Inquiry and deliberation in judicial systems

The problem of jury size

Angere, Staffan; Olsson, Erik J; Genot, Emmanuel

Published in:
Perspectives on Interrogative Models of Inquiry

DOI:
10.1007/978-3-319-20762-9_3

2016

Inquiry and Deliberation in Judicial Systems: the Problem of Jury Size

Staffan Angere, Erik J. Olsson, and Emmanuel Genot

Abstract: We raise the question whether there is a rigorous argument favoring one jury system over another. We provide a Bayesian model of deliberating juries that allows for computer simulation for the purpose of studying the effect of jury size and required majority on the quality of jury decision making. We introduce the idea of jury value (*J*-value), a kind of epistemic value which takes into account the unique characteristics and asymmetries involved in jury voting. Our computer simulations indicate that requiring more than a >50% majority should be avoided. Moreover, while it is in principle always better to have a larger jury, given a >50% required majority, the value of having more than 12-15 jurors is likely to be negligible. Finally, we provide a formula for calculating the optimal jury size given the cost, economic or otherwise, of adding another juror.

1. Introduction

The size of deliberating juries in court varies somewhat for different countries.[1] In the English speaking world, the number is usually 12, except in Scotland which has a 15-juror system. Yet there is a growing debate regarding the possibility of downsizing juries. A bigger jury is more expensive and difficult to administer than a smaller one, and, at least in smaller countries, a big jury can be difficult to assemble given the constraint that the same juror should not serve in consecutive trials. The pressure to downsize has led to some court cases where it has been ruled that smaller juries are admissible. Thus in the case *Williams v. Florida* (399, U.S. 78, 1970), the US Supreme Court ruled that the relevant part of the constitution, the Sixth Amendment, does not require juries to be composed of any specific number of jurors. In particular, six jurors should be allowed because "the essential feature of a jury obviously lies in the interposition between the accused and his accuser of the common sense judgment of a group of laymen" and, furthermore, "[t]he performance of this role is not a function of the particular number of the body that makes up the jury". The court added: "And, certainly the reliability of the jury as a factfinder hardly seems likely to be a function of its size."[2]

This ruling, which overturned an earlier Supreme Court decision in *Thompson v. Utah* (170, U.S. 343, 349, 1898) to the effect that the jury guaranteed by the Sixth Amendment consists "of

[1] Acknowledgement: this paper was written by Angere and Olsson, except the second part of section 6, which was written by Genot.

[2] *Williams v. Florida*, reprinted as pp. 3-70 in Jacobstein and Mersky (1998).

twelve persons, neither more nor less", stands in stark contrast to a recent evaluation of the Scottish 15 jury system which found that system to be, in the words of Cabinet Secretary for Justice Kenny MacAskill, "uniquely right".[3] In the consultation process, some advantages of 15-person juries were noted as being that people still have confidence in the system, larger juries lead to fairer verdicts, they are less likely to be influenced by prejudice, they allow for majority verdicts and are composed of a greater cross section of the public. Against this were arguments that 15-person juries often lead to unwieldy discussions and that the juror pool is being stretched by the requirement of having so many jurors for each trial.[4]

Given what seems to be a deep disagreement on the relationship between jury size and jury competence, it would be desirable to find a rigorous argument for either position, one than both parties to the debate were rationally obliged to accept. Obviously, we want jury deliberation to be as reliable a process as possible: we want someone to be convicted just in case he or she in fact did it. These considerations suggest the use of the famous Condorcet jury theorem, stating – among other things – that a larger voting body gives rise to a more reliable majority vote. It would seem, in the light of this mathematical result, that a deliberating body should be as large as possible, time and money permitting.

Unfortunately, the application of the Condorcet theorem to deliberating bodies is highly problematic. Condorcet's assumptions include that of independence of voting, which tends to be violated by deliberating bodies: in the process of deliberation, jurors will become increasingly influenced by each other's views. Furthermore, the theorem, in its standard formulation, requires everyone's likelihood of individually coming to the right answer to be above ½. While one may, optimistically, hold that individual jurors tend, on average, to be right more often than not, it is not clear how the presence of the occasional statistical outliers affects the result. It is obviously not enough that a *majority* of the jurors have a chance of more than ½ to be right, since this is compatible with almost 75% of the votes finally cast to be wrong.[5]

In an effort to overcome some of these limitations this paper proposes a different model, called Laputa, which allows for a process of group deliberation and inquiry. The model does not assume that jurors cast their final votes independently but only that jurors, where they contribute to the deliberation process, do so based on their own evidence rather than based on

---

[3] MacAskill as quoted in an article in *The Scotsman* (Forsyth and Macdonell, 2009).

[4] Forsyth and Macdonell (2009).

[5] See List and Goodin (2001) for generalized versions of the Condorcet theorem. See also Goodin (2003) for an extended discussion.

the evidence they received from others in the group.[6] Laputa is fundamentally Bayesian and decision-theoretic in nature. Naturally, the jury process has been investigated from similar perspectives before, beginning with Kaplan (1968). However, these studies, as far as we can tell, have not taken into account the deliberation process and its possible effect on the voting outcome which is not surprising given the mathematical complexity any such study would have to grapple with. In the present article, the computational problem is solved by focusing on the method of computer simulation rather than on that of analytical proof. In our understanding of juror inquiry, we settle for a model which in certain respects generalizes Jakko Hintikka's well-known interrogative model of inquiry. We will offer some remarks about the relations between Hintikka's model and our modeling assumptions, and give further details in the discussion section.

2. A probabilistic model of jury deliberation

There are several features that set juries apart from many other deliberative bodies and that will play a role in motivating our model:

Random selection of jurors. While the exact process whereby the jurors are selected varies widely, the usual case is that they are selected randomly from a precompiled list of eligible jurors. There may be various screening processes designed to exclude jurors that whose impartiality could be questioned. Also, it is considered desirable that the jurors come from varied backgrounds and provide a representative sample of the population. For example, a jury consisting of only Wall Street bankers, or only Mexicans, or only women, would be considered inappropriate.

Layman jurors. The jurors are supposed to be laymen and not experts.

Binary question. The jury's task is to deliberate on the question whether the accused is guilty or not. There is, in general, no third alternative.[7]

---

[6] For more details on the Laputa model and its interpretation, see Olsson (2011, 2013). See also Olsson and Vallinder (2013), Vallinder and Olsson (2013a, 2013b).

[7] In the Scottish legal system, a jury can also give the verdict "not proven". As some commentators (e.g. Luckhurst, 2005), have noted, including this verdict alongside the not guilty verdict has no legal consequence: in both cases the accused goes free and cannot be tried again for the same offence. Instead it is common to argue that the value of the not proven verdict is not legal but social: it allows the jury, for better or worse (probably the latter), to acquit the defendant while leaving a stain on his or her character. Alternatively, it can be used to "expose a poor investigation and highlight the failings of an incompetent

Restricted evidence. There are some restrictions on the evidence that the jury can appeal to in the process of deliberation. The jury is supposed to be present in court to hear all the evidence presented there. This evidence includes not only the written or spoken material presented but also the observed reactions of the accused, the witnesses, and so on. Juries are often instructed to avoid learning about the case from any source other than the trial (such as from media accounts) and to refrain from conducting their own investigations (such as independently visiting a crime scene). Parties, lawyers, and witnesses are not allowed to speak with jury members. Jurors are, however, allowed to appeal to their own general life experience in the process of deliberation.

Public announcements within the jury. Finally, while in the deliberation room, any contribution to the discussion made by some juror is available to all the other jurors. It would be unusual, and probably inappropriate, for some jurors to discuss matters "in private" without the knowledge of the other jurors.

As we propose to model jury deliberation, at every point in time a juror may, with varying degrees of competence, conduct inquiry, communicate with the other jurors, or both. Conducting inquiry here means consulting memory or notes about what happened at the trial or about other relevant things, such as the juror's own life experience. It does not include conducting investigations outside the court. Inquiry results in a reason for or against the guilt of the accused. As we conceive of reasons, they need not be interpreted as conclusive. If a juror has conducted inquiry, he or she may announce the result to the other jurors in the form of a pro or con reason (vis-à-vis guilt). These other juror's will react to the information by updating their cognitive states. This process will continue until time is up, at which point the jurors cast their individual votes.[8]

In the light of this initial characterization of the jury deliberation process we need to represent the following in the language of probability theory: (a) a juror's reliability, (b) a juror's cognitive state, and (c) how a juror's cognitive state is updated as the effect of receiving a

---

prosecutor" (Luckhurst, 2005). Since the not proven verdict has no legal consequence we have decided not to take it into account in this study of legal decision making.

[8] In some jury systems, such as the American, the condition specifying when the deliberation has come to an end does not refer to time but to some other feature of the situation, such as the jurors having reached a unanimous verdict. We have decided to leave the study of such jury systems for another paper. Having said this, the simulation results we present below count indirectly against the American system, and it seems unlikely, in the light of those results, that the latter should be a serious competitor e.g. to the Scottish jury system as regards quality of collective decision making.

pro/con reason. Let us start with jurors' reliability. A juror can be more or less reliable in retrieving information from memory or notes. A juror's reliability in this regard can be modeled as the (objective) probability that any result of inquiry is true. At the outset, we allow for different jurors to have different levels of competence – from being wrong all of the time, to being right all of the time, and everything in-between.

We assume that a juror's cognitive state consists of three things: an assessment of the accused's guilt/innocence, a self-assessment, and an assessment of others. The assessment of guilt or innocence is represented as a subjective probability ("credence") in the proposition that the accused is guilty, i.e. a number between 0 and 1. A number close to 1 means that the juror thinks the accused is probably guilty. A number close to 0 means that the juror thinks the accused is probably innocent. The self-assessment records how reliable (trustworthy) the juror considers his or her own inquiry to be. Here we generalize a common assumption of Hintikka's interrogative model of inquiry by allowing an inquirer to be less than fully confident in the results or her inquiry (see section 6 for a detailed discussion). The assessments of others records, for each other juror, how reliable (trustworthy) the juror in question considers those other jurors to be.

While it is easy to represent the assessment of the accused's guilt or innocent in probabilistic terms, it is less clear how to model probabilistically a juror's self-assessment or assessment of others. Our main idea is that a juror's trust in a source (own inquiry or other jurors) can be represented as a *credence in the reliability of the source*. Thus, a juror's self-assessment can be thought of as the juror's credence in the proposition that she is a reliable inquirer. We assume that a juror's trust in a source to be represented as a *trust function*, i.e. an assignment of a credence to every possible degree of reliability. For instance, a juror may assignment a credence of 0.7 to the hypothesis that the source is telling the truth 90% of the time. Trust functions offer a probabilistic representation of a critical aspect of Hintikka's interrogative model, namely that reasoning from any evidence whatsoever always takes into account its source (cf. section 6).

Let us now turn to the question of how juror's cognitive states should be updated. A juror reacts to reasons emanating from inquiry or other jurors by only taking into account (a) whether the reason is a pro or con reason (vis-à-vis guilt), (b) her own (prior) credence in the guilt of the accused, and (c) her (prior) trust in the source. Internal details of reasons or arguments are abstracted from. This is an idealization yet one without which the model would probably become utterly, and unworkably, complex. Here our model departs slightly from Hintikka's own, which usually emphasizes the fine structure of reasons in insisting on strategic aspects of reasoning. But this apparent departure actually allows us to generalize Hintikka's model, as will be explained in §6. Independent support for making this idealization can be found in the

Persuasive Argument Theory (PAT) tradition in social psychology, as explained in Olsson (2013).[9] Moreover, it receives some support from the fact that jurors are supposed to be laymen and not experts: experts are more likely to care about the fine structure of reasons than are laymen. Above all, this way of construing the updating of cognitive states in response to reasons is supported by our statistical approach to the jury problem, as soon to be explained.

The single source case. Let g be the proposition that the accused is guilty. Suppose that a juror receives a pro reason from a source α. We can now compute the posterior credence in g (i.e. the credence in g after receiving information from some source) as well as the reliability of the source:

$C_{t+1}(g) = C_t(g \mid \alpha \text{ gives a pro/con reason})$

$C_{t+1}(\alpha \text{ is reliable to degree } r) = C_t(\alpha \text{ is reliable to degree } r \mid \alpha \text{ gives a pro/con reason})$

The many sources case. Suppose that a juror receives information from many sources $\alpha_1, \ldots, \alpha_n$ at the same time. How can we calculate the following probabilities?

$C_{t+1}(g) = C_t(g \mid \alpha_1 \text{ gives a pro/con reason}, \ldots, \alpha_n \text{ gives a pro/con reason})$

$C_{t+1}(\alpha_i \text{ is reliable to degree } r) = C_t(\alpha_i \text{ is reliable to degree } r \mid \alpha_1 \text{ gives a pro/con reason}, \ldots, \alpha_n \text{ gives a pro/con reason})$

We recall that the jurors have been chosen randomly from the population of eligible candidates that are representative of the entire population. In the normal course of events, this selection process should ensure a certain degree of independence of thinking among the jurors, so that the fact that one juror at a given point in the deliberation notes or remembers something from the trial will not by itself make it more likely that another juror will note or remember that same thing. We also recall that jurors give pro/con reasons directly as they find evidence in their own notes or recollections. These two considerations together justify assuming *source independence*:

(SI) Each juror assumes that the other jurors are reporting independently (conditional on the truth/falsity of g).

[9] For more on the PAT tradition, see Isenberg (1986).

Using source independence, the result of receiving information from multiple sources is calculable from data about the individual sources, just as in the single-source case (cf. Olsson 2013). The bottom line is that assuming source independence makes the model computationally workable and at the same time it seems reasonably realistic given the way in which jurors are selected and assumed to interact.

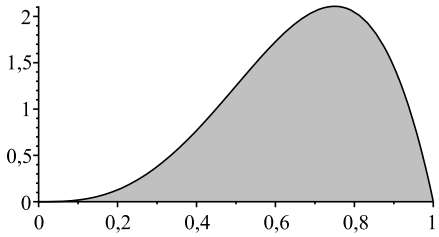
Now that we have a probabilistic model of the deliberation process, let us return to our main problem: to evaluate the effect of jury size on the jury's competence. Clearly we cannot solve this problem by looking at *just a few* deliberation processes while varying the size of the jury. If we do, we would not know whether the effect of adding more jurors was due to the size or to something else (difference in initial credence, individual competence, and so on). We need a way to study *the effect of size per se*. The solution to this problem is to study a large number of varied deliberation processes for a jury of a particular size and assess the average competence over all these processes. The competence pertaining to the jury size under consideration is the average, or expected, competence over all these particular deliberation processes.

This suggestion raises a worry regarding the practical possibility of performing all these competence calculations by hand. We propose to solve the computational issue by means of computer simulation. Our Laputa model has been implemented in a computer program that bears the same name and which automatically generates juries, allows the members to deliberate, in the idealized sense previously described, and, finally, collects data about the average reliability of the juries of the given size. Laputa can study millions of juries and deliberation processes in this fashion. Such considerations of scale also give an additional justification for treating reasons as "black boxes" without any internal structure because any persuasive effect that derives from the internal structure of reasons will be but a drop in a vast statistical ocean, or so we conjecture.

When Laputa generates a jury and a deliberation process it has to select initial values for various parameters. These parameters are, for each juror:

- prior credence in g, i.e. credence in g after court proceedings but before jury deliberation
- competence, i.e. probability that a result of inquiry is correct
- inquiry activity level
- communication activity level
- trust function for inquiry
- trust functions for other jurors

We configured Laputa to select these values according to a beta distribution with mean 2/3 and mode 3/4. This corresponds to the values $\alpha = 4$, $\beta = 2$, and its shape is plotted below.



The Beta distribution is congenial to Bayesianism, and has several useful properties:

1. Unlike the normal distribution, it is naturally clamped to [0, 1], and so does not need to be truncated. The normal distribution is not, as it is, possible to use to generate numbers in the interval [0, 1], but beta distributions with $\alpha \approx \beta$ are similar to normal distributions.
2. It simplifies several calculations, since it interacts well with conditionalization.
3. It has a straightforward statistical interpretation: for an inquirer beginning with a uniform distribution on all possible frequencies of a property *P* in a population, the beta distribution with parameters α, β gives the credence that inquirer should assign each frequency, given that he or she has observed α - 1 instances of *P* and β - 1 instances of not-*P* in that population.

For these reasons, we will use the above distribution whenever we want one whose expected value and peak are both between ½ and 1, symbolizing "somewhat better than average". This modeling decision corresponds to the limited degree of optimism embodied in applications of the Condorcet jury theorem, although it does not place any restrictions on the competences of individual jurors, rather than on their statistical mean.

3. Epistemic value in a jury situation

We will refer to the kind of epistemic value we aim to study as *Jury value* (*J-value*, for short). *J*-value should take into account: (i) the fact that it is the final state and not the difference between the final and initial states that is important, (ii) the fact that it is the *majority*'s opinion that counts, rather than the *average* opinion, and (iii) the fact that the jury situation is importantly

asymmetric, as embodied in Blackstone's principle "better that ten guilty persons escape than that one innocent suffer".[10]

Since we are dealing with majority voting it is important to settle on what we are to mean by "majority". Different justice systems differ in how large a majority is required for a verdict, from a simple >50% majority, up to unanimity. Requiring unanimity among 20 jurors will result in a fewer verdicts than requiring unanimity among, say, three jurors. A justifiable expectation, therefore, is that the value of having a certain number of jurors may depend on the size of the required majority. Hence, we will have to take into account different required majority sizes when we measure the expected *J*-value of a certain jury size.

Another parameter that is important for *J*-value is the credence required for voting for or against the guilt of the accused. In most legal traditions, a greater confidence is required for a conviction than for an acquittal. In a survey of American judges, the mean credence associated with the concept "beyond a reasonable doubt" was about 90% certainty (McCauliff 1982). For this reason we have chosen 0.9 as the credence in the guilt of a suspect required for a given juror to vote accordingly.

In the kind of trial we are dealing with, there are five possible relevant outcomes of a round of deliberations:

- conviction of the guilty (*CG*)
- conviction of the innocent (*CI*)
- acquittal of the guilty (*AG*)
- acquittal of the innocent (*AI*)
- no verdict (*NV*)

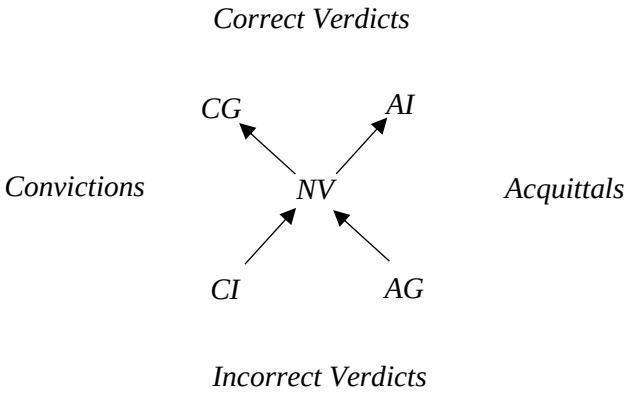
In the last case, we assume that the deliberations have to continue for another round. In epistemological terms, this corresponds to *status quo*, an outcome that may itself be connected with various costs.

We refer to the *J*-value of outcome *X* as *J*(*X*). Since it is always better to get the right verdict than no result at all, and always better to get no verdict than to get the wrong verdict, we postulate that

$$J(CG) > J(NV), J(AI) > J(NV), J(NV) > J(\text{AG}), \text{ and } J(NV) > J(CI).$$

[10] This principle has reappeared in many guises both before and after Blackstone, with a varying number of guilty acquittals held to be better than one innocent conviction (cf. Volokh, 1997).

These inequalities give rise to the following qualitative structure among *J*-values, where an arrow from outcome O_1 to outcome O_2 signifies that O_2 has a higher value than O_1:



How can we determine the outcome values more specifically? We could, of course, simply assign them conventionally, but we think that a better approach would be to try to ground them in specific features of the jury process. Since *J*-value is to be interpreted as a kind of *epistemic* value, it is not the *practical* consequences of the various outcomes that are to be assessed. We can think of the epistemic value of an outcome as the value it would have, from the point of view of an idealized judge, to be told the corresponding verdict. Still, the practical consequences are connected to the epistemic ones: the judge is generally *obliged* to follow the verdict of the jury, so if the judge is given the verdict that the suspect is guilty, the judge has to convict him or her, purely on basis of the epistemic situation.

This means that, from the perspective of the idealized judge, epistemic and practical value coincide. This is fortunate for us since it means that we can identify the *J*-values using decision theory. In general, utilities are determined only up to an affine transformation, and so it should be possible to assign two of the values arbitrarily. Interestingly, the particulars of the jury situation suggest more structure. The Blackstone ratio says that we should prefer acquitting 10 guilty men to convicting one innocent. What does this mean in terms of utilities? In order for Blackstone's principle to be interpretable at all, we have to assume that these are *additive* across cases. Thus the combined *J*-value of two verdicts will have to be the sum of the *J*-values of the individual verdicts.

Additive quantities have a clearly defined zero, which is the value of a type of situation *S* such that $J(nS) = J(S)$, where *nS* is the occurrence of *n* instances of *S*. In our case, *NV* is such a situation: it gives the judge no information at all and two verdicts, both of which are uninformative, contain exactly the same information as one. Therefore, we may set $J(NV) = 0$.

We still have the freedom to choose a *scale* for *J*-value arbitrarily. For reasons of mathematical simplicity we settle for $J(CI) = -10$. Using this value, together with the Blackstone ratio, we may draw the conclusion that the value of AG must be such that

$$J(CI) < 10 \cdot J(AG).$$

Since we assumed $J(AG) < J(NV)$, it follows that $J(AG)$ must be between 0 and -1. Given the intuitive disvalue in acquitting the guilty, we set $J(AG)$ to -1. While this is tantamount to judging that convicting the innocent is *just as bad* as letting 10 guilty men go, it only constitutes an infinitesimal deviation from the Blackstone principle.

$J(CG)$, the value of a correct conviction, is difficult to assess, and there seems to be little empirical work upon which one could rely for guidance. $J(CG)$ should certainly exceed $J(NV)$, the value of not arriving at a verdict, but how it should relate to $J(AI)$, the value of an innocent acquittal, seems impossible to determine on an *a priori* basis. Indeed, the literature contains arguments for $J(CG) > J(AI)$ (Tribe 1971) as well as for $J(CG) < J(AI)$ (Milanich 1981), and even for $J(CG) \approx J(AI)$ (Connolly 1987).

Since we have assumed additivity, there is an alternative way in which we characterize $J(CG)$. Let n be the total number of *guilty* suspects sent through the jury system for which a verdict is reached, and let c and a be the number of convictions and acquittals, respectively. By definition, we have $n = c + a$. The value $J(CG)$ can be calculated as the limit, as $n \to \infty$, of the ratio $\lambda = c / -a$ such that one should be indifferent between (a) adopting the jury system in question and (b) not making any verdicts at all. In short, λ records the number of guilty convictions it takes to undo the disvalue of a guilty acquittal.

One *J*-value remains to be assessed: $J(AI)$, the value of acquitting the innocent. We have already decided upon a degree of reasonable doubt. As it turns out, this degree in conjunction with the other *J*-values are sufficient to fix $J(AI)$ as well. To be rational, any juror should vote for conviction whenever the expected utility of doing so is greater than that of acquittal. Letting p be the juror's credence in the suspect's guilt, we should therefore have that

$$p \cdot J(CG) + (1 - p) \cdot J(CI) > p \cdot J(AG) + (1 - p) \cdot J(AI) \qquad (*)$$

iff $p > 0.9$. From this we derive that we therefore must have must have $J(AI) = 9\lambda - 1$.

Collecting our findings, we get the following table of *J*-values for the various outcomes:

| *Suspect*

Verdict	Guilty	Innocent
Conviction	λ	-10
No verdict	0	0
Acquittal	-1	9λ-1

Setting λ to 1, we get $J(AI) = 8$, corresponding to an assessment according to which each correct acquittal is as good as 8 correct convictions. For $J(AI) = J(CG)$, we need to set $\lambda = 1/8$. A lower value produces assessments for which a correct conviction is better than a correct acquittal. However, such low values of λ make the value of a correct conviction, as compared to an incorrect acquittal, strangely low, as pointed out by Connolly (1987).

Finally, the asymmetry between guilt and innocence means that the ratio of suspects who are actually guilty to those who are actually innocent will influence the result. Unfortunately, this is a figure which is extremely hard to assess in the present context. Despite its imperfections, the legal process is the best source we have for assessing the ratio in question. However, that source is unavailable in the present context because it is precisely the legal process that is currently under scrutiny. As an approximation, however, we may use the *conviction rate*, i.e. the percentage of cases brought to a jury which finally lead to conviction rather than acquittal. While this number varies from country to country, and also varies depending on the type of crime in question, it lies around 80% both in the U.S. and the U.K. (United States Courts 2010, Ministry of Justice 2011). Even if the actual number of guilty defendants deviates from this number, we have no evidence to suggest that such deviation would vary systematically in either direction. Given our limited knowledge, using 80% as an approximation of the percentage of guilty defendants seems to be at least a reasonable option.

4. Simulations based on *J*-value

We instructed the simulation program Laputa to compute, for each jury size, the average expected *J*-value over 1,000,000 juries of that size, each deliberating for 15 steps ("round table discussions"), with λ set to 1. We refer to such an expected value, for *n* jurors, as $E[J_n]$.

Running the simulation, we get the following figure (with number of jurors along the *x*-axis, and the resulting expected *J*-value, for different majority sizes required, along the *y*-axis).

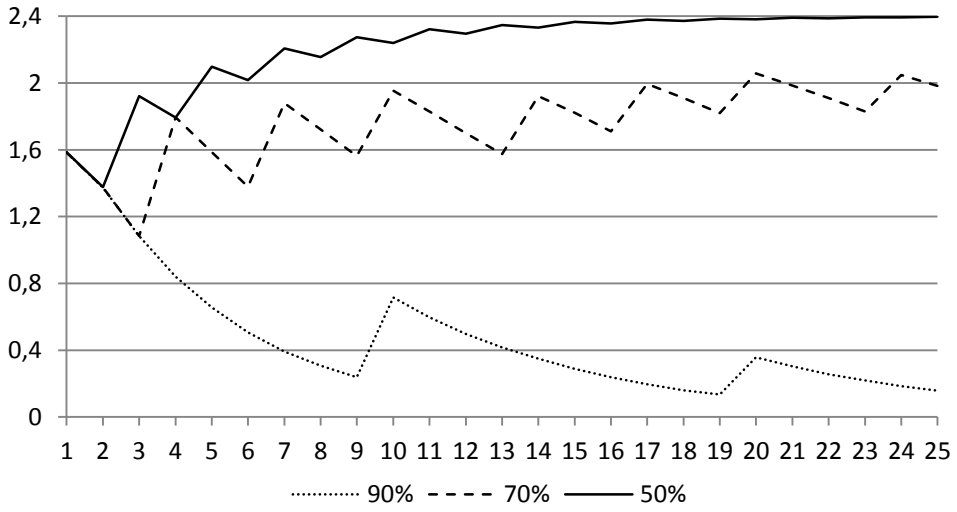


Figure 1: Expected *J*-values of different majorities and number of jurors.

The addition of more jurors clearly increases the *J*-value, at least for >50% and 70% required majority. When we require a 90% majority, the difficulty of getting a conviction means that less deliberations will lead to a verdict, and since this has a *J*-value of 0, the expected *J*-value will go to 0 as well. For a >50% required majority, adding more jurors makes the *J*-value approach 2.4, which is the theoretical maximum for the case where 80% of the defendants are actually guilty and $\lambda = 1$. For a 70% required majority, the maximum seems to lie around 2.0. As we see, the advantage of adding more than 15 jurors should, in many cases, be negligible.

One curious feature of the data is the "sawtooth" appearance of all curves in Figure 1. We can explain this effect as follows. A >50% required majority translates into a bigger majority required for a jury with an even, as opposed to an odd, number of jurors. With a 2-member jury, the only way to achieve >50% majority is through unanimity, whence there will be fewer verdicts than with just a single juror. For 4 members, it translates into 75%, while for 5, it requires only 60%. Hence, there will be fewer verdicts for an even number of jurors. Since the *J*-value of no verdict is zero this will tend to decrease the expected *J*-value for cases involving an even number of jurors, thus accounting for the sawtooth appearance of the curve.

To substantiate this hypothesis, the probability of not reaching a verdict can be measured using Laputa:

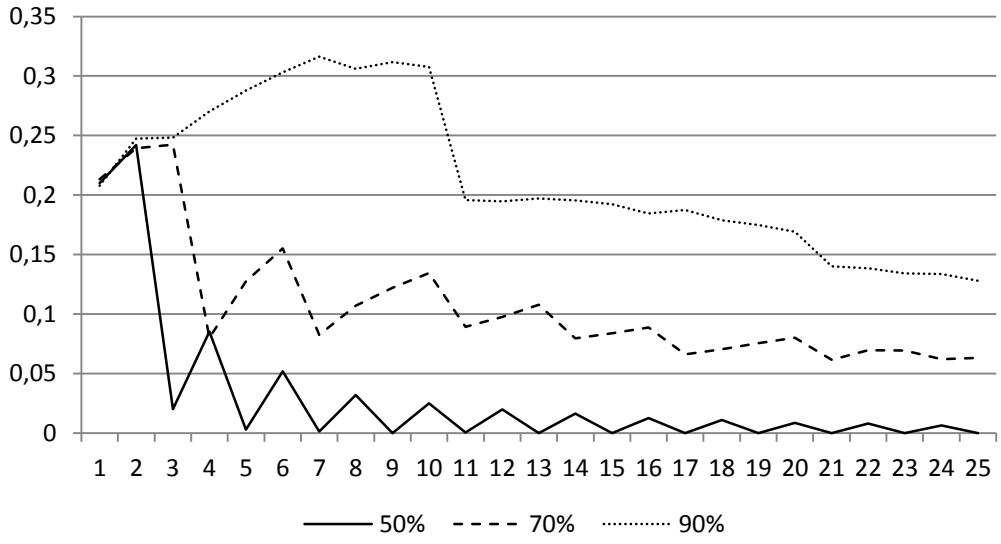


Figure 2: Probability of *NV* for different majorities and number of jurors

As expected, higher requirements on majorities give rise to a greater probability of not reaching a verdict. What may not be quite as expected, however, is that this probability decreases as the number of jurors is increased, in sharp contrast to what would be the case if the inquirers voted independently.

Our results so far indicate that no more than a >50% majority should be required for a conviction, even in criminal cases. We may further strengthen the support for this conclusion by showing that it holds independently of the proportion of defendants who are actually guilty in relation to all defendants. Below we have plotted the same data as in figure 1 under the assumption that *no* defendants are guilty.

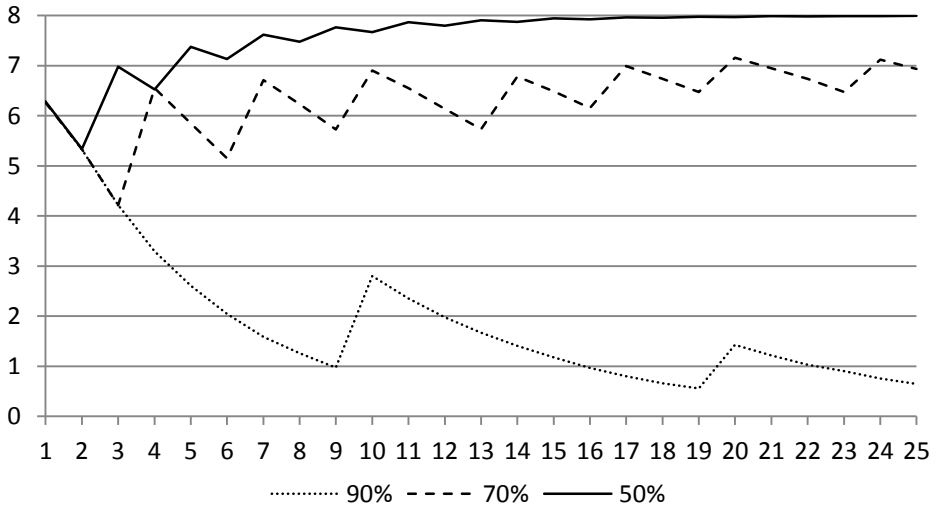


Figure 3: Expected *J*-values when defendant is innocent

Apart from the maximum *J*-value being 8 (the value of a correct acquittal) in this case the curves are almost indistinguishable from those in figure 1. This adds further support to the validity of our method since, as we noted, the proportion of actually guilty suspects is in general difficult to approximate in a non-circular manner.

Altering the parameters so that $\lambda = 0.125 = J(AI)$ and rerunning the experiments gives us the following result:

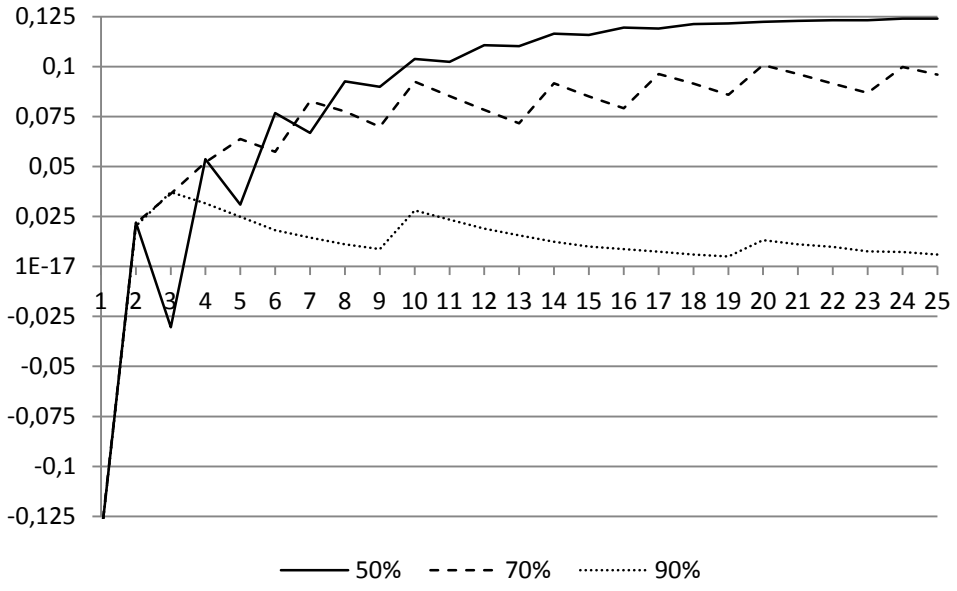


Figure 4: Expected *J*-values for $\lambda = J(AI) = 0.125$.

Here the scale is different and the maximum expected *J*-value attainable is 0.125 rather than 2.4. Apart from this, the graph is reminiscent of the one preceding it. The main difference lies in the fact that, when $\lambda = 0.125$, a very small number of jurors tends to give *negative J*-value, whence a jury with a single juror (or with three jurors, in case we require only >50% majority) is worse than no jury at all. This is due to the fact that, as λ decreases, correct convictions begin to affect the result more than correct acquittals. Since voting for conviction requires greater certainty than voting for acquittal, there will always be fewer correct convictions than correct acquittals. Making the latter count for less will therefore make it harder to offset the cost of incorrect verdicts.

There are several reasons why adding more jurors is beneficial to jury competence. One of them is that since more jurors means more results of inquiry, and these results get communicated to the whole jury, everyone will be better informed. But there is also the factor

that, generally, discussion tends to strengthen everyone's held opinions, and thus push their beliefs farther into certainty territory. Thus, after the deliberation process, more jurors will be willing to vote for guilt, reducing the number of unsuccessful attempts to reach a verdict as well as the number of erroneous acquittals.

The fact that discussion itself tends to strengthen prior opinion is obvious when a juror hears his or her own view echoed by the other jurors. But even hearing a divergent view can strengthen a juror's prior opinion, if he or she is willing to attribute the divergence to a general lack of credibility or even distrust on the part of the juror expressing the contrary opinion. For example, when a juror is convinced that *p*, hearing that not-*p* from some other juror may be interpreted by the first juror, via his or her trust function, as evidence to the contrary. This follows from the Bayesian treatment of trust used in Laputa and is, we believe, in accordance with human psychology.

5. Calculating the optimal jury size

Since, at least in the case of a >50% required majority, the addition of further jurors is conducive to epistemic jury competence, the question of an *optimal* jury size will have to involve a weighing against *other* values. While economy is an obvious value that may need to be given due weight, there are further values that concern the judicial process without being epistemological in kind. For instance, it is of interest for the defendant as well as the prosecutor that the trial proceeds as quickly as possible, and a greater number of jurors tends to slow down the process.

To simplify the problem, we will assume that the combined costs of adding more jurors are *linear* for each round of deliberation. When it comes to economic costs, this is probably indeed the case. With regards to other types of cost, it may at least be an admissible approximation. Let c be the non-epistemic disvalue of adding another juror; thus the total value of adding n jurors will be $-n \cdot c$. The interesting case will be when $c > 0$, as this will require an actual weighing of *J*-value against other values.

There is of course an extensive literature in value theory and economics about how to weigh or combine values.[11] A central theorem in this context was proved by Harsanyi (1955): when combining independent utilities, each of which satisfies the von Neumann-Morgenstern axioms, the only consistent choice is to use a weighted sum. We have already assumed *J*-value to be such a utility, and in the absence of any other well-developed theory of value, it is reasonable to take

[11] See, for example, Keeney and Raiffa (1976) and Broome (1991).

non-*J*-value to be in this class as well. Since we are only combining two forms of value, the weighing will be determined by a single number $w = w_c / w_J$, where w_J is the weight attached to jury value, and w_c the weight attached to other values. But this means that we can simply include w in c by measuring non-*J*-value using the same scale as *J*-value, so the total expected value of a practice, when applied to n persons, will be $V(n) = E[J_n] - n \cdot c$.

This is applicable primarily when the majority required is 50%. For higher majorities, the probability of *NV* becomes significant, and each such verdict also carries the costs of another round of deliberations, so the full formula would be given by the equation

$$V(n) = E[J_n] - n \cdot c + P(NV) \cdot V(n)$$

which can be solved to yield

$$V(n) = \frac{E[J_n] - n \cdot c}{1 - P(NV)}$$

The probabilities $P(NV)$ was given in figure 2 above. In order to be able to calculate a maximum, we need to represent both these functions and $E[J_n]$ as a continuously differentiable. This will, of course, involve a conventional choice of which function to use on our part. Among the usual functions available, those of the shape

$$A + B \cdot e^{C \cdot n + D}$$

turn out to approximate the functions we want to model best. Fitting such an exponential functions to the data points of the >50% required majority series of figure 1 gives $J^*(n) = 2.4 - 1.1753\ e^{-0.2150n}$, with a root mean square error of 0.075. We have plotted both $E[J_n]$ and J^* in the figure below.

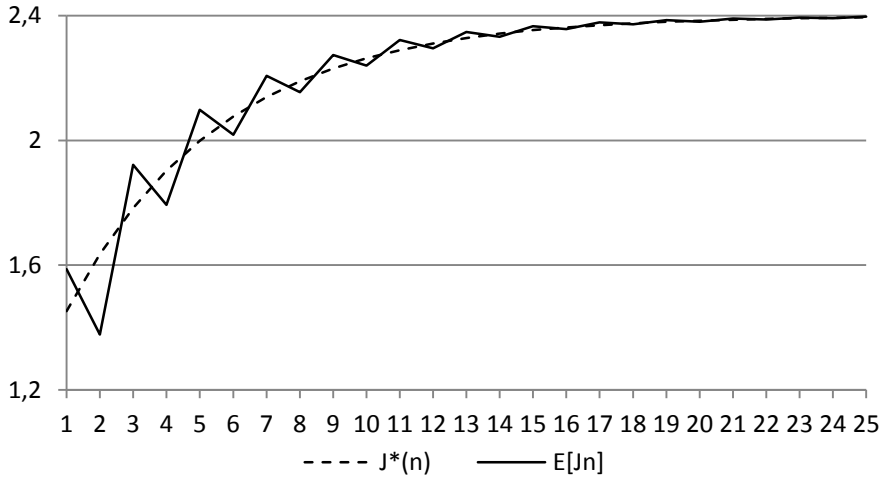


Figure 5: E[J_n] and $J^*(n)$ for 50% majority

Using J^* and similar continuous approximations of the J-value, we can find the optimal jury size by a simple optimization. Differentiating $V(n)$ with respect to n and setting this to zero to find the maximum, for each possible cost c of adding a single juror, gives the following figure.

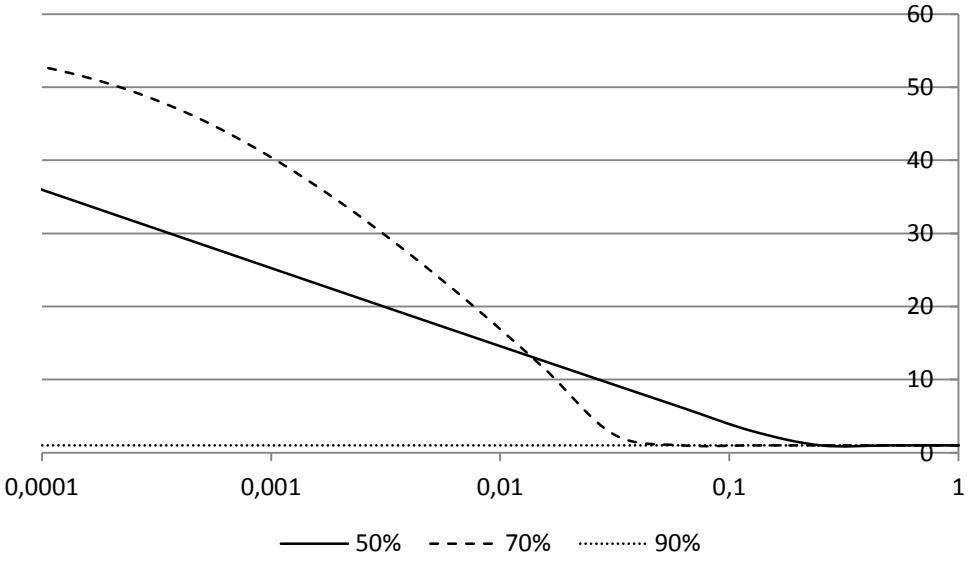


Figure 6: Optimal Jury sizes depending on cost of adding a new juror

We have assumed that there has to be at least one juror. With a 90% majority required, this is also the best number of jurors to have. For 70% and 50% majorities, the optimal number depends on c. For example, if the addition of a single juror has practical disvalue equal to a hundredth of the value of obtaining a correct conviction, i.e. $c = 0.01$ (remembering that we have

assumed λ = 1), a 50% majority system is best served by having around 15 jurors, and a 70% majority system by having around 18.

As in the case of λ, the determination of c will depend on personal values as well as on particularities of the specific justice system, such as the expense involved in adding a further juror. For this reason, it may very well be the case that what jury size is optimal differs not only between different countries but also within the courts belonging to one and the same country. What the present model gives us is a way to calculate such optima in a way that depends on these particular circumstances.

6. Discussion

In this section, we first discuss the consequences of various aspects of our model, and second, explain how it can be seen as generalizing Hintikka's model of interrogative inquiry (IMI) in certain respects. As we noted, several legal theorists have proposed to use formal decision theory for the purposes of investigating the jury process (Kaplan 1968, Connolly 1987, Arkes & Barbara 2002). Such attempts were severely criticized in Tribe (1971) for illegitimately disregarding the ritual aspects of a trial. This objection may indeed be well-founded so long as the purpose of a formal treatment is to *replace* the jury system, in this case with one based on decision theory. The purpose of our study is not to replace judicial procedure but to suggest possible ways in which that procedure could be improved.

For instance, our study indicates that requiring more than 50% majority should be avoided. This is a very stable recommendation which holds even if we count an incorrect conviction as a hundred times worse than a correct one. For another example, we suggested that having more than 15 jurors should be expected to add little perceptible epistemic value to the deliberation process. In the same vein, we could ask what degree of certainty *should be* required for a juror to vote for guilt. In the American justice system, jurors are informed about the "beyond reasonable doubt" requirement. In some states, they are, in addition, instructed how to interpret it (see Diamond 1990). Such instructions could potentially be based on simulations of the type we have been studying.

Since J-value is connected to the degree of certainty required for conviction through eq. (*), changing this value affects the relationship between the values of $J(CG)$ and $J(AI)$ as well. When we allow p (the required credence in question) to vary, we have the more general determination

$$J(AI) = \frac{p(\lambda + 1)}{1 - p} - 10$$

of the value of acquitting the innocent, given a value of λ. This is useful, since despite the fact that people generally *report* 90% certainty as what they require for reasonable doubt, actual studies show that they tend to *vote* on much lower certainties. According to Dane (1985), measuring the jurors' value judgments and then calculating the threshold from these results in an astonishingly low threshold of roughly 52%. As Dhami (2008) shows, the same result is obtained *even if the jurors were told to judge the defendant innocent unless they were 90% certain of his truth*. Not only is this an excellent illustration of how badly we tend to estimate our own degrees of belief; it also highlights the importance of doing experiments with a wide range of parameter values, especially if we are interested in measuring the effectiveness of actual juries as opposed to merely ideal ones.

If, following the findings of Dane and Dhami, p is set to 52%, we get the following relationship between λ and $J(AI)$:

$$J(AI) = \frac{13\lambda - 107}{12}$$

From this it follows that as long as λ is at least $8\,{}^{3}/_{13}$, $J(AI)$ will be positive, and at $\lambda = 107$, $J(CG)$ and $J(AI)$ will be equal. The resulting expected J-values for the latter case are plotted below, for the majority amount of 50%.

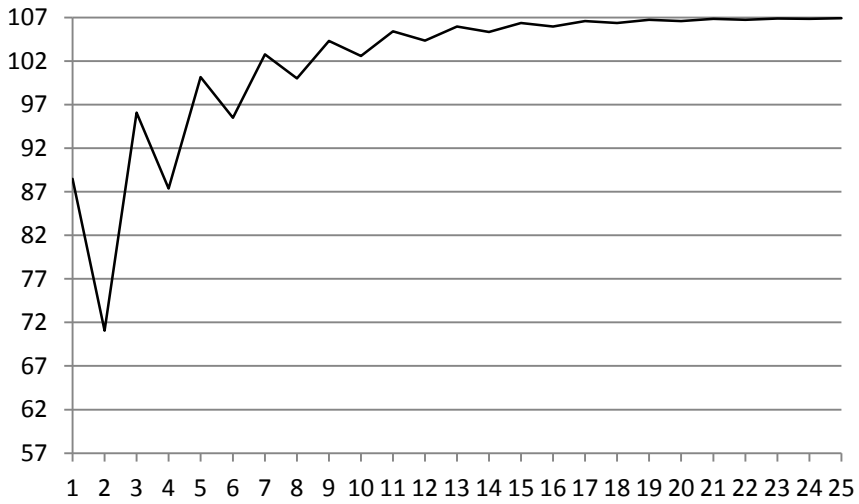


Figure 7: Expected J-values when p=0.52

The *shape* of the curve is certainly similar to the shape of the >50% required majority curve in the earlier figures, which means that our choice of an inverse exponential function as an approximation for use in the optimization problem remains valid.

However, because eq. (*) connects p with λ, it is hard to compare *cardinal* values with the case $p = 0.9$. At first sight it might, for instance, seem like setting $p = 0.52$ would be much better than setting it at 0.9, since the expected *J*-values are significantly higher for each possible number of jurors. But which part of this increase is caused by lowering p and which part is caused by increasing the values of *CG* and *AI*? There seems to be no way to separate these factors.[12]

So what if we were *not* to adjust λ, but only p? This would give us *J*-values in the same interval as before, but it would mean that we require jurors systematically to contradict decision-theoretic rationality. It also would not solve the fundamental problem: subjective probabilities and values are conceptually linked, so an adjustment of probabilities is generally impossible unless we adjust our values as well (cf. Jeffrey 1990).

It is important to see why this does not affect the conclusions we have reached so far: we have only compared jury methods using the *same J*-value assignments to one another, and in these cases the method we have given for calculating the optimal size of a jury remains valid. The difficulty arises only when we try to evaluate scenarios not only on the basis of the values the jurors *have*, but also on the basis of the values the jurors *should* have. Then it seems that we would need some kind of second-order value judgment which might be difficult to elicit in an objective manner.

Now for the second topic in this discussion. Let us explain why we consider our model to be a generalization, in certain respects, of the interrogative model of inquiry. In Hintikka's standard model, a lone inquirer attempts to answer some principal research question, using her background knowledge and answers to instrumental questions. The model essentially deals with the case of *pure discovery*, "a type of inquiry in which all we need to do is to find out what the truth is [and] we do not have to worry about justifying what we find" (Hintikka, 2007, p. 98). In such cases, inquiry terminates when the inquirer's background knowledge, together with the answers to instrumental questions she has gathered, implies deductively one of the answers to the principal question. The IMI illuminates the strategic role of deduction in the selection of questions and how the goal of inferring deductively an answer from strengthened assumptions guides the selection of instrumental questions.

---

[12] It is, of course, always possible to *scale* the *J*-values so that they have the same maximum and minimum, thereby achieving an illusion of comparability. But without an independent argument for why these maxima and minima *should* be the same such an approach would seem woefully *ad hoc*.

Jury deliberation, as modelled here, departs from pure discovery in at least two respects. The first concerns an assumption of restricted evidence. Evidence is essentially restricted to what transpired in court. In IMI terminology, at the time of deliberation, it is neither possible to ask new instrumental questions, nor to obtain answers to such questions previously asked. The second – the potential unreliability of information sources – is captured by assigning juror's assigning credence to information coming from inquiry or other jurors. Simply put, jury deliberation, unlike pure discovery, requires taking into account information both incomplete and uncertain.

While the IMI already accommodates reasoning from uncertain answers, it does so either by introducing probabilities, attached to uncertain answers, as reflecting their relative justification (Hintikka, 1987), or by introducing means to disregard (possibly provisionally) or "bracket" some background assumptions or instrumental answers when their justifications are questioned (Hintikka, 1998; Genot, 2009). Thus there is a sense in which IMI, unlike the present model, pays attention to what we referred to as the "finer structure of reasons". A common feature of these mechanisms, though, is that they encapsulate information about the sources of these answers. It has been argued that tracking multiple sources, IMI style, can account for reasoning patterns that *prima facie* violate Bayesian rationality (Hintikka, 2004), or vindicate some of the controversial axioms of AGM style belief revision in some contexts, but not in others (Genot, 2009).

An approach to jury deliberation based on the above mechanisms is possible in principle, but would in practice require tracking the many parameters that contribute to a single juror's epistemic evaluation. Our model represents the situation using only three parameters: the (current) credence assigned to the proposition that the accused is guilty, the (current) self-assessment of reliability; and the (current) assessment of other jurors' reliability. These three parameters allow us to abstract from the finer structure of reasons in the case of individual reasoning, but it is presumably more a difference in the level of process description, than a true divergence between models.

Abstracting from the details of the process by which jurors arrive at possibly uncertain answers to the principal question of guilt allows us "zoom out" to features that are specific to the multi-agent case, and to represent them explicitly. Simply put, one juror's preliminary answer to the principal question, at a given stage of deliberation, is at the next stage publicly announced, and becomes for all jurors part of the evidence to consider. New items of evidence are considered in the light of the trust one has in their sources (modeled by trust functions), and the total information is aggregated into a new preliminary answer, and a new assessment of trust. Hence, our model remains, we believe, compatible with the main tenets of Hintikka's

interrogative model. In addition, it generalizes Hintikka's model to the multi-agent case, and is to our knowledge the first systematic attempt at proposing and implementing formally such a generalization.

7. Conclusion

We have given a Bayesian model of deliberating juries for the purpose of studying the effect of jury size on group competence. We introduced the notion of *J*-value which takes into account the unique characteristics, asymmetries and values involved in jury voting. Our simulation results indicate that requiring more than a >50% majority should be avoided. Of the jury systems currently in use, it seems that only the Scottish system does not require more than a >50% majority. The British system, by contrast, requires a 10-2 (or 83%) majority, whereas the American prescribes unanimity. A further result of our study is that while it is in principle always better to have a larger jury, given a required majority of >50%, the value of having more than 12-15 jurors is likely to be negligible. More specifically, the optimal size of a jury appears to depend logarithmically on the non-epistemic cost of adding another juror. The Scottish system could potentially be further motivated by setting the value of a correct conviction to be the same as the disvalue of an incorrect acquittal, and the disvalue of adding a further juror to be a tenth of the value of a correct conviction. However, when different values are considered, different jury systems emerge as optimal.

These remarks are meant to be little more than suggestive hints as to how our approach could be relevant in practical cases. The extent to which our results apply to actual jury systems is an open question that we hope to be able to pursue in future work. Such an investigation would presumably involve addressing two limitations of our study. One concerns the fact that a jury trial is naturally divided into two stages: one stage at which the jurors listen to evidence presented at the court proceedings, and another at which they engage in closed room deliberations. It would be interesting to try to mimic these two stages in future simulations. A second limitation has to do with the problem of freeriding. Forming an independent judgment as to whether or not the defendant is guilty requires the weighing of evidence for or against the proposition in question, which in difficult cases can be a time and resource consuming activity. It is therefore attractive for a juror to decide to rely on the judgment of the other jurors rather than to form an independent opinion. If every juror delegates responsibility to the others, we have a serious freeriding problem, which may make the jury unable to reach a reliable majority verdict. Conceivably, as the size of the jury grows, the temptation to free ride increases, thus negatively affecting group competence. Various measures can be taken to counteract this

mechanism of social psychology, e.g. regularly reminding the jurors during the deliberation process of the great responsibility involved in serving in a jury, the importance of making an independent assessment and the dangers of group think. Our model as presented presupposes that such steps have been successfully taken. However, it might be interesting to take a more general strategy where the possibility of freeriding is part of the model.[13]

References

Arkes, Hal R. & Mellers, Barbara A. 2002. "Do Juries Meet Our Expectations?", *Law and Human Behavior* 26: 625-639.

Blackstone, Sir William. 1769. *Commentaries on the Laws of England*.

Broome, John. 1991. *Weighing Goods: Equality, Uncertainty and Time*, Blackwell.

Connolly, Terry. 1987. "Decision Theory, Reasonable Doubt, and the Utility of Erroneous Acquittals", *Law and Human Behavior* 11: 101-112.

Dane, Francis C. 1985. "In Search of Reasonable Doubt", *Law and Human Behavior* 9: 141-158.

Dhami, Mandeep K. 2008, "On Measuring Quantitative Interpretations of Reasonable Doubt", Journal of Experimental Psychology: Applied 14:353-363.

Diamond, Henry A. 1990. "Reasonable Doubt: To Define or Not To Define", *Columbia Law Review* 90: 1716-1736.

Forsyth, J., and Macdonell, H. 2009. "Scotland's Unique 15-strong Juries will not be Abolished", *The Scotsman*, 11 May.

Genot, E. 2009. "The Game of Inquiry: The Interrogative Approach to Inquiry and Belief Revision Theory", *Synthese*, 171:271-289

Goodin, R. E. 2003. *Reflective Democracy*, Oxford University Press.

Hintikka, J. 1987. "The Interrogative Approach to Inquiry and Probabilistic Inference", *Erkenntnis* 26(3): 429-442.

Hintikka, J. 2004. "A fallacious Fallacy?", *Synthese*, 140 (1-2): 25-35.

Hintikka, J. (2007), *Socratic Epistemology*, Cambridge Universty Press.

Hintikka, J., Halonen, I. & Mutanen, A. (1998). "Interrogative Logic as a General Theory of Reasoning", in Woods, J. & Johnson, R. (eds.) *Handbook of Applied Logic*, Kluwer.

Isenberg, D. 1986. "Group Polarization: A Critical Review and Meta-Analysis", *Journal of Personality and Social Psychology* 50 (6): 1141-1151.

[13] We owe the observation that there might be a free-riding problem in larger juries to Andrzej Wiśniewski.

Jeffrey, Richard. 1990. *The Logic of Decision*, 2nd ed., University of Chicago Press.

Jacobstein, J. M., and Mersky, R. M. 1998. *Jury Size: Articles and Bibliography from the Literature of Law and the Social and Behavioral Sciences*, Rothman & Co: Littleton, Colorado.

Kaplan, John. 1968. "Decision Theory and the Factfinding Process", *Stanford Law Review* 20: 1065-1092.

Keeney, R. L. and Raiffa, H. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, Wiley: New York.

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

Lockhurst, T. 2005. "The Case for Keeping 'Not Proven' Verdict", *The Sunday Times*, March 20.

McCauliff, Catherine M. A. 1982. "Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitutional Guarantees?", *Vanderbilt Law Review* 35: 1293-1335.

Milanich, Patricia G. 1981. "Decision theory and Standards of Proof", *Law and Human Behavior* 5: 87-96.

Ministry of Justice. 2011. "Criminal Justice Statistics. Quarterly Update to December 2010". Available online at http://www.justice.gov.uk/downloads/publications/statistics-and-data/criminal-justice-stats/criminal-stats-quarterly-dec10.pdf

Olsson, E. J. 2011. "A Simulation Approach to Veritistic Social Epistemology", *Episteme* 8 (2): 127-143.

Olsson, E. J. (2013), "A Bayesian simulation model of group deliberation and polarization", in Zenker, F. (ed.), *Bayesian Argumentation*, Synthese Library, New York: Springer, 113-134.

Olsson, E. J., and Vallinder, A. (2013), "Norms of assertion and communication in social networks", *Synthese* 190: 1437-1454.

Tribe, Lawrence H. 1971. "Trial by Mathematics: Precision and Ritual in the Legal Process", *Harvard Law Review* 84: 1329-1393.

United States Courts. 2010. "U.S. District Courts—Criminal Defendants Disposed of, by Type of Disposition and Offense (Excluding Transfers), During the 12-Month Period Ending March 31, 2010". Available online at http://www.uscourts.gov/Viewer.aspx?doc=/uscourts/Statistics/FederalJudicialCaseloadStatistics/2010/tables/D04Mar10.pdf

Vallinder, A., and Olsson, E. J. (2013a), "Do computer simulations support the argument from disagreement?", *Synthese* 190(8): 1437-1454.

Vallinder, A., and Olsson, E. J. (2013b), "Trust and the value of overconfidence: a Bayesian perspective on social network communication", *Synthese*, Online First: http://link.springer.com/article/10.1007%2Fs11229-013-0375-0

Volokh, Alexander. 1997. "*n* Guilty Men", *University of Pennsylvania Law Review* 146: 173-216.