Should Juries Deliberate?

Brian Hedden

1 Introduction

Trial by jury is widely considered a fundamental feature of democratic governance, used in all common law countries, many civil law countries, and a number of Asian democratic countries whose legal systems do not fall neatly into either category.[1] The jury trial grants the responsibility for determining guilt or innocence to a group of ordinary citizens, rather than to a single judge representing the state. Jury decision-making is also sometimes seen as a model for democratic governance in general. Estlund (2000), for instance, gives an epistemic argument for democracy that is explicitly modeled on an analogy with juries. And jury decision-making plays a role in citizens' education about our democratic way of life; many Americans will remember enacting mock trials and watching the classic film *12 Angry Men* in high school civics classes.

But what form should jury decision-making take? In particular, should juries be permitted—even encouraged or required—to make their decisions by means of free, unstructured deliberation? In this paper, I argue that juries should not engage in such

[1] See Hans (2008) and 'The Jury is Out,' *The Economist* (2009) for discussion of the range of countries that used juries, and in what form. As noted, all common law countries use jury trials. Among civil law countries, some employ all-citizen juries (e.g., Spain and Austria) and still others (e.g., Germany, France, and Italy) use 'mixed tribunals' or 'mixed juries' consisting of both professional judges and lay citizens for more serious offenses. While mixed juries raise some special issues, the arguments of this paper will apply to them as well. Asian democracies that make some at least use of jury trials include Japan and South Korea. Democratic countries that do not use jury trials include South Africa and many countries in Latin America. Finally, jury trials are also used in some non-democratic countries, such as China.

deliberations on the way to reaching a verdict. In particular, I argue that jury deliberation is problematic on both theoretical and empirical grounds. On the theoretical front, deliberation destroys the independence of jurors' judgments that is needed for certain theoretical results (in particular, the Condorcet Jury Theorem) to be applicable. On the empirical front, there is evidence from both legal and non-legal contexts that unstructured group deliberation leads to group judgments that are worse in a number of respects than judgments generated by non-deliberative methods of judgment aggregation. Finally, I examine some possible alternatives to unstructured deliberation, including the constrained, structured deliberation embodied in the DELPHI method, voting (without deliberation), and averaging of probabilistic judgments.

This proposal can be seen as part of a broader movement advocating a 'less is more' approach to decision-making in a variety of contexts. It is tempting to think that more is more—that decision-making will benefit from having more choices, more information, more deliberation, and the like. In the legal case, Larry Laudan (2006) gives voice to this sentiment in criticizing rules of evidence that prevent the jury from seeing various sorts of evidence, such as rules not requiring testimony from the defendant, not requiring testimony from people with certain special relationships to the defendant (spouses, doctors, social workers, and the like), and ruling out non-voluntary confessions and illegally obtained evidence. He proposes that 'The triers of fact—whether jurors or judges in a bench trial—should see all (and only) the reliable, nonredundant evidence that is relevant to the events associated with the alleged crime' (2006, 121).[2]

[2] To be fair, Laudan does acknowledge that if 'we were to discover that there is a certain kind of relevant evidence (hearsay, for example) whose importance juries are apt to overestimate, then excluding it might be appropriate' (2006, 120). But I am inclined to think that this caveat kicks in more frequently than Laudan suspects and that we do indeed have evidence that jurors are apt to systematically overweight certain kinds of evidence. See Benforado (2015) for extensive evidence and references to relevant studies.

Laudan also notes that there may be non-epistemic reasons for excluding certain types of evidence, such as ethical considerations. But he generally finds these wanting. For instance, against the claim that the inadmissibility of illegally obtained evidence is necessary so as not to incentivize police misconduct, Laudan holds that the evidence should be admissible but that the wrongs of illegal searches

But we have evidence from psychology that people are apt to made all sorts of mistakes, and that these mistakes are often systematic and predictable. Often, this evidence suggests that we could achieve better outcomes by blinding decision-makers to certain sorts of evidence and otherwise constraining their decision-making. For instance, hiring committees would do well to remove names from applications to prevent the influence of various implicit biases (cf. Goldin and Rouse (2000), who present evidence that blind auditions increase the chances of females being hired by orchestras). Related proposals include increasing reliance on checklists rather than human judgment (Gawande 2010), relying on simple, mechanical rules for medical diagnoses and other predictions (Dawes et al 1989; Grove et al 2000), and using prediction markets rather than deliberation in governmental and corporate settings (Sunstein 2011).[3] The proposal to have non-deliberating juries is of a piece with these proposals in that it limits the information available to jurors and constrains their decision-making powers, with the aim of increasing accuracy. Nevertheless, the present proposal does not stand or fall with these other related proposals; it should be evaluated on its own merits.

2 Why Deliberate?

Jury deliberation is entrenched not only in public consciousness, but also in legal doctrine. I know of no jurisdiction that uses juries but does not have them reach a verdict through group deliberation. Some jurisdictions include model (or 'pattern') instructions that specifically instruct jurors to discuss the case with their fellow jurors in attempting to reach a verdict. For instance, the model instructions of the US First Circuit state:

> I come now to the last part of the instructions: the rules for your deliberations. When you retire you will discuss the case with the others to reach

should be remedied separately, for instance through civil suits.

[3] Another proposal is the use of 'virtual trials,' which will be briefly considered in Section 4.4.

> agreement if you can do so. You shall permit your foreperson to preside over your deliberations, and your foreperson will speak for you here in court. (Section 6.01)
>
> Each of you must decide the case for yourself, but you should do so only after considering all the evidence, discussing it fully with the other jurors, and listening to the views of the other jurors. (Section 6.03)

Other US Circuits' model instructions, along with those of a number of other countries, follow a similar pattern, instructing jurors to deliberate as a group so as to arrive at a verdict.[4]

Aside from historical inertia, why should our legal system include requirements that juries deliberate? There are both epistemic and legal considerations that might be used to support jury deliberation.

[4] For example, consider the cases of Australia, Canada, and the Crown Courts of England and Wales. In Australia, the High Court's ruling in Black v. the Queen (1993) recommends jury instructions that include the following:

> You also have a duty to listen carefully and objectively to the views of every one of your fellow jurors. You should calmly weigh up one anothers' opinions about the evidence and test them by discussion. Calm and objective discussion of the evidence often leads to a better understanding of the differences of opinion which you may have and may convince you that your original opinion was wrong.

The Canadian Judicial Council's model jury instructions say:

> You should make every reasonable effort, however, to reach a verdict. Consult with one another. Express your own views. Listen to the views of others. Discuss your differences with an open mind. Try your best to decide this case.

Finally, the *Crown Court Bench Book* (2010) on Directing the Jury includes the following:

> Subject to the application of section 17 Juries Act 1974, the jury must return a unanimous verdict. Section 17 enables a majority verdict to be returned after the jury has been deliberating for at least two hours. In practice, the minimum period is 2 hours and 10 minutes. By section 17(4) the trial judge, before considering a majority verdict, should allow such period for deliberation as the nature and complexity of the case requires. In long and complex, and multi-handed, cases it may be appropriate not to consider a majority verdict direction until the jury has been deliberating for well over a day and, perhaps, longer. It is good practice (and good manners) for the trial judge to invite observations from the advocates when a majority verdict direction is under consideration.

2.1 Epistemic Reasons for Deliberation

There are three epistemic reasons for deliberation. First, it is natural to think that group deliberation is a good way of taking advantage of individuals' particular backgrounds, perspectives, and expertise. Discussion gives each juror an opportunity to make his or her own distinctive expertise and reasoning, and their bearing on the case, available to each other juror. It is important to note, however, that some non-deliberative aggregation procedures, such as those discussed in Section 4, also take advantage of individuals' diverse perspectives, suggesting that this motivation does not supporting deliberation over other methods of reaching a verdict.[5]

Second, it is tempting to think that group deliberation profits from the pooling of evidence, which in the case of juries largely amounts to improving jurors' memories of trial proceedings (since jurors are supposed to focus on evidence presented at trial). The US Supreme Court, in Apodaca v. Oregon (1972, p. 386), endorses this motivation, stating that, 'Because they have imperfect memories, the forensic process of forcing jurors to defend their conflicting recollections and convictions flushes out many nuances which otherwise would go overlooked.' However, this purported benefit of deliberation may be overstated. A study by Pritchard and Keenan (2002) found that memory improvements resulting from deliberation in mock juries were real but small, and that deliberation also increases jurors' confidence regardless of their accuracy.[6]

Third, and relatedly, deliberation may help jurors resolve misunderstandings regarding language used in the presentation of evidence (involving ambiguity, vagueness, context-dependence, and the like) and settle on an understanding of the relevant standard of proof.[7]

[5] Cf. Hong and Page (2012) on diversity of backgrounds and its bearing on non-deliberative aggregation methods such as averaging of individuals' estimates.

[6] For deliberation increasing confidence without increasing accuracy, see also Heath and Gonzalez (1995) and Baron et al (1996).

[7] Thanks to an anonymous referee for suggesting this point.

2.2 Unanimity

There is a further consideration in favor of jury deliberation that is specific to the legal context, namely the requirement that jury verdicts be unanimous. If jury verdicts must be unanimous, it is difficult to see how juries could reach a verdict without engaging in deliberation, except in the rarest of cases. (This could just mean many more acquittals, except that in many jurisdictions, a unanimous judgment of not guilty is required for acquittal.) Indeed, the quest for unanimity seems to be the main consideration in favor of jury deliberation in various legal writings. For instance, the aforementioned First Circuit model instructions tell jurors to discuss the case among themselves 'to reach agreement if you can do so,' and the next sentence after the quoted text from Section 6.01 states, 'Your verdict must be unanimous.' Many of the other circuits' model instructions also mention jury deliberation and the unanimity requirement at the same time.

It is important to note, however, that not all jurisdictions require unanimity; many allow for majority verdicts (which vary with respect to the size of the majority required). But even then, many jurisdictions permit such non-unanimous majority verdicts only after the jury has engaged in deliberation for some specific period of time (usually between two and six hours). This is the case in Ireland, England and Wales, and a number of states in Australia (Drabsch 2005, 14-16).

Now, deliberation may not be strictly necessary for achieving some sort of consensus. There are models of consensus-building that allow for consensus to be achieved without any face-to-face interaction, but instead through iterated updating of individuas' probability estimates (e.g., Wagner and Lehrer 1981). However, these consensus models are controversial and, in my view, problematic, so I relegate a brief discussion to a footnote.[8]

[8] Here is a brief overview of the consensus model of Wagner and Lehrer (1981). Suppose that at the

More to the point, there are good reasons for dropping the unanimity requirement (though arguing that these reasons are decisive would go beyond the scope of this paper). To begin with, the unanimity requirement likely increases the frequency of hung juries, which is costly in both monetary and human terms. It also creates an added risk of 'rogue jurors' who unreasonably hold out against the more reasonable position of the majority, and makes it more difficult for a small minority of jurors to maintain a reasonable position in the face of group-think (see Section 3.2 for further discussion).

Moreover, arguments in favor of a unanimity requirement are unpersuasive. First, one of the main arguments in favor of the unanimity requirement is that it improves deliberation. For instance, the *amicus* brief supporting unanimity from the American Bar Association in the Supreme Court case Bowen v. Oregon claims that jury delib-

outset (round 0), there are n individuals with probabilities $p_1^0, p_2^0, ..., p_n^0$ in some proposition. Suppose also that each individual i assigns weights of respect w_{ij} to each individual j in the group (including herself), where the weights are non-negative and sum to one. These weights can be understood, roughly, as representing her assessments of the reliabilities of the various members of the group with respect to the topic at issue. At the first round of iteration (round 1), we take each individual's new probability to be a weighted linear average of the individuals' round 0 probabilities, with the weights being her weights of respect for the individuals in the group. So, for each individual i, $p_i^1 = \sum_j p_j^0 \times w_{ij}$. Now, suppose that some individual assigns a positive weight of respect to herself, and each individual can be connected to every other through a chain of assignments of positive weights of respect (they call this *communication of respect*). Then, iterated updating by weighted linear averaging is guaranteed to result in convergence to a consensus probability assignment.

Does this model point to a way of reaching consensus without deliberation? There are two problems. The first is that it is doubtful whether (as Wagner and Lehrer claim) individuals are rationally obligated to update their credences by taking such weighted linear averages of the group members' credences, in particular because updating by taking such a weighted linear average sometimes conflicts with the Bayesian norm of updating one's probabilities by conditionalization (Laddaga and Loewer 1985). The second is that it is unclear how to implement the model in practice. In particular, it is unclear where to get the weights of respect from. There are many options. We could ask jurors to rate each other, but it is questionable whether their estimates of each others' reliability should be taken very seriously. Alternatively, we could impose weights of respect from outside. We might, for instance, impose on them uniform weights of respect (so that each juror's weights for herself and others are all equal), or use weights that correspond to each juror's score on some set of test questions (along the lines of Cooke 1991). But imposing weights of respect from the outside in this way further undermines the claim that the resulting group-level probability judgment really constitutes any kind of consensus. Nevertheless, there are advocates of using the Wagner and Lehrer consensus model in real-world committee decision-making (Regan et al 2005), and there may be some justification for attempting to bring the model to bear on juries as well.

erations with a unanimity requirement are likely to be better than deliberations under a non-unanimous supermajority requirement. There is merit to this contention. The brief points to studies of mock juries (especially Hastie et al 1983) in which juries with a unanimity requirement deliberated for longer, took more ballots, and evaluated the evidence more thoroughly than those with a mere supermajority requirement. However, these studies only compared deliberating juries with a unanimity requirement against deliberating juries without a unanimity requirement; they did not establish that deliberating juries perform better in any respects than juries that reach a verdict by some non-deliberative aggregation method. Without independent motivations for unanimity or for deliberation, what we have is a circle: we need juries to deliberate because that's the only way they can achieve unanimity, and we need a unanimity requirement because it improves juries' deliberations.

Second, one might also think that unanimity is important because it serves as protection for defendants. It raises the bar, relative to mere supermajority verdicts, for how compelling the prosecution's case must be in order to obtain a conviction. And one might think that unanimity is required by the defendant-favoring *beyond a reasonable doubt* standard of proof. The thought goes that if one juror has a reasonable doubt, then *ipso facto* there is a reasonable doubt, and so a verdict of 'guilty' is unwarranted.

However, it is important to emphasize that unanimity is standardly required not only for conviction, but also for acquittal (France is an exception to this rule[9]). This means that unanimity does not clearly provide protection to the defendant, since a hung jury will simply result in a retrial. Also, the fact that lack of unanimity does not result in acquittal means that the courts cannot be seen as endorsing the thought that if one juror has reasonable doubts, then reasonable doubts exist, and hence conviction is unwarranted. Moreover, unanimity is not required for the legal system to have a

[9] See Drabsch (2005, 16-17).

high bar for conviction; it could adopt an even higher standard of proof than 'beyond a reasonable doubt,' or it could use supermajority verdicts with larger juries; a 23 out of 24 majority might be more difficult to achieve than 12 out of 12, for instance.

Finally, unanimity might be important because we want the legal system to 'speak with one voice,' to use Dworkin's (1986) phrase. Dworkin adopts this as a general desideratum for the legal system, in large part because it is important that the law be consistent, predictable, and non-arbitrary (though he does not use this consideration to argue for the unanimity rule). And unanimous jury verdicts might be thought necessary to uphold public confidence in the justice system.[10] Interestingly, however, we allow for non-unanimous rulings from collegial courts and non-unanimous verdicts from civil juries (as well as criminal ones in some jurisdictions, as noted above). It is not clear that this results in any significant erosion of public confidence in the legal system. Moreover, if we are concerned about the possible erosion of public confidence in the criminal justice system, we might allow non-unanimous verdicts but keep the vote totals hidden from the public. This might seem like an objectionable lack of transparency, but we already keep the jury's deliberations secret from the public, and these deliberations can be just as relevant as the final vote tally in terms of what the public should make of the case. Even with a unanimity requirement, it is epistemically relevant, for instance, whether uanimity was achieved immediately or only after much discussion to convince a few holdouts of the majority's view.

I conclude that while deliberation and unanimity go together naturally as a package deal, there are good reasons not to require unanimity in the first place. As noted above, some jurisdictions have already dropped the requirement of unanimity and allow for

[10] In addition to public confidence, it is worth noting that in studies of mock juries, jurors themselves seem to be more satisfied with the verdict and the quality of the deliberation than jurors on mock juries with majority decision rules (Hastie et al 1983, 78-9). These effects do matter and must be weighed against other considerations. However, effects of a given procedure on juror satisfaction pale in significance compared to their effects on accuracy.

majority or supermajority verdicts, at least for some cases. If unanimity is not required, one of the principal arguments in favor of having juries deliberate disappears.

3 Against Jury Deliberation

Reasons against having juries deliberate include theoretical and empirical considerations. I take these up in turn.

3.1 Theoretical Reasons

By deliberating, each juror's opinion is influenced by the expressed opinions of the others. Deliberation makes their opinions non-independent of each other, in the sense that even conditional on the factual guilt or innocence of the defendant, and conditional on the evidence presented in court, the probability that one juror gets the right answer will not be independent of the probability that some other juror gets the right answer. (A quick caveat: deliberation need not always destroy independence of jurors' judgments, for instance if none are at all swayed by the viewed expressed by others. But to the extent that jurors influence each other's judgments during the course of deliberation, deliberation will tend to erode independence.)

This loss of independence due to deliberation is important for two reasons. First, it conflicts with one aspect of standard scientific methodology, which is to use multiple independent tests to confirm a hypothesis. If multiple tests yield the same experimental result, the support they confer on a hypothesis will generally be stronger if the tests are independent than if they are non-independent. This is the main argument made by Dawkins (2000) against trial by jury.[11]

[11] Somewhat oddly, to my mind at least, Dawkins does not propose doing away with jury deliberation entirely. Instead, his proposed remedy is simply to have two separate deliberating juries, with guilty verdicts from both of them required for conviction.

Second, and relatedly, the loss of independence undercuts one particular theoretical consideration in favor of using juries rather than single judges to decide verdicts, namely the Condorcet Jury Theorem, or CJT (Condorcet 1785; see also Grofman et al 1983 for a review of related results). In general, this theorem provides some justification for having decisions made by large groups of people rather than by a single individual or only a few. Suppose that, for some proposition, individuals' judgments as to whether that proposition is true or not are probabilistically independent of each other (the *independence* condition). Suppose further that individuals are all better than random at determining whether or not the proposition is true. That is, if it is true, each individual is more likely than not to judge that it is true, and if it is false, each individual is more likely than not to judge that it is false (the *competence* condition).[12] The CJT states that, if the independence and competence conditions are met, then the probability that a majority of the individuals' judgments will be correct increases with increasing group size. In the limit, the probability of a correct majority judgment goes to 1.

Now, before going further, it is worth noting that even if the CJT were to apply to real life juries (i.e. if the competence and independence conditions were met), it would not by itself be sufficient to justify reliance on juries. After all, the CJT does *not* say that juries of any particular size will be tremendously accurate; the probability that the majority of any particular group judges correctly depends both on the size of the group and on each individual's competence level (or probability of judging correctly). It also does not say that a jury can be expected to be more accurate than a single

[12] A clarificatory note: while standard presentations of the CJT assume that all individuals have the same competence level, extensions of the CJT weaken this condition. Grofman et al (1983) show that Condorcet-like results can still obtain if individuals are heterogeneous with respect to competence. Their Theorem VI states that for heterogeneous groups (where individuals need not all have the same competence level), if each individual has a competence level above 0.5, then the greater the probability that a majority judgment is correct. Moreover, their Theorem V allows that some individuals may have a competence level below 0.5. It states that if the distribution of individuals' competence levels is symmetric, then results analogous to the CJT can be obtained, with the average competence level in place of the competence level that was previously assumed to be the same for everyone.

judge, if that judge has a higher competence level than some of the jurors. So there are already obstacles to using the CJT to justify relying on juries.[13] Nevertheless, given the independence and competence conditions, the CJT does entail (i) that for any particular judge, there is some n such that a jury of size n has higher expected accuracy than the judge, and (ii) that larger juries have higher expected accuracy than smaller ones (holding fixed jurors' competence levels).

But even this limited justification for juries is blocked if juries deliberate before reaching a verdict. After all, if they deliberate, then jurors' judgments will very likely no longer be probabilistically independent of each other. Deliberation erodes independence.[14] Thus, even granting the competence condition, the CJT does not apply to real-life juries if they engage in group deliberation prior to reaching a verdict.

Now, one might object that juror's judgments will be non-independent even in the absence of deliberation. After all, as Dietrich and List (2004) note, jurors' judgments will still be non-independent in virtue of their having seen the same evidence presented in court. The latest common cause of their judgments will not be the state of the world (the defendant's guilt or innocence) but rather the shared body of evidence that they received at trial. In order for the CJT to apply and say that the probability of a correct majority verdict goes to 1 with increasing group size, the juror's judgments must be independent given the state of the world, not merely independent given the shared body of evidence.

There are two, closely related, things to say in response. First, let us suppose that our jury does not engage in deliberation, so that jurors' judgments are indeed independent given the shared evidence, but non-independent given the actual state of

[13] Another obstacle to using the CJT to justify reliance on juries is that the theorem concerns the probability of a *majority* judging correctly, whereas juries are often subject to unanimity decision rules.

[14] Rawls (1999, 315) makes this point in the context of justifying majority rule in political affairs. Another relevant possible cause of non-independence is the presence of opinion leaders. See Grofman et al (1983) and Estlund (1994) for discussion of opinion leaders and independence in the context of the CJT.

the world. Then, Dietrich and List show that it's no longer true that the probability of a *correct* majority verdict increases with increasing group size (going to probability 1 in the limit). But they also show that it is the case that the the probability of a *valid* majority verdict (i.e. a verdict matching the rational response to the evidence presented in court) does still increase with increasing group size. And as group size increases, the probability that the majority will give the correct verdict converges on the probability that the evidence presented in court is not misleading (i.e. that the evidence points toward guilt if and only if the defendant is in fact guilty). That is still a comforting thought.

Moreover, the fact that the jurors are exposed to a shared body of evidence only means that their judgments are non-independent if we think of them as judging whether or not the defendant is guilty. But we could also think of them as making a judgment about the higher-order proposition that the admissible evidence supports a guilty verdict. That is, we treat the 'state of the world' as referring not to actual guilt or innocence, but to whether the admissible evidence points toward guilt or innocence. Exposure to a shared body of evidence erodes independence of judgments with respect to the question of whether the defendant is guilty, but not with respect to the question of whether that evidence supports a guilty verdict. And it is natural to think of the latter question as the one to which juror's should direct their attention and the one on which they should be seen as voting. After all, it would be inappropriate for a juror to vote 'guilty' if he or she thought that while the defendant was guilty, the evidence admissible in court wasn't sufficient to show this.[15] Thus, again provided that the jury does not deliberate, the CJT still applies to say that if the competence condition is met with respect to the question of whether the admissible evidence supports a guilty verdict, then the probability that a majority correctly judges whether guilt has been

[15] For instance, the juror might be convinced by some piece of evidence that was presented but then ruled inadmissible.

established beyond a reasonable doubt increases with increasing group size (and goes to 1 in the limit).

So, the CJT does not apply to real-life juries if they engage in deliberation. But it can apply (provided the competence condition is true, and given suitable interpretations of the question that they are seen as addressing) if they do not deliberate.

I don't want to overstate my case. First, as noted above, the CJT is not by itself sufficient to justify reliance on juries rather than single judges. So the fact that deliberation and the resultant non-independence of jurors' judgments renders the CJT inapplicable may be no great loss. Second, and more importantly, the CJT does *not* say that non-deliberating groups (or, more generally, groups whose members' judgments are independent) will do better than deliberating groups. A friend of deliberation might say, for instance, that while deliberation typically erodes independence, it increases individuals' competence. Moreover, the loss of independence may be slight (making the CJT apply 'approximately' if not strictly; though note that if a unanimity requirement on the jury's verdict is in place, the loss of independence will likely be great). If so, the benefits of deliberation in the form of increased individual competence might outweigh the costs in the form of decreased independence such that deliberation on balance yields more accurate verdicts.

Now, whether deliberation really does increase individual competence so as to yield more accurate group judgments on balance is a straightforwardly empirical question, though we will see evidence in the next section that paints a pessimistic picture.

3.2 Empirical Reasons

A significant body of social scientific evidence suggests that deliberating groups do worse in a variety of contexts than non-deliberating ones. To begin with, in addition to numerous cases of deliberating groups doing poorly, we can point to lots of examples of non-deliberating groups doing well. Cases of surprisingly good performance by non-deliberating groups are well-known from the literature on the so-called wisdom of crowds. Perhaps most famously, there is Francis Galton's (1907b) famous case in which the average of nearly 800 individual estimates of the weight of a prize ox was only one pound off from the actual value (1,197 pounds vs. the actual 1,198 pounds). There are also well-attested successes of prediction markets, which have been incredibly accurate in predicting the outcomes of elections, as well as winners of the Oscars and opening weekend box office receipts.[16] (This is not to say that no interpersonal deliberation at all took place in these cases; it is likely that small numbers of people in Galton's case and in the prediction markets did discuss their judgments with each other.[17]) There are also many cases of group deliberation yielding poor estimates or poor decisions (see Janis 1982 and Sunstein 2011).

Group deliberation can fail for a variety of reasons, some or all of which might apply in any given case. First, there are social pressures. Some people might not speak up, or speak less forcefully, for fear of social sanction. This is especially likely when the person is in the minority, and most extreme when the person is a lone dissenter. And of course, social pressures are likely strongest when unanimity is required, as in the case of many juries. While a unanimity requirement may improve deliberation in the sense of making deliberation last longer and involve consideration of a greater range of facts, it also increases the social costs of holding and maintaining a dissenting view. And while jurors may indeed be well-intentioned and take their jobs with the utmost seriousness, good intentions will not make them immune to succumbing to social

[16] See Sunstein (2011).

[17] Moreover, even without any deliberation, the judgments of participants in a prediction market will not be independent of each other, since they are influenced by signals sent by market prices. Group deliberation is neither necessary nor sufficient for the non-independence of participants' judgments. Individuals can discuss matters with each other without their judgments becoming probabilistically dependent (e.g., if they are not at all influenced by the discussion), and individuals can have probabilistically dependent judgments even in the absence of group deliberation (e.g., in the prediction market case).

pressures or imposing such pressures on others. After all, this is not the sort of thing that is entirely under one's conscious control (any more than the influence of myriad other biases and heuristics can be avoided by simply thinking hard).

Moreover, there is some evidence that real jurors do indeed succumb to social pressures. In a study of approximately 3,500 jurors in four large urban courts, Waters and Hans (2009) found that a startling number of juries (all using a unanimity rule) included at least one juror who reported that if it were entirely up to them, they would have reached a different verdict. Since there are often multiple charges in a trial, Waters and Hans defined a 'general verdict measure' which takes into account these multiple charges and 'summarizes the predominant outcome of the jury trial' (520). They found that 38% of juries contained at least one juror who privately would have voted against the jury's general verdict but nonetheless joined with the others. And 54% of juries had a least one juror who privately disagreed with the final vote of the jury (some of whom agreed with the 'general verdict' of the jury, despite differing on certain charges). While the evidence does not show that these jurors went along with the others despite private disagreement due to specifically social pressures arising during deliberation (as opposed to a desire to get the job done and head home in time for dinner, say), this is certainly a plausible conclusion to draw.

Second, there are purely informational reasons why group deliberation can fail. The fact that a group of people (say, those who speak up first) all express a belief in a given proposition (or present evidence for that proposition) constitutes an epistemic reason (that is, evidence) to come to believe that proposition oneself, perhaps even a reason strong enough to outweigh one's initial inclination to believe the opposite.

The role of informational pressures is most clearly illustrated with informational cascades. In one experiment, subjects had to guess whether the experiment used an urn containing two red balls and one white one, or instead an urn containing two white balls and one red one.[18] Each subject was privately shown a ball drawn from the urn (with the drawn ball being replaced each time). After each subject's draw, the subject announced to the group her guess of which urn was being used. Crucially, while the first subject had only the evidence of her own draw to go on, subsequent subjects could base their guess on both the result of their own draws and also on the announced guesses of previous subjects. And so after the first few subjects announced their guesses, subsequent subject's guesses rationally ought to follow whatever the majority of previous announcements was. This can create cascades (and indeed such cascades do result, as the experiment showed), where for instance if the first two subjects each see a red ball, the third one will (rationally) announce a guess of the first (majority red) urn even if she sees a white ball, and so with the fourth, and the fifth, and so on. Now, in the case of juries, jurors don't really have such private information, except to the extent that, due to their fallible memories, each juror will likely remember a different subset of the evidence presented at trial. But cascades still present a risk. If the first few speakers, or the first few people to raise their hands in a straw poll, go one way, that constitutes evidence in favor of their view which might lead others to go along, even if they were inclined to go the other way prior to exposure to evidence of the first few jurors' views.

Worryingly, both social and informational pressures are likely to have a disproportionate impact on 'low status' jurors, that is, females, members of minority ethnic groups, jurors with less education, jurors of low socioeconomic status, and the like. There is substantial evidence of unequal participation in group deliberations, both in juries and in non-legal contexts. For instance, it has been found that 'low-status' group members speak less and exert less influence in medical teams (Christensen and Abbott 2000). On juries, males have been found to speak more than females, and higher socioeconomic status jurors participate at higher rates than lower status jurors (see Hastie

[18] See Anderson and Holt (1997).

et al (1983, 28) and references therein). In addition, in cases where the jury foreman is elected by the jury (as opposed to appointed by the judge), 'juror sex, social status, and seat location are correlated with election to the foreman role. Males, higher classes, and end seating are overrepresented in the role' (Hastie et al 1983, 28). And there is a concern that these same sorts of people may be taken less seriously by others and also may be more likely to give up their initial view when confronted by those with opposing views. Miranda Fricker (2007) dubs this phenomenon 'epistemic injustice.'[19]

In addition to social and information pressures, there is evidence that group deliberation can amplify rather than correct individuals' errors. Sunstein (2011, 320) cites evidence that 'If individual jurors are biased because of pretrial publicity that misleadingly implicates the defendant, or even because of the defendant's unappealing physical appearance, juries are likely to amplify rather than correct those biases' and are also likely to 'be more affected by the biasing effect of spurious arguments from lawyers.'[20]

The discussion so far has focused on accuracy. But another important value is consistency, by which I mean predictability (or low variability) rather than logical consistency. We want it to be the case that if we repeat the same trial over and over again, we will tend to get the same verdict each time. Consistency is important for accuracy; if ten trials of the same case result in five guilty verdicts and five not guilty verdicts, then half of the trials yielded inaccurate verdicts. But consistency may also be also valuable in its own right. It is part of the concept of the rule of law that legal (or more generally, governmental) actions are non-arbitrary. As Rawls (1999,

[19] On a more optimistic note, Sommers (2006) found that racially diverse juries do better in various respects than more homogeneous ones. In particular, racially diverse juries exchanged a wider range of information than homogeneous ones, and this wasnt wholly attributable to the performance of blacks on the juries; white participants on diverse juries also cited more facts, made fewer errors, and were more amenable to discussion of racism than whites on homogeneous juries. Crucially, however, this just provides evidence that diverse deliberating juries do better (in certain respects) than homogeneous deliberating juries. In no way does it suggest that deliberating juries (whether diverse or not) do better in any respects than non-deliberating ones.

[20] See MacCoun (ms, 116, 121) for the first point, and Schumann and Thompson (ms) for the second.

208) puts it, 'The rule of law implies the precept that similar cases should be treated similarly.' Among other things, this means that individuals should not be subjected to chancy or otherwise arbitrary treatment if at all possible. But high unpredictability of verdicts is in tension with this aspect of the rule of law.[21] Consistency may also be important for deterrence, as individuals should be able to reason about the likely costs of committing some crime.[22] It is unclear, however, whether the value of consistency vis-à-vis deterrence goes beyond its contribution to accuracy; it may be that deterrence primarily requires accurate verdicts, and consistency is only important to deterrence insofar as it is important for overall accuracy of the criminal justice system.

But regardless of whether consistency is valuable in its own right or merely due to its contribution to accuracy, it is still valuable. And while it might be thought that jury deliberation will yield more predictable, less variable verdicts, in fact the opposite seems more likely. To begin with, it would offhand be surprising if deliberation increased the predictability of verdicts, simply due to the existence of many different possibilities for how deliberation might proceed. Different choices of foreman, different orders in which people speak up, and whether deliberation involves periodic ballots (and if so, whether they are secret or not) might all affect the final outcome.

We need not simply speculate. There is evidence that jury deliberation increases unpredictability in civil cases (Schkade et al 1999, 2000). In a study of mock juries, Schkade and his colleagues had jurors read summaries of civil trials and determine whether to award punitive damages (in addition to compensatory damages) and also what those punitive damages awards should be. They found that deliberation led to higher variability in dollar awards. Dollar awards made by the jury after deliberation were less consistent and predictable than the mean or median of the awards that were

[21] But see Lewis (1989) for an argument that the concept of a chancy punishment can justify the differing sentences for murder and for attempted murder.

[22] Compare Rawls (1999, 208), who writes that 'Men could not regulate their actions by means of rules if this precept [that similar cases should be treated similarly] were not followed.'

preferred prior to deliberation. (Dollar awards were also systematically higher post-deliberation, with 27% of mock juries reaching dollar verdicts that were as high or higher than the any of the individuals' predliberation judgments. This polarization phenomenon may also show up in criminal trials. Hastie et al (1983, 59) found, in their study of mock juries, that 'there was a shift from the favorite predeliberation verdict of manslaughter to the final modal jury verdict of second degree murder.') Now, the fact that mock civil juries' post-deliberation dollar awards were less predictable than taking the mean or median of their predeliberation awards does not criminal juries' post-deliberation verdicts will likewise be less predictable than, say, taking the average of their pre-deliberation probability judgments of the defendant's guilt. It might be, for instance, that in the mock civil juries, deliberation increased unpredictability in large part due to the absence of an upper bound on dollar awards. But in the absence of empirical evidence to the contrary, I think it is more likely than not that deliberation increases, rather than decreases, unpredictability of verdicts in criminal cases as well as civil ones. So insofar as predictability and low variability is important in the law, this constitutes another reason against having criminal juries deliberate on their way to reaching a verdict.

4 How Should Juries Decide?

If juries are not to deliberate in reaching a verdict, what should they do? In this section, I sketch three possible methods: voting, averaging of probability judgments, and the DELPHI method. Before beginning, however, I want to emphasize that I take it to be a largely empirical question which method of jury decision-making should be adopted. Ethical and perhaps specifically legal concerns may somewhat constrain our choice of decision-making method. But aside from these sorts of ethical and legal concerns, our choice of jury decision-making procedure should be determined on the basis of empirical evidence about which procedure yields the most consistent and accurate group judgments. While theoretical results in the theory of judgment aggregation can no doubt do some of the work in determining which procedures are most likely to be consistent and accurate, empirical evidence does the bulk of the work in this regard. Therefore, while I will sketch some considerations in favor of certain decision-making procedures below, any recommendations are open to be being overturned by further evidence, especially evidence about their use in jury-specific contexts.

4.1 Voting

The simplest procedure would be to have jurors simply vote by secret ballot, without deliberating with each other, in favor of a guilty verdict or a not guilty one. Of course, there are many different ways of going from a set of votes to a group verdict that will then feed into the rest of the legal process. The most conservative option (in the sense of least departure from the *status quo*) would be to adopt an 11-1 or 10-2 supermajority rule. This sort of decision rule is already used in many jurisdictions (see Section 2), though they still require the jury to reach that verdict through group deliberation. There is a further choice to be made as to whether such a supermajority should be required for both conviction and for acquittal or just for conviction (so that a non-supermajority vote either way would result in acquittal rather than a hung jury). My sentiments lie with the latter option, though I won't say more about this choice point here.

An alternative, more revisionary proposal would have the sentence vary depending on the final vote tally. This would amount to a system of scaled punishments (Laudan 2010; Wansley 2013). With scaled punishments, the severity of the sentence is discounted by the jury's confidence in the defendant's guilt (reflected in the final vote tally or in some other way). A full defense of this proposal would go beyond the scope of this paper, but I will note two advantages of this system: transparency and proportionality. It conveys to the public more fine-grained information about the strength of the evidence against the defendant, as opposed to the *status quo* in which acquittals, for instance, indicate only that the case fell somewhere in the range between showing the defendant to be definitely innocent to falling just shy of proving his or her guilt beyond a reasonable doubt. And it treats similar cases similarly, instead of having a sharp discontinuity right around wherever we set our standard of proof. There are of course many objections to a system of scaled punishments, but that is not our main topic, so I will just mention a few in a footnote.[23]

But I want to note a third advantage of a system of scaled punishments that is relevant to the present topic, namely that it mitigates the threat of strategic voting. Suppose that we have a non-scaled system of punishments, with unanimity required for conviction (but not for acquittal); the same argument goes through if we have a supermajority decision rule instead of unanimity. Then, a juror may have good reasons not to vote in the way that expresses her genuine beliefs about the defendant's guilt or innocence, even if her only concern is that the jury's overall verdict be accurate. She has two relevant cases to consider: either her vote is decisive, or it is not. If it is not, then it doesn't matter which way she votes. If it is decisive, however, that means

[23] Perhaps the main concern about scaled punishments involves the issue of low-probability verdicts. The worry is that defendants will be subject to punishment under such a system even if the jury reports that the defendant is, say, only 20% likely to be guilty. This also creates a worry about potential abuse, in which prosecutors harrass political or other targets, hoping to inflict some measure of punishment on them. Wansley (2013, 353-4) argues that there are a number of obstacles to such abuse, including limited prosecutorial resources, political accountability for district attorneys, the appeals process, and threats of lawsuits for malicious prosecution. Regarding the intuitive repugnance of punishing defendants as a result of low-probability convictions, Wansley argues that the scale of punishments should be sharply non-linear, largely due to the decreasing marginal disutility of prison time and other punishments (one year in prison is far more than half as bad as two years in prison). Therefore, below some fairly high probabilistic threshold, punishments would probably involve no prison time and instead involve at some some sort of probation or supervision. Other concerns involve whether this system would increase net incarceration and that it would undermine public confidence in and support for the criminal justice system.

that all 11 other jurors were in favor of conviction. If she supposes that their votes sincerely express their beliefs, then she could rationally conclude that, supposing her vote is decisive, then the defendant is very likely to be guilty. So, she can reason that if her vote is not decisive, it doesn't matter which way she votes, while if her vote is decisive, she should vote guilty, since in that case all 11 others believe the defendant is guilty, which is strong higher-order evidence for the defendant's guilt.[24] So, she should vote guilty, even if her present unconditional probability that the defendant is guilty is far below whatever might be thought of as the threshold of reasonable doubt.[25] Under a system of scaled punishments with the final vote tally providing the scaling, however, there is no division of cases into ones where the juror's vote is decisive and ones where it doesn't matter. Instead, her voting guilty will always increase the defendant's punishment, and her voting not guilty will always decrease it, and so the setup incentivizes her to vote in accordance with what she genuinely believes about the case.

4.2 Probability Averaging

A second possibility would be to have jurors report their probabilistic judgments that the defendant is guilty, i.e. to report on a scale of 0 to 1 (or perhaps 0% to 100%, since that scale may be more familiar) how confident they are that the defendant is guilty, and then take the average of those reported judgments to be the jury's overall judgment. (Outlier probability judgments could perhaps be discarded first, but empirical evidence on the frequency of outlier judgments, whether genuine or manipulative, would be needed to assess the merits of this move.[26])

[24] Of course, if she thinks that the other jurors will go through the very same reasoning, then this gives her at least some grounds for doubting whether their votes will express their sincere beliefs.

[25] See Feddersen and Pesendorfer (1998) and List and Pettit (2011, 114-119) for further discussion.

[26] Cf. Galton's (1907a) argument against using the average of individuals' estimates to serve as the group estimate: 'That conclusion is clearly *not* the *average* of all the estimates, which would give

Such averaging has its merits. Like voting (without deliberation), it is simple and intuitive and preserves independence of jurors' judgments. Averaging of individuals' estimates has also proven highly successful in many cases, for instance in the aforementioned experiment of Francis Galton's, in which the average of nearly 800 individuals' estimates of an ox's weight differ by only one pound from its actual value. And the use of averaging to arrive at a group-level estimate shows up in some attractive theoretical results, such as the Diversity Prediction Theorem (Hong and Page 2012), which says that the accuracy of the group-level estimate of a given value (e.g., the truth value of the proposition that the defendant is guilty) is equal to the average accuracy of the individuals' estimates minus the variance of those individual estimates. Because variance must be non-negative, this means that the group is guaranteed to do at least as well as the individuals do on average, and that increased judgment diversity in the group (i.e. increased variance of estimates) increases the degree to which the group outperforms the individuals. This result only goes through, however, when the group-level estimate is taken to be the linear average of the individuals' estimates.[27]

One might worry that individuals are bad at working with probabilities, and so their probability judgments will be unreliable. But while this might justifiably make us

a voting power to "cranks" in proportion to their crankiness. One absurdly large or small estimate would leave a greater impress on the result than one of reasonable amount, and the more an estimate diverges from the bulk of the rest, the more influence would it exert.' Galton himself favored taking the median estimate to serve as the group's estimate. See also Bassett and Persky (1999) and Levy and Peart (2002). I leave open whether discarding outliers or instead taking the median would be better.

[27] As noted by Laddaga and Loewer (1985), linear averaging of probabilities does not commute with the Bayesian norm of Conditionalization, which states that one's probability for H after learning E (and nothing stronger) should equal one's previous conditional probability for H given E. Suppose that we take the linear average of a set of individuals' probability functions, and that then they all learn E (and nothing stronger), and then we take the linear average of their post-learning probability functions. The new group-level probability function will not in general equal the old group-level probability function conditionalized on E. How worrying this is depends on why we think Conditionalization is important. Many authors (e.g. Russell et al 2015) motivate Conditionalization by appeal to a diachronic Dutch Book argument, which shows that violating Conditionalization can lead one to accept a set of bets offered at different times that together guarantee a loss. But this should not be concerning in the jury context, for juries last only a short time and (typically) make just one decision, and anyway they are not vulnerable to exploitation as other groups, like corporations, might be.

wary of the use of explicitly probabilistic evidence in the courtroom (not to say that it should be ruled out), I do not think that it should make us skeptical of jurors' ability to accurately report probabilistic degrees of confidence. On the proposal under consideration, jurors are simply asked to report their confidence on a bounded scale. They are not asked to do any complex manipulations of probabilities of the kind where people tend to make errors. And people are, I think, reasonably familiar with probabilities playing this sort of role. For instance, sports fans are well practiced at consuming and reporting probabilities that a given team will win.

Moreover, even if people sometimes make mistakes in reporting their probabilistic judgments, this does not mean that a system where jurors report probabilities would be worse than the present system, in which jurors are asked to decide whether a given claim has been established by preponderance of the evidence, by clear and convincing evidence, or beyond a reasonable doubt. While jurors may have some uncertainty about how to understand probabilities, my suspicion is that this uncertainty pales in comparison with their uncertainty about how to interpret these baroque legal expressions. Indeed, Kagehiro and Stanton (1985) found evidence that people understand probabilistic verdicts better than the traditional legal ones. They found that when subjects were asked to use probabilistic analogues of the three traditional legal standards of proof (with 0.51 for preponderance of the evidence, 0.71 for clear and convincing evidence, and 0.91 for beyond a reasonable doubt), the frequency of verdicts of guilty (or liable) decreased as the standard of proof became stricter, as desired. But this effect was considerably weaker when subjects were asked to use the traditional legal standards of proof. This suggests that people's understanding of probabilistic standards of proof, while no doubt imperfect, may still be better than their understanding of the traditional non-quantitative legal ones.

Once jurors report probabilities, and these probabilities are averaged, there is a further question of what to do with that average. First, we could set a threshold such that a jury average probability above that threshold amounts to conviction, and below that threshold is acquittal. Legal scholars, when pressed, typically peg 'beyond a reasonable doubt' as probability 0.95 or higher, although many judges and scholars are reluctant to propose any numerical gloss. This option faces some objections. One might worry that any threshold, whether 0.95 or something else, will be arbitrary. And one might worry that giving an explicitly probabilistic threshold for conviction would make explicit our willingness to occassionally convict innocent people (Tribe 1971). One's probability assignments are *well-calibrated* if, for every 100 cases in which one judges the probability of an event to be n, the event obtains in $n \times 100$ of them. So if we set the threshold for conviction at 0.95, then if juries' probabilistic verdicts are well-calibrated, as many as 5% of convictions will be false.[28]

I do not regard these objections as decisive. To begin with, we should think of the relevant threshold not as some arbitrary number pulled out of a hat, but rather as something to be debated among citizens as we attempt to settle the relative values or disvalues to assign to true convictions, false convictions, true acquittals, and false acquittals (DeKay 1996; Laudan 2006, 2008). The threshold is only arbitrary to the extent that assignment of relative values to these different possibilities is arbitrary. Once we think of the threshold in that way, there should be (*contra* Tribe) no great objection to making explicit this socially, democratically determined choice. And in any event, everyone acknowledges that innocent people have been, are, and will be

[28] No conclusions about the likely rate of false convictions follow if we don't assume well-calibration. The rate of false convictions also depends on the ratio of truly guilty to truly innocent people who are brought to trial, as well as the probabilty that the total evidence at trial is misleading (i.e. the probability that the evidence supports guilt, given that the defendant is in fact guilty, and the probability that the evidence supports innocence, given that the defendant is in fact innocent). For instance, even with a 0.95 threshold for conviction, no false convictions will result if no innocent people are brought to trial, or if the the evidence always points in favor of innocence (and the jury is able to pick up on this fact) whenever the defendant is in fact innocent. See DeKay (1996) and Laudan (2008) for discussion.

mistakenly convicted, so I am not concerned that any great upheaval will result from saying as much out loud.

Second, instead of setting a threshold, we could adopt a system of scaled punishments in which the sentence is weighted by the jury's probability that the defendant is guilty. We have already discussed the merits of a system of scaled punishments, and see fn 23 for brief discussion of objections.

Before closing this section, I want to flag that what is really distinctive about this possible procedure is that it elicits probabilistic judgments from jurors just once and then combines those probabilities. But these probabilities could be combined in ways other than linear averaging. There is an extensive literature on other ways of combining probabilistic judgments (geometric averaging, various ways of assigning weights to the different probabilistic judgments, and so on), and I won't attempt to survey them here.[29] If theoretical and empirical considerations show one of these alternatives to be superior to linear averaging, then it should be adopted instead, but the same issues of how to go from a group probabilistic judgment to a sentence (setting a threshold vs. adopting scaled punishments) will still arise. I leave it to others to debate the merits of linear averaging versus alternative methods.

4.3 The DELPHI Method

The final procedure I consider is the DELPHI method, developed by the RAND Corporation in the early 1950s (Cooke 1991). There are many variants, but the basic procedure involves (i) anonymous elicitation of judgments (in this case, probabilities or binary verdict preferences) from group members, followed by (ii) a summary of these judgments being shown to the group members, followed by (iii) another round of elic-

[29] See Genest and Zidek (1986) and Cooke (1991) for surveys.

itation of judgments.[30] The standard DELPHI method stops at this second round, though it could be extended to any number of rounds. The final elicited judgments are then aggregated in some way (e.g., by some voting rule, in the case of binary verdict preferences, or by linear averaging or some alternative method like geometric averaging, in the case of probability judgments). And the same choice will then have to be made between setting a threshold vote total or probability for conviction or instead adopting scaled punishments.

The DELPHI method, unlike the previous two methods considered, does not preserve independence of the jurors' judgments as required by Condorcet-style theorems. The flip side is that like unstructured deliberation, DELPHI is likely to decrease variation between individuals' judgments, and so insofar we think jury cohesion is important (so that e.g., the legal system can be seen as speaking with one voice), DELPHI offers an improvement over the two procedures we have already considered. In other respects, DELPHI corrects for the deficiencies of unstructured deliberation. Judgments are made anonymously, thus mitigating the threat of social pressures and outsized influence by certain jurors. And judgments are made at the same time and reported simultaneously, thus eliminating the threat of informational cascades in which the first few opinions expressed affect the next ones, which then affect the next ones, and so on.

The DELPHI method is now widely used in a variety of contexts in place of unstructured deliberation, and studies have shown it to yield accurate group judgments, and in particular more accurate judgments than unstructured deliberation and simple averag-

[30] A notable variation of the standard DELPHI method allows individuals to discuss their reasons for their initial judgments after being shown a summary of individuals' initial judgments. This may resemble certain particularly tightly structured jury deliberations in which jurors take periodic secret ballot straw polls with discussion in between. This of course threatens to bring back many of the bad features of group deliberation (though it reduces the influence of social pressures by preserving anonymity and, presumably, not requiring total consensus in the end). This sort of variant, often referred to as Estimate-Talk-Estimate has been shown to be successful in many contexts, and so is also worthy of consideration, even though it involves an element of group deliberation. See Burgman (2015) for discussion of this sort of variant on DELPHI.

ing of initially elicited judgments (Dalkey 1969; Dalkey and Brown 1971; Woudenberg 1991; Rowe and Wright 1999; Graefe and Armstrong 2011), though see Gustafson et al (1973) for a study in which the DELPHI method did poorly compared to alternatives.

4.4 Jury Size and the Wisdom of Crowds

One of the main lessons from theoretical and empirical work on the wisdom of crowds is that group accuracy is generally improved as the number of individuals in the group increases. But group deliberation is most natural with relatively small group sizes. It would be impossible, or at least chaotic, to attempt unstructured deliberation with an auditorium full of people. Given the *status quo* of face-to-face jury deliberation, it would be problematic to attempt to further benefit from the wisdom of crowds by increasing jury size. But with non-deliberative judgment aggregation methods like voting, averaging, and DELPHI, larger jury sizes are possible and would very likely improve accuracy.

Of course, larger juries are also more costly, with more citizens having to take time off work and longer *voir dire* processes (though larger juries would make each juror's impact on the verdict smaller, which might make the *voir dire* process less important). Nevertheless, there are ways in which we might increase jury size without substantially increasing costs. To begin with, eliminating deliberation would itself reduce such costs by shortening the process. Cost reduction would be minimal in most cases, however, saving jurors only a couple of hours, but cost reduction would be more substantial of course in the case of prolonged deliberations.[31]

[31] In their famous study of American juries, Kalven and Zeisel (1966) found that in nine out of ten trials, the eventual jury verdict matched the pre-deliberation verdict preferences of a majority of jurors. If one concludes on that basis that deliberation is irrelevant (and no often nefarious, as I have suggested), then cost-saving considerations alone support doing away with jury deliberation. However, it is not clear that their finding really shows that deliberation is irrelevant. It could be, for instance, that nine out of ten trials are 'easy cases' (so it would be surprising if the ultimate verdict diverged from the pre-deliberation majority preference) while the tenth is a tough case where deliberation plays

But other changes could further reduce costs in ways that might allow for larger juries. One option would be the use of virtual trials (Benforado 2015), in which trial proceedings occur largely through computers, with an animated scene of the courtroom and avatars replacing human participants. The main potential benefits of virtual trials include blinding jurors and other participants to features like the ethnicity of the defendant, the rhetorical flourishes of the attorneys, and the body language of witnesses, which are likely to be misleading. Virtual trials also allow jurors to never be exposed to inadmissible evidence like hearsay, as opposed to hearing it but then being instructed to ignore it. But virtual trials also have the potential to cut down on costs. In principle, jurors could serve without having to travel to the courthouse (except perhaps for *voir dire*) and could perform their work in their own time, without necessarily needing to take time off work. This may enable much larger juries without substantially increased costs. While no doubt there are potential downsides to using virtual trials, they are deserving of consideration and should be studied further.

Another possibility for increasing jury size in a sense would be to make use of the so-called 'crowd within.' Herzog and Hertwig (2014) found that people could make more accurate estimates by making two estimates (ideally with a time delay in between, or with instructions to 'consider the opposite,' or play devil's advocate in one's mind, prior to the second estimate) and then combining them with linear averaging or some other method (see also Vul and Pashler 2008 and Ariely et al 2000). This allows potential increases in the number of probabilistic judgments without actually having a larger jury. It could thereby serve to improve jury accuracy while adding only negligible costs. As with virtual trials, the literature on the crowd within is probably too new to warrant a full-throated endorsement, but it is likewise worth considering.

a major role, for good or for ill.

5 Conclusion

Considerable evidence from psychology suggests that unstructured group deliberation is a poor way of making group judgments. Groups can suffer from groupthink and polarization. Social and informational pressures can bias the deliberation in non-truth-conducive ways by making it sensitive to evidentially irrelevant factors like which members speak up early and often, and also which convey the most confidence (which is known not to correlate well with accuracy; see references in fn 6). Worse, members of traditionally disadvantaged groups tend to speak less and carry less influence than others. Moreover, deliberation erodes the probabilistic independence of members' judgments, rendering a number of theoretical results on the wisdom of crowds inapplicable.

And yet we still entrust one of the most weighty tasks in the legal system—determining the guilt of a criminal defendant—to this procedure. Not only do we permit juries to engage in deliberation, but we instruct and even require them to do so.

I have argued that we should reconsider this approach and have sketched alternative methods of having juries reach verdicts in ways that involve either no interaction (voting and averaging) or very limited, tightly constrained interaction (DELPHI). While I take no stand on which of these alternatives (or some other one not considered here) is superior, I contend that one of them should replace the current system.

In arguing that juries should not engage in deliberation and instead aggregate their judgments by some non-deliberative means, I am in effect arguing for one particular way of limiting both the information available to juries and the powers they have at their disposal. As noted at the outset, my proposal is in this respect opposed to the general approach favored by some theorists such as Larry Laudan. His specific concern is with rules of evidence and criminal procedure rather than jury deliberation, but he writes that his general preference 'would be for strengthening rather than weakening the powers of the jury' (2006, 215). And he points out that, historically, juries have had more power and been exposed to more information than today:

> The English have been using jury trials as the instrument for determining guilt at least since the thirteenth century. Through most of that time, juries were both active and robust. An early English jury would normally draft into its ranks those who had witnessed the crime. It would conduct its own inquiries, often including interviews outside of the courtroom proper...Beginning in the eighteenth century, the rot began to set in with the rise of professional prosecutors and defense attorneys. Judges came to insist that it was their role, not the jury's, to interpret the relevant questions of law that touched on the case before the court. Prosecutors and defense attorneys conspired to take away from jurors all of their independent investigative functions. The ideal juror came to be conceived as a person who, instead of knowing something about the crime and about the principals in the case, was completely ignorant of such matters...In sum, the jury was rendered inert, stripped of its investigative and interrogatory powers, firmly told that it was to decide questions of fact and not questions of the law (which had become the judge's territory), and that it was to do all this shielded by the rules of evidence from seeing and hearing much of the evidence relevant to the case. Jurors came to be subject to elaborate instructions from the judge about the law and about how they were to deliberate. They were given mandatory presumptions to 'aid' them in drawing inferences. They were firmly told to ignore certain items of evidence or to suppress memory of certain testimony they had heard. (ibid, 215-6)

Laudan regards this change in the role of the jury, in which it is blinded to certain information and limited in its powers, as an unfortunate development. I disagree. While I am quite sympathetic to many of Laudan's criticisms of specific aspects of evidence law, I think that the overall direction of change he identifies over the past few centuries is precisely the direction we should be heading in. While the particular ways in which the juror's powers and access to information were limited may have included many mistakes, that does not impugn the overall idea that we should be systematic in constraining what jurors are permitted to see, hear, and do in reaching a verdict. It just means that we must do it better, basing these constraints on the best empirical evidence and theoretical work available.

In favoring limiting the information and powers of the jury, my proposal is in line with recent proposals for civil trials made by Sunstein.[32] He is concerned by the evidence cited above that civil juries' post-deliberation punitive damages awards are unpredictable (and less predictable than if we took the mean of the jurors' pre-deliberation award preferences) and sometimes alarmingly high (and in particular often higher than any of the jurors' pre-deliberation award preferences), along with further evidence that civil juries fail to follow instructions designed to have punitive damage awards provide optimal deterrence (for which, see many of the papers in Sunstein et al 2002). In response, he proposes having judges 'take a firmer role in overseeing jury awards' and 'moving away from the jury and toward a system of civil fines, perhaps through a *damages schedule* of the sort that has been used in many areas of the law, including workers' compensation and environmental violations' (2002, 242). Somewhat oddly, he does not propose eliminating the role for jury deliberation in the civil case, focusing only on reducing the jury's power to determine the size of the damages award. In this respect, he endorses a structural parallel between the civil and criminal systems in which juries, through deliberation, determine guilty or liability, and then judges play a large or even exlusive role in determining punishment. I would instead propose that while juries should determine guilt or liability, they should do so non-deliberatively (leaving it open how the determination of punishment should be divided between judge and jury). Even so, Sunstein's proposals for civil trials are of a piece with my own proposal for criminal trials.

Importantly, in these proposals it is not the case that 'jurors are treated as simpletons,' as Laudan (2006, 217) complains is the case with respect to many rules of evidence law. Or at least, in proposing that juries not engage in deliberation, we are no more treating jurors as simpletons than we are treating teachers as simpletons in

[32] In his contribution 'What Should be Done?,' chapter 13 of Sunstein et al (2002). See also Schkade et al (1999).

proposing that they do blind grading, or treating doctors and pilots as simpletons in requiring them to use checklists. We are not treating such people as simpletons, but rather as fallible, limited agents who make certain systematic errors when reasoning both alone and in groups. And this is exactly what the evidence suggests that they (i.e. we) are. In the case of juries as elsewhere, we should respond to this evidence by, among other things, attempting to devise procedures so as to eliminate or at least mitigate the unfortunate effects that stem from these cognitive limitations. Ending the system of jury deliberation is just a start.

References

Anderson, L. and Holt, C. 1997. 'Information Cascades in the Laboratory.' *American Economic Review* (87): 847-62.

Apodaca v. Oregon. 1972. 406 U.S. 404.

Ariely, D., Au, W. T., Bender, R.H., Budescu, D.V., Dietz, C.B., Gu, H., and Zauberman, G. 2000. 'The Effects of Averaging Subjective Probability Estimates Between and Within Judges.' *Journal of Experimental Psychology: Applied* (6): 130-47.

Baron, R. et al. 1996. 'Social Corroboration and Opinion Extremity.' *Journal of Experimental Social Psychology* (32): 537-60.

Bassett, G.W. and Persky, J. 1999. 'Robust Voting.' *Public Choice* 99(3-4): 299-310.

Benforado, A. 2015. *Unfair: The New Science of Criminal Justice.* New York: Crown Publishers.

Brief for the ABA as *amicus curiae.* Scott David Bowen v. Oregon, United States Supreme Court, No. 08-1117, May 28, 2009.

Burgman, Mark. 2015. *Trusting Judgments: How to Get the Best out of Experts.* Cambridge: Cambridge University Press.

Canadian Judicial Council. 'Model Jury Instructions.' Available: https://www.nji-inm.ca/index.cfm/publications/model-jury-instructions/

Christenson, C. and Abbott, A. 2000. 'Team Medical Decision Making.' In Chapman and Sonnenberg (eds) *Decision Making in Health Care*, New York: Cambridge University Press, 273-6.

Condorcet, Marquis de. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Paris: Imprimerie Royale. (Reprinted New York: Chelsea Publishing, 1972).

Cooke, Roger. 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science.* New York: Oxford University Press.

Crown Court Bench Book, Mar. 2010. Available: https://www.judiciary.gov.uk/wp-content/uploads/JCO/Documents/Training/benchbook_criminal_2010.pdf

Dalkey, N.C. 1969. 'The Delphi Method: An Experimental Study of Group Opinions.' Report No. RM-5888-PR. The Rand Corporation.

Dalkey, N.C. and Brown, B. 1971. 'Comparisons of Group Judgment Techniques with Short-Range Predictions and Almanac Questions.' Report No. R-678-ARPA. The Rand Corporation.

Dawes, R.M., Faust, D., and Meehl, P.E. 1989. 'Clinical versus actuarial judgment.' *Science* 243(4899): 166874.

Dawkins, R. 2003. 'Trial by Jury.' In *A Devil's Chaplain*, Mariner Books. 38-41.

DeKay, M.L. 1996. 'The Difference Between Blackstone-Like Error Ratios and Probabilistic Standards of Proof.' *Law and Social Inquiry* 21(1): 95-132.

Dietrich, F. and List, C. 2004. 'A Model of Jury Decisions Where All Jurors Have the Same Evidence.' *Synthese* (142): 175-202.

Drabsch, T. 2005. 'Majority Jury Verdicts in Criminal Trials.' NSW Parliamentary Library Research Service Briefing Paper no. 15/05.

Dworkin, R. 1985. *Law's Empire.* Cambridge, MA: Belknap Press of Harvard University Press.

Estlund, D. 1994. 'Opinion Leaders, Independence, and Condorcet's Jury Theorem.' *Theory and Decision* 36: 131-62.

Estlund, D. 2000. *Democratic Authority.* Princeton: Princeton University Press.

Feddersen, T.J., and Pesendorfer, W. 1998. 'Convicting the Innocent.' *American Political Science Review* (92): 23-35.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford: Oxford University Press.

Galton, F. 1907a. 'One Vote, One Value.' *Nature* 75: 414.

Galton, F. 1907b. 'Vox Populi.' *Nature* 75: 450-1.

Gawande, Atul. 2010. *The Checklist Manifesto.* New York: Metropolitan Books.

Genest, C. and Zidek, J.V. 1986. 'Combining Probability Distributions: A Critique and an Annotated Bibliography.' *Statistical Science* 1(1): 114-35.

Goldin, C. and Rouse, C. 2000. 'Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians.' *American Economic Review* 90 (4): 715-41.

Graefe, A. and Armstrong, J. 2011. 'Comparing Face-to-Face Meetings, Nominal Groups, Delphi and prediction markets on an estimation task.' *International Journal of Forecasting* (27): 183-95.

Grofman, B., Owen, G., and Feld, S. 1983. 'Thirteen Theorems in Search of the Truth.' *Theory and Decision* 15(3): 261-78.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., and Nelson, C. 2000. 'Clinical versus Mechanical Prediction: A Meta-Analysis.' *Psychological Assessment* 12(1): 19-31.

Gustafson, D., Shulka, R., Delbecq, A., and Walster, A. 1973. ‘A Comparative Study of Differences in Subjective Likelihood Estimates Made by Individuals, Interacting Groups, Delphi Groups, and Nominal Groups.’ *Organizational Behaviour and Human Performance* (9) 280-291.

Hans, V. 2008. ‘Jury Systems Around the World.’ *Cornell Law Faculty Publications.* Paper 305. http://scholarship.law.cornell.edu/facpub/305

Hastie, R., Penrod, S., and Pennington, N. 1983. *Inside the Jury.* Cambridge, MA: Harvard University Press.

Heath, C. and Gonzalez, R. 1995. ‘Interaction with Others Increases Confidence but Not Decision Quality: Evidence against Information Collection Views of Interactive Decision Making.’ *Organizational Behavior and Human Decision Processes* (61): 305-26.

Herzog, S. and Hertwig, R. 2014. ‘Harnessing the Wisdom of the Inner Crowd.’ *Trends in Cognitive Sciences* (18): 504-6

High Court of Australia. 1993. Black v. Queen. HCA 71.

Hong, L. and Page, S. 2012. ‘Some Microfoundations for Collective Wisdom.’ In Landemore, H. and Elster, J. (eds), *Collective Wisdom: Principles and Mechanisms*, Cambridge: Cambridge University Press.

Kagehiro, D and Stanton, W. 1985. ‘Legal vs. Quantified Definitions of Standards of Proof.’ *Law and Human Behavior* (9): 159-78

Kalven, H. and Zeisel, H. 1966. *The American Jury.* Boston: Little, Brown.

Janis. I. 1982. *Groupthink: psychological studies of policy decisions and fiascoes.* Boston: Houghton Mifflin.

Laddaga, R and Loewer, B. 1985. ‘Destroying the Consensus.’ *Synthese* (62): 79-95.

Laudan, L. 2006. *Truth, Error, and Criminal Law: An Essay in Legal Epistemology.* Cambridge: Cambridge University Press.

Laudan, L. 2008. ‘The Elementary Epistemic Arithmetic of Criminal Justice.’ *Episteme* 5(3): 282-94.

Laudan, L. 2010. ‘Need Verdicts Come in Pairs?’ *International Journal of Evidence and Proof* 14: 1-24.

Levy, D.M. and Peart, S. 2002. ‘Galton’s Two Papers on Voting as Robust Estimation.’ *Public Choice* 113(3/4): 357-65.

Lewis, D. 1989. ‘The Punishment that Leaves Something to Chance.’ *Philosophy and Public Affairs* 18(1): 53-67.

List, C. and Pettit, P. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford: Oxford University Press.

MacCoun, R. 2002. 'Comparing Micro and Macro Rationality.' In Fox, J. and Gowda, R. (eds) *Judgments, Decisions, and Public Policy*, Cambridge: Cambridge University Press, 116-37.

Pattern Criminal Jury Instructions for the District Courts of the First Circuit. 2014 ed. rev. Available http://federalevidence.com/pdf/JuryInst/1st.Crim.6.2014.pdf

Pritchard, M. and Keenan, J. 2002. 'Does Jury Deliberation Really Improve Jurors' Memories?' *Applied Cognitive Psychology* (16): 589-601.

Rawls, John. 1999. *A Theory of Justice*, revised edition. Originally published 1971. Cambridge, MA: Belknap Press of Harvard University Press.

Regan, H., Colyvan, M., and Markovchick-Nicholls, L. 2006. 'A Formal Model for Consensus and Negotiation in Environmental Management.' *Journal of Environmental Management* 80: 167-76.

Rowe, G. and Wright, G. 1999. 'The Delphi Technique as a Forecasting Tool: Issues and Analysis.' *International Journal of Forecasting* (15): 353-75.

Russell, J., Hawthorne, J., and Buchak, L. 2015. 'Groupthink.' *Philosophical Studies* (172): 1287-1309.

Schkade, D., Sunstein, C., and Kahneman, D. 1999. 'Are Juries Less Erratic than Individuals? Deliberation, Polarization, and Punitive Damages.' John M. Olin Law and Economics Working Paper no. 81.

Schkade, D., Sunstein, C., and Kahneman, D. 2000. 'Deliberating about Dollars: The Severity Shift.' *Columbia Law Review* 100 (4).

Schumann, E. and Thomson, W.C. 1989. 'Effects of Attorney's Arguments on Juror's Use of Statistical Evidence.' Unpublished manuscript.

Sommers, S. 2006. 'Diversity and Group Decision Making: Identifying Multiple Effects of Racial Composition on Jury Deliberations.' *Journal of Personality and Social Psychology* (90): 597-612.

Sunstein, C. 2011. 'Deliberating Groups versus Prediction Markets (or Hayek's Challenge to Habermas).' In Goldman and Whitcomb (eds) *Social Epistemology: Essential Readings*, New York: Oxford University Press.

Sunstein, C., Hastie, R., Payne, J., Schkade, D., and Viscusi, K. 2002. *Punitive Damages: How Juries Decide.* Chicago: University of Chicago Press.

'The Jury is Out.' *The Economist.* 14 Feb, 2009: 65 (EU). Retrieved 8 July, 2016.

http://www.economist.com/node/13109647

Tribe, L. 1971. ‘Trial by Mathematics: Precision and Ritual in the Legal Process.’ *Harvard Law Review* (84): 1329-93.

Vul, E. and Pashler H. 2008. ‘Measuring the Crowd Within: Probabilistic Representations within Individuals.’ *Psychological Science* (19): 645-7.

Wagner, C. and Lehrer, K. 1981. *Rational Consensus in Science and Society.* Dordrecht Reidel.

Wansley, M. 2013. ‘Scaled Punishments.’ *New Criminal Law Review* (16): 309-63.

Waters, N. and Hans, V. 2009. ‘A Jury of One: Opinion Formation, Conformity, and Dissent on Juries.’ *Cornell Law Faculty Publications*, Paper 114.

Woudenberg, F. 1991. ‘An Evaluation of Delphi.’ *Journal of Economic Perspectives* (40): 131-50.