1 up to *i*-1, is
  $l(X = 1, V_i = 1 \mid E) = l(X = 0, V_i = 0 \mid E) = l(X = 1, V_i = 1) = l(X = 0, V_i = 0)$
  $= \log(p / (1\text{-}p))$,
  where *E* refers to the voting pattern across jurors 1 up to *i*-1.
- On the updating strategy, the conditional degree of support juror *i*'s vote gives to *H* or ¬*H*, given the votes of jurors 1 up to *i*-1, is
  $l(X = 1, V_i = 1 \mid E) = l(X = 0, V_i = 0 \mid E) = 0$,
  supposing that juror *i* is not pivotal for the majority among jurors 1 up to *i*, i.e. supposing that jurors 1 up to *i*-1 are not tied between '$X = 1$' and '$X = 0$'. This result is presented in more detail in the appendix.

Table 1 summarizes implications B and C.

Juror *i*'s strategy	The *unconditional* degree of support juror *i*'s vote for '$X = 1$' gives to the hypothesis that $X = 1$	The *conditional* degree of support juror *i*'s vote for '$X = 1$' gives to the hypothesis that $X = 1$, given the votes of jurors 1, ..., *i*-1
The private judgment strategy	$\log(p / (1\text{-}p))$ (1)	$\log(p / (1\text{-}p))$ (2)
The updating strategy	$\log(P_i/(1\text{-}P_i))$ (3)	0 (supposing that juror *i* is not pivotal)(4)

Table 1

Note that, for any odd number *i* greater than 1, the value in box (3) is greater than the one in box (1), while the value in box (4) is smaller than the one in box (2). This means that, under the updating strategy, juror *i*'s individual (unconditional) epistemic performance will be better than under the private judgment strategy, but his or her epistemic contribution to the collectivity will drop down to 0 (supposing that juror *i* is not pivotal for the majority judgment). Crucially, all of this depends on the assumption that jurors 1 up to *i*-1 reveal their private judgments truthfully.[20]

This means that, if all jurors act on the assumptions described in this section, jurors beyond the first few will typically not make an additional epistemic contribution. We can imagine a case where there are 10 jurors, each of whom has a competence level of 0.6. Let us suppose the state of the world is $X = 1$ and the pattern of private judgments across these jurors is as shown in table 2.

$i =$	1	2	3	4	5	6	7	8	9	10
$J_i =$	1	0	0	1	1	0	1	1	1	0

Table 2

Under the idealized behavioural assumptions of this section – which, as we haved noted, may not hold in equilibrium – the order in which the jurors reveal their judgments – cast their votes – may crucially affect the outcome.

If the order in which votes are taken is 1, 4, 5, 2, 7, 3, 8, 9, 10, 6, then a majority for '$X = 1$' will form after the first few jurors and be supported by further jurors acting on the updating strategy. If the order is 2, 3, 6, 1, 4, 5, 7, 8, 9, 10, then a majority for '$X = 0$' will form after the first few jurors and be supported by further jurors acting on the updating strategy.

So the dilemma is that the updating strategy, while apparently individually rational under certain conditions once Condorcet's assumptions are understood and accepted, will undermine Condorcet's happy result at the collective level.

4. Beyond the problem

How serious are the reflexivity problems discussed here? We have seen that the effect of updating in truthful deliberation is relatively benign, especially when it is transparent that such updating takes place only at the voting stage, after the individuals have truthfully expressed their private judgments at the deliberation stage. The problems of free-riding, by contrast, are potentially more serious; and, as the differences between conditions leading to updating in truthful deliberation and ones leading to free-riding are only subtle, there may be a slippery slope between the more benign cases and the more serious ones.

Are there any escape routes from the reflexivity problems? We consider two possible such routes. The first is what we think of as a happy flaw in our character as reasoners, the second a particular institutional design.

4.1 A happy flaw

The problems characterized will arise only if one or more group members are indeed disposed to update their own judgments on learning the judgments of the others. The individuals will be so disposed only if they believe that all group members have a competence level that is sufficiently high to sustain Condorcet's result, so that the majority is more likely to make a correct judgment than any one individual. If they each think that others have the same competence level as they have, and that that competence level is greater than 0.5, then Condorcet's result will certainly apply. This is the condition under which the individuals, if they know about the jury theorem, may be inclined to update their judgments on learning the judgments of the others, and under which a reflexivity effect – either of the benign kind or of the more serious kind, depending on the scenario in question – may arise.

But here is where the happy flaw may cut in. Although, at first, the individuals each ascribe the same competence level to all others on a given issue, they may nonetheless find it difficult to give up their own view on that issue when the majority judgment conflicts with their own private judgment. They may find it more difficult to give up their own view than to revise the belief that others have the same competence level as they do. If this is so, then they may be disposed, in cases of conflict between

majority opinion and their own one, not to revise their own judgment but rather to revise their belief that others have the same competence level as they do.

This phenomenon would be a rational flaw. If I assign the same competence level to others as to myself on a certain issue, then I do so, not conditional on a majority of others agreeing with me, but rather unconditionally. Thus I should not revise the competence assignment to others in the event of finding that I am in the minority on a given question. We hypothesise, however, that in the practice of responding to being in the minority, I may find it difficult to follow Condorcet's theorem – a difficulty I may not have fully anticipated.

It is one thing to acknowledge that others are as likely as I am to make a correct judgment, in ignorance as to what their judgments are; here I may just look at our respective information, training, talent, track record, and the like. It is another to stick with this belief on finding that most others disagree with me, in a situation where Condorcet's logic would require me to change my own judgment on a given issue. I will naturally assess those others in the light of my own judgment on the issue in question and, given my natural disposition to stick with my own judgment – that represents, after all, how things seem to me to be – it will be tempting to decide that actually not all of those in the majority have as high a level of competence as I do.[21]

Thomas Hobbes (1991) supports this hypothesis about the difficulty of thinking that others are just as likely as we are to be right on some question, at least when their views differ from ours. He stresses that, while each of us will be conscious in such a case of what moves us, and why, we will not have that same intimate access to the reasoning of others:

> "such is the nature of men, that howsoever they may acknowledge many others to be more witty, or more eloquent, or more learned; Yet they will hardly believe there be many so wise as themselves: For they see their own wit to hand, and other men's at a distance." (Hobbes 1991, p. 87)

If this line of thought is right, then the reflexivity problems discussed may be less troublesome than they might appear at first. At the point where I am considering whether to update my judgment, I may find that I am unable to make the kind of competence assignment required by the jury theorem, unable to think therefore that the majority has a

chance of making a correct judgment that is greater than mine, and so unable to be relaxed about not voting independently.

In our first scenario – the one in which individuals engage in truthful deliberation prior to voting – the happy flaw may prevent unanimity from arising at the voting stage, even when a clear majority forms at the deliberation stage.[22] It is of course debatable whether updating in truthful deliberation is a desirable effect or not. But if updating is regarded as a desirable effect, as on many accounts, then what we described as a happy flaw may actually turn into an unhappy obstacle. In our second scenario – the one in which individuals care about being individually correct – the happy flaw may lead individuals to stick to their own private judgment, as they would have greater confidence in their own private judgment than in the judgment of the majority. In the third scenario – the one in which individuals care primarily about the group being collectively correct – it might mean, among other effects, that individuals do not necessarily prefer the case where all of the others judge independently to the case where none of them does so. If I am one of the individuals, I might wish not only to stick to my own private judgment, but perhaps even to persuade the others to update their judgments according to mine. In that scenario, of course, there may be a different motive for free-riding, associated with the group size and the small chance that my vote will make a difference, but there will not be a free-riding problem of a kind that stems specifically from people's awareness of Condorcet's result.

4.2 An institutional design strategy

Is there any way of guarding against the problems mentioned, short of relying on this facet of human nature to block their appearance? One way around the problems might be to ensure that no group member knows the judgments of the others. This would require voting to occur all at once, or to be veiled by secrecy procedures, without prior revelation of any private judgments among the group members. Would it mean eliminating all group deliberation, in so far as individuals might reveal their judgments by how they reason? It would certainly require eliminating the sort of deliberation that reveals private judgments but it might tolerate the sort that doesn't. This might consist, for example, in each person putting on the table considerations that they think may have

been ignored by others, whether or not they themselves assign a high weight to those considerations.

It is important to notice, though, that this institutional strategy will not always be desirable. To the extent that our first scenario – updating in truthful deliberation – might be seen as an attractive one, the institutional strategy would block not only reflexivity of the worrisome kind – the one that undermines the main prediction of the Condorcet jury theorem – but also reflexivity of the more benign, deliberative kind. Deliberative democrats might therefore argue that the institutional strategy requires us to pay an unacceptably high price for the avoidance of free-riding.

The second point to notice is that the institutional strategy will not always be feasible with the problem in our second scenario – the one in which individuals care about being individually correct – or not at least in the version of that scenario that we illustrated by reference to a community of scientists. The reason is that in this sort of case, people do not make their judgments all at once, as they might do in a collective body like a committee, and, once made, their judgments are not readily capable of being shrouded from others. The example of a scientific community is perhaps closest to our case of sequential voting, where the individuals express their judgments one by one in a sequence, and where each individual can observe the judgments expressed by those individuals who come earlier in the sequence. Thus, after a certain point, the individuals will be in a good position to tell what the majority is likely to think and they will therefore be exposed to the temptation to go with the majority view.

But institutional veiling might work in the third scenario – the one in which individuals care primarily about the group being collectively correct. Since that scenario involves taking a vote at a fixed point in time, veiling of the vote would be feasible and it would deny all group members a knowledge of where the majority vote is likely to lead. For committees of a certain size it might be a very useful check. But again, it will not come without a price. In requiring the absence of the sort of interpersonal deliberation that reveals private judgments, it may remove the check provided by the individuals' interrogating and testing one another's judgments. And in doing this, particularly with relatively large committees, it may give rise to the distinct free-riding problem associated

with members coming to think of their votes as insignificant. No one's vote will be likely to make a difference in a large committee and, absent the need to deliberate publicly about their voting intentions, no one will find it significant in social terms either (Brennan and Pettit 1990; Brennan and Lomasky 1993). Thus individuals may be inclined to vote without serious thought or reflection. Again, this is a situation of rational ignorance (Brennan and Lomasky 1993, ch. 7).

However, while veiling the vote may block those reflexivity problems that arise from individuals' updating their judgments based on what they learn about the judgments of others, it will not always block a related, but somewhat different kind of reflexivity problem that has received attention in the literature, namely one that arises from particular strategic considerations. The problem can be illustrated by the case where a jury is required to be unanimous about a guilty verdict, as analysed by Fedderson and Pesendorfer (1998).

Under their analysis, Condorcet's finding may lead each juror to reason as follows. Suppose I am one of the jurors and I believe that Condorcet's conditions – competence, independence of private judgments and private judgment voting – are fulfilled by the other jurors. If these others disagree among themselves about the guilt of the defendant, or if they all judge the defendant innocent, then I need not worry about how I vote; my vote will not make any difference, as unanimity is required for a guilty verdict. The only case where my vote will make a difference is the one where all of the others favour a guilty finding. But if all the others favour a guilty verdict, and if the conditions of the jury theorem are satisfied – as by hypothesis they are – then the chance of their being right is much higher than the chance of my being right, whatever my private judgment; moreover, if the competence level of each juror is sufficiently high, then, depending on my threshold of reasonable doubt, I should believe that guilt has been established beyond reasonable doubt. And so I should vote as the others do: that is, guilty. Thus, no matter what others do, voting guilty seems to make sense – at least under the assumption that others vote independently. It will make no difference to the outcome in the cases where others are divided or they think the defendant is innocent. And it will represent the judgment that is likely to be correct beyond reasonable doubt in the case where the others all vote for guilty.

Veiling the votes or blocking deliberation will have no useful effect in this case, since the reasoning just rehearsed – in essence a kind of dominance reasoning – does not require a knowledge of how the majority is likely to go. To reduce the threat of the problem, Feddersen and Pesendorfer advocate the use of special majority voting rather than unanimity rule. However, while the unanimity rule sharpens the strategic structure of the case, even the use of simple majority voting will not in general remove all incentives for misrepresentation of private judgments, as Austen-Smith and Banks (1996) have shown.

Notice that, like our cases of reflexivity, the cases analysed by Feddersen and Pesendorfer and Austen-Smith and Banks arise on the assumption that Condorcet's conditions are fulfilled, but they arise for somewhat different reasons.

5. Conclusion

The problem of reflexivity, as described by Popper and others, is that a theory may cease to be true of a given population – cease to apply there – when it is popularized among members of that population. Those who proclaim the danger of reflexivity haven't often considered it in the context of the sort of theory illustrated by Condorcet's jury theorem. We hope that this paper may show that the problem has a particular resonance in this case.

Consider how the jury theorem might work with a population, not of human agents, but of diagnostic machines. Let these machines operate independently of one another and let them each have the same, greater than 0.5 chance of being right on a certain issue; let them each have a competence level of 0.6. There is no doubt that the Condorcet jury theorem yields a useful and exploitable result in relation to such machines. It provides us with the remarkable assurance that the chance of a majority of the machines being correct in any judgment on that sort of issue is higher than the chance of any individual machine's being correct and that it approaches one as the number of machines in the population increases.

What our paper has shown, we hope, is that no such reassuring result will be as straightforwardly available with a population of human beings, if they come to be aware of the result itself; not, at least, in conditions where the happy flaw fails to operate and

where there is no suitable institutional design in place. Let such awareness materialize and the application of the theorem may be put in jeopardy. Human beings are liable to be tempted to take an epistemic free ride on the efforts of others, thereby violating the condition of mutual independence that the theorem presupposes. And to the extent that they succumb to that temptation they will undermine the application of the theorem in their own case; they will cook the goose that might have laid a golden egg. Their capacity to reflect on the theorem – the very capacity that marks them off from mere machines – is in this respect a disadvantage. Oscar Wilde once quipped that nothing succeeds like excess. Not in this case, alas; not when the excess is an excess of reflection.

References

Austen-Smith, D. and J. S. Banks (1996) "Information aggregation, rationality and the Condorcet jury theorem." American Political Science Review **90**: 34-45.

Austen-Smith, D. and T. Feddersen (2002) "Deliberation and Voting Rules." Working Paper, Kellogg Graduate School of Management, Northwestern University.

Bikhchandani, S., D. Hirshleifer, I. Welch (1992) "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." Journal of Political Economy **100**: 992-1026.

Borland, P. J. (1989) "Majority systems and the Condorcet jury theorem." Statistician **38**: 181-189.

Brennan, G. and L. Lomasky (1993) Democracy and Decision: The Pure Theory of Electoral Preference. Oxford, Oxford University Press.

Brennan, G. and P. Pettit (1990) "Unveiling the Vote." British Journal of Political Science **20**: 311-33.

Condorcet, M. d. (1976) Condorcet: Selected Writings. Indianopolis, Bobbs-Merrill.

Dietrich, F. and C. List (2002) "A Model of Jury Decisions where All Jurors have the Same Evidence." Nuffield College Working Paper in Economics 2002-W23.

Estlund, D. (1994) "Opinion leaders, independence and Condorcet's jury theorem." Theory and Decision **36**: 131-162.

Fedderson, T. and W. Pesendorfer (1998) "Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting." American Political Science Review **92**: 23-35.

Fitelson, B. (2001) "A Bayesian Account of Independent Evidence with Applications." Philosophy of Science **68** (Proceedings): S123-S140.

Grofman, B. G. Owen and S. L. Feld (1983) "Thirteen theorems in search of the truth." Theory and Decision **15**: 261-278.

Harsanyi, J. C. (1976) Essays on Ethics, Social Behavior, and Scientific Explanation. Dordrecht, D. Reidel.

Hobbes, T. (1991) Leviathan. Cambridge, Cambridge University Press.

Lahda, K. (1992) "The Condorcet jury theorem, free speech and correlated votes." American Journal of Political Science **36**: 617-634.

List, C. and R. E. Goodin (2001) "Epistemic Democracy: Generalizing the Condorcet Jury Theorem." The Journal of Political Philosophy **9**: 277-306.

List, C. (2003) "On the Significance of the Absolute Margin." British Journal for the Philosophy of Science, forthcoming.

Nozick, R. (1981) Philosophical Explanations. Oxford, Oxford University Press.

Pettit, P. (1986) "Free Riding and Foul Dealing." Journal of Philosophy **83**: 361-379.

Popper, K. R. (1963). The Poverty of Historicism. London, Routledge and Kegan Paul.

Appendix

Suppose juror i adopts the updating strategy, i.e. $V_i = 1$ *if and only if* $V_1+\ldots+V_{i-1}+J_i > i/2$, where juror i has observed the votes of jurors 1 up to i-1.

Let E denote the voting pattern across jurors 1 up to i-1. The conditional degree of support juror i's vote, say $V_i = 1$, gives to H conditional on E is:

$$l(H, V_i = 1 \mid E) := log\left(\frac{Pr(V_i = 1 \mid H \wedge E)}{Pr(V_i = 1 \mid \neg H \wedge E)}\right).$$

By the updating strategy, we have $V_i = 1$ if and only if $V_1+\ldots+V_{i-1}+J_i > i/2$. Suppose E is the evidence that *precisely* h among jurors 1 up to i-1 have voted for '$X = 1$'. Then juror i's private judgment is *pivotal* if and only if $h = (i-1)/2$, i.e. if and only if jurors 1 up to i-1 are tied between '$X = 1$' and '$X = 0$'.

If juror i's private judgment is pivotal (i.e. $h = (i-1)/2$), then $V_i = 1$ if and only if $J_i = 1$, i.e.

$$Pr(V_i = 1 \mid H \wedge E) = P(J_i = 1 \mid H) = p$$

and $Pr(V_i = 1 \mid \neg H \wedge E) = P(J_i = 1 \mid \neg H) = 1\text{-}p$.

If juror i's private judgment is not pivotal (i.e. $h \neq (i\text{-}1)/2$), then $V_i = 1$ if and only if $h > i/2$, i.e. $V_i = 1$ depends on H (or $\neg H$) only *through* E, and hence

$$Pr(V_i = 1 \mid H \wedge E) = Pr(V_i = 1 \mid E)$$

and $Pr(V_i = 1 \mid \neg H \wedge E) = Pr(V_i = 1 \mid E)$.

Let π be the probability that juror i's private judgment is pivotal. Then

$$Pr(V_i = 1 \mid H \wedge E) = \pi p + (1\text{-}\pi)\, Pr(V_i = 1 \mid E)$$

and $Pr(V_i = 1 \mid \neg H \wedge E) = \pi (1\text{-}p) + (1\text{-}\pi)\, Pr(V_i = 1 \mid E)$.

Therefore

$$l(H, V_i = 1 \mid E) := log\left(\frac{\pi p + (1\text{-}\pi)\, Pr(V_i = 1 \mid E)}{\pi (1\text{-}p) + (1\text{-}\pi)\, Pr(V_i = 1 \mid E)} \right).$$

If juror i is always pivotal ($\pi = 1$), then $l(H, V_i = 1 \mid E) = l(H, V_i = 1) = \log(p / (1\text{-}p))$. If juror i is never pivotal ($\pi = 0$), then $l(H, V_i = 1 \mid E) = 0$.

The lower the probability π that juror i is pivotal, the lower is the value of $l(H, V_i = 1 \mid E)$, i.e. the lower is the conditional degree of support juror i's updated vote gives to H, given the other jurors' votes.

---

[1] Christian List, Department of Government, London School of Economics; c.list@lse.ac.uk. Philip Pettit, Departments of Politics and Philosophy, Princeton University; ppettit@princeton.edu. We are grateful to Campbell Brown for comments and discussion.

[2] Reflexivity need not be a bad thing. Consider a theory which shows that present social behaviour leads to certain undesirable consequences. The understanding of such a theory may lead the members of a given society to adjust their behaviour in such a way as to avoid the occurrence of conditions under which the theory applies. The theory will then have been undermined by reflexivity, but with an outcome that is more desirable than the one that would have occurred if the predictions of the theory had become true. Harsanyi (1976, p. 83) made a suggestion along these lines: "Keynesian economics has enabled us to make much better predictions about the effects of various economic policies in conditions of mass unemployment. By this means, it has also enabled us to eliminate these very conditions, and to create a completely novel economic situation of continuing high employment, in which Keynesian predictions may no longer work."

[3] Whether or not the independence assumption is undermined by jury deliberation – and, if so, how exactly the jury theorem is affected – is a debated issue. See, for example, Estlund (1994) and Dietrich and List (2002).

[4] The Condorcet jury theorem is robust to certain relaxations of the assumptions. A version of it still holds in certain cases where different jurors have different competence levels, but where the average competence

is greater than ½ (e.g. Grofman, Owen and Feld 1983; Borland 1989), and in cases where there are certain dependencies between different jurors' votes (ibid.; Ladha 1992; Estlund 1994; but see Dietrich and List 2002). Since the concern of this paper is not technical, however, we will here stick to the jury theorem in its simplest, classical form.

[5] What do we mean by saying that the Condorcet jury theorem is undermined by reflexivity? The jury theorem can be stated as the following conditional: If conditions C hold, then prediction P follows; in short, C→P. Conditions C are the assumptions of the theorem: competence, independence of private judgments, and private judgment voting. Prediction P is the proposition that the probability of a correct majority judgment converges to 1 as the jury size increases. Since the conditional C→P is a mathematical truth, reflexivity clearly cannot undermine the truth of that conditional. We say that the jury theorem is undermined by reflexivity in the following case: (i) Initially, conditions C hold, and so prediction P is also true. (ii) The individuals in question learn that C→P. (iii) The understanding that C→P leads the individuals to change their behaviour in such a way that conditions C cease to hold.

[6] Note that we are here considering an idealized limiting case. We do no claim that rational individuals will always reveal their private judgments truthfully in equilibrium. Recent game-theoretic work suggests that rational individuals may not in general do so. See Austen-Smith and Feddersen (2002) for a detailed investigation of the incentives on whether or not to reveal private judgments truthfully in group deliberation prior to voting.

[7] Note that this simple reasoning depends on the assumption that all individuals assign an equal prior probability of 0.5 to both possible states of the world. In the absence of this assumption, the picture is much more complicated. If we relax the assumption, two individuals may still disagree even after having observed the same pattern of private judgments across the group as a whole. An individual who assigns a prior probability of 0.5 to $X = 1$ will be convinced, after observing a majority for $X = 1$, that the posterior probability of $X = 1$ is above 0.5. By contrast, an individual who assigns an extremely low prior probability to $X = 1$ may still think, even after observing a substantial majority for $X = 1$, that the posterior probability of $X = 1$ is below 0.5.

[8] Goodin notes that this effect is a fairly robust prediction of a Bayesian framework. In a reasonably sized group, the epistemic force of the majority should convince the individuals in the minority to change their beliefs. However, he also notes that, in practice, the effect typically does not occur as predicted, and we are thus faced with a "paradox of persisting opposition". The reasons he discusses as to why opposition persists perhaps also suggest that the phenomenon of reflexivity discussed here is less severe than a purely Condorcetian perspective might lead us to think.

[9] This is clearly an idealized assumption. We do not claim that rational individuals, who are aware of the incentive structure of the problem, will always believe that others express their private judgments truthfully. Indeed, once an individual is tempted not to express his or her private judgment truthfully, he or she should see that others may feel the force of the same temptation. See Austen-Smith and Feddersen (2002).

[10] We here leave aside such unlikely cases as the one where all odd-numbered individuals privately judge that $X = 1$ while all even-numbered individuals privately judge that not $X = 0$.

[11] Again, note the idealizing character of this assumption. The belief that no others are free-riding may not be rationally sustainable in equilibrium.

[12] And, as we have noted, the belief that no others are free-riding may be rationally unsustainable in equilibrium.

[13] This result does not hold if the prior probability r takes a value different from 0.5.

[14] Again, the result ceases to hold if $r \neq 0.5$.

[15] Proof:
$$\begin{aligned} l(H, E) &= l(H, V_1 = 1) + l(H, V_2 = 1) + \ldots + l(H, V_h = 1) \\ &\quad + l(H, V_{h+1} = 0) + l(H, V_{h+2} = 0) + \ldots + l(H, V_n = 0) \\ &= h \log(p / (1-p)) + (n-h) \log((1-p) / p) \qquad \text{(by (*) and (**))} \\ &= (h - (n-h)) \log(p / (1-p)) \\ &= m \log(p / (1-p)), \qquad \text{where } m = h - (n-h). \end{aligned}$$

[16] What is the implication of this result for the probability that H is true given the evidence E that h out of n jurors have voted for '$X = 1$'? By the formula stated above, we have

$$Pr(H|E) = \frac{r}{r + (1-r)\, exp(-l(H, E))}$$

$$= \frac{r}{r + (1\text{-}r)\ exp(\text{-}m \log(p / (1\text{-}p)))},$$

$$= \frac{r}{r + (1\text{-}r)\ (^{(1-p)}/_{p})^{m}},$$

which is exactly Condorcet's formula (see List 2003).

[17] Compare our previous note on the idealizing character of this assumption.

[18] Again our simple result requires the assumption that the two possible states of the world have the same prior probability 0.5. Assuming a different prior probability or heterogeneous prior probability assignments across different jurors would make the picture much more complicated.

[19] As before, note the assumption that the two possible states of the world have the same prior probability 0.5.

[20] Again note the idealizing character of this assumption. If our argument here is correct, then it also shows that this assumption will not generally hold in equilibrium. Suppose that it holds for jurors 1 up to *i*-1. Then, as we have seen, juror *i* may have an incentive not to reveal his or her private judgment truthfully. So the fact that the assumption holds for jurors 1 up to *i*-1 will here prevent it from holding for juror *i* too.

[21] There is an analogy with the phenomenon of confirmatory bias here, our "tendency to attach more credence to evidence confirming rather than disconfirming our existing beliefs" (Goodin 2002, p. 128).

[22] Compare our remarks above on Goodin's "paradox of persisting opposition".