The neurobiology of opinions: can judges and juries be impartial? [*]

Isabelle Brocas
University of Southern California
and CEPR

Juan D. Carrillo
University of Southern California
and CEPR

Abstract

In this article we build on neuroscience evidence to model belief formation and study decision-making by judges and juries. We show that physiological constraints generate posterior beliefs with properties that are qualitatively different from traditional Bayesian theory. In particular, a decision-maker will tend to reinforce his prior beliefs and to hold posteriors influenced by his preferences. We study the implications of the theory for decisions rendered by judges and juries. We show that early cases in a judge's career may affect his decisions later on, and that early evidence produced in a trial may matter more than late evidence. In the case of juries, we show that the well-known polarization effect is a direct consequences of physiological constraints. It is more likely to be observed when information is mixed, as behavioral evidence suggests, and when prior beliefs and preferences are initially more divergent across jurors.

Keywords: Neuroeconomic theory, neurobiology, information processing, Bayesian learning, polarization, jury selection.

[*] Address for correspondence: Isabelle Brocas or Juan D. Carrillo, Department of Economics, University of Southern California, Los Angeles, CA 90089, `<brocas@usc.edu>` or `<juandc@usc.edu>`. We are grateful to participants in the CSLP symposium on "Modeling human decision making in the law" for useful comments.

1 Introduction

Biases in beliefs and behavior have long been noted in many fields of Psychology and Economics. Although it is possible to categorize some specific biases and associate them with specific situations, the fundamental causes for biases are not well-understood. Moreover, even though the standard Bayesian framework does not capture the way beliefs are revised after exposure to evidence, it is difficult to pinpoint an alternative model that does strictly better in a large range of situations.

In a recent study (Brocas and Carrillo, 2012a), we have postulated that biases in beliefs and in decision-making may be mediated by physiological constraints that place a bound on the amount of information processed by neurons. The idea is intuitive. The information about the world is encoded in the sensory system and processed and decoded by a variety of systems before being mapped into a decision. If the encoding/decoding procedure is constrained, then the information transmitted to the decision-maker will differ from the information actually used to make a decision. In turn, the action of the decision-maker will appear biased to an outside observer. Interestingly, the neurobiology literature shows that such constraints exist and act in a very specific way. Brocas and Carrillo (2012a) link the neurobiological evidence about information processing to systematic biases in decision-making and beliefs. Brocas and Carrillo (2012b) extend the analysis to situations in which several decision-makers interact after privately making sense of the information they are exposed to. The objective of this article is to build on those studies to analyze the behavior of judges and juries. We will rely on results that have been demonstrated in Brocas and Carrillo (2012a, 2012b) and illustrate how they apply to this particular case.

The task of judges and juries consists in interpreting the evidence produced and make a decision based on it. In a purely Bayesian world, judges may take different decisions due to different opinions or intensity of preferences. However, posterior beliefs should be entirely driven by priors. In particular, two judges with the same prior belief and exposed to the same evidence should hold the same posterior. Similarly, in the Bayesian framework, members of a jury facing the same evidence should all update their beliefs in the same direction or else have views converging but we should never observe polarization. We show that these properties are not necessarily true when physiological constraints are taken into account.

More specifically, we show that physiological constraints generate biases in beliefs: decision-makers will tend to form posteriors that confirm their priors and that are affected by the magnitude of their payoffs or preferences (Proposition 1). Therefore, two individuals with the same prior and observing the same evidence will end up holding different posteriors if their preferences differ. Also, two individuals having different priors but the same preferences may update their beliefs in opposite directions (Proposition 2). These results offer a framework to address various implications on decisions rendered by judges and juries.

We show that physiological constraints make the order in which evidence is received critical (Proposition 3). Therefore, cases analyzed in the early career of a judge may affect the decisions that this judge will take on later a priori independent cases. Also early evidence produced in a trial may matter more than late evidence. Hence, it is not the same to be exposed first to strong evidence a crime has been committed, than to listen first to the childhood story of the criminal.

The case of juries is also interesting. In particular, we show that the distribution of preferences in a jury affects the way information is interpreted by individual jurors. If jurors are all willing to make the correct decision but have different priors or different preferences (and are therefore inclined to take to different decisions), each will interpret the available evidence differently. This may result in polarization, defined as the fact that two subjects with either different preferences or different priors may move their belief farther apart after being exposed to identical mixed evidence (Proposition 4). Such an outcome is in line with a long series of studies, and cannot be reconciled with traditional Bayesian theories of decision-making.

[xxx if we remove section 4.2 we should remove this paragraph xxx] Our study can also address a series of questions concerning the design of rules. We study the effect of the rule specified in case an unanimous verdict is not reached by a jury, and we show that it also affects the way information is interpreted by individual jurors (Proposition 5). Even though our model focuses on a particular case, it suggests that the rule affects strongly the probability of polarization.

To illustrate our theory in the simplest possible terms, we consider a problem with two underlying states of the world (e.g., a crime has been committed or not). Evidence is produced, aggregated and transmitted. It is positively but imperfectly correlated with the true state. Individuals who observe this evidence *interpret* it to make a decision (a verdict in the case of a judge or a recommendation in the case of a juror). Decisions are implemented yielding state-dependent payoffs for all the subjects involved.

The article is organized as follows. In section 2, we present the model consistent with the neurobiology evidence. In section 3, we study the case in which a decision is delegated to a judge. In section 4, we investigate recommendations by a jury. In section 5, we offer some concluding remarks. We will keep mathematical notations to a minimum in the main text. The reader should refer to the Appendix as well as to Brocas and Carrillo (2012a, 2012b) for the details of the model and proofs of the results.

2 The neurobiology of decision-making

To best fit the application to legal environments, we consider the following stylized situation. A person recently arrested for a crime is either guilty (state A) or innocent (state B). This person has to be evaluated by individual i (for example, a judge) who holds a prior belief p_i that the state is A. When evidence is produced, i interprets the information to update his belief, and takes the decision to convict (action a) or acquit (action b) the person. In this simple model, a is the correct action in state A (convict a guilty person) and b in state B (acquit an innocent person). We assume that i always wants to take the correct action. This is captured with the following payoffs:

$$U_i(a;A) = G_A^i > 0, \quad U_i(b;B) = G_B^i > 0, \quad U_i(b;A) = 0, \quad U_i(a;B) = 0$$

In words, wrong decisions are normalized to 0 and correct decisions yield positive payoffs. The parameters G_A^i and G_B^i capture the incremental utility of choosing correctly ($U_i(a;A) - U_i(b;A)$ and $U_i(b;B) - U_i(a;B)$). In this model, there is no scope for partisan recommendations: all individuals want to make the objectively correct decision and would do so under full information (a if $\Pr(A) = 1$ and b if $\Pr(B) = 1$).

We borrow the framework developed in Brocas and Carrillo (2012a) to formalize the actual decision-making process. Following the evidence from neurobiology, we model information processing and decision-making as a threshold mechanism in which the evidence produced is first encoded in the sensory system and then interpreted in reference to a threshold. We now introduce it more formally.

Let us first take a step back and abstract from our application to describe the conceptual framework developed in neurobiology. It is based on experimental research in which subjects are asked to extract relevant information from noisy evidence before making a decision. A typical paradigm is subjects having to identify a color or the direction of a movement and report their perception. They are rewarded when their answer is correct. Both behavior and neural activity are recorded and correlated.

The evidence produced is a signal about the underlying state, which is first encoded in the sensory system. For instance, the decision-maker must detect the color of an object placed in front of him and is given only a glimpse at the object under a certain light condition (the signal). Assume there may be two possible colors, ‘black’ or ‘white’. Neurons detecting each color will react according to the strength of the signal. In particular, the light intensity and conditions will affect cell firing.[1] We represent the encoded evidence by $c \in [0,1]$, which corresponds to the neuronal activity in the sensory system. The variable c can be interpreted as the ratio of neurons that detect the 'black' color. The signal is imprecise but informative: a high cell firing in favor of 'black' is (stochastically) more likely to occur when 'black' is the true color, while a low cell-firing in favor of 'black' is (stochastically) more likely to occur when 'white' is the true color. The same mechanism applies for the case we are interested in, with c representing the fraction of neurons supporting the hypothesis a crime has been committed given the evidence produced.

The next step is to understand how a decision is made based on the encoded information c. In a classical study, Hanes and Schall (1996) use single cell recording to analyze the neural processes responsible for the duration and variability of reaction times in monkeys. The authors find that movements are initiated when *neural activity reaches a certain threshold activation level*, in a winner-takes-all type of contest. This evidence suggests that the process can be schematically represented by a decision-threshold mechanism: it is as if there exists a threshold x such that action b is triggered when $c < x$ (release the person when there is enough evidence he is not guilty), and action a is triggered when $c \geqslant x$ (convict the person when there is enough evidence he is guilty). At the same time, it filters information out. In other words, the mechanism provides an 'interpretation' of the information. The sensory system collects c and the decision system interprets it as either $c < x$ or $c \geqslant x$, where the former is evidence of B and the latter is evidence of A. The decision system compares alternatives via this mechanism (see Shadlen et al. (1996), Gold and Shadlen (2001) and Ditterich et al. (2003) for further evidence). This type of threshold mechanism is widely used in neurobiology to account for behavior and neural activity.

It is important to realize that the threshold x represents actual neuronal thresholds and synaptic connections. Those are the physical elements through which the threshold mechanism is implemented: they filter information just as the threshold does. A neuron fires if it receives enough input activity from neurons in previous layers, which requires a strong synaptic connectivity between neurons. Depending on the level of neuronal thresholds and the strength of the synaptic connections, neuronal activity will be stopped or propagated along a given path and trigger one action or the other. This mechanism is economical in that it requires scarce knowledge to reach a decision.

We adopt this threshold mechanism and assume that the threshold x can be *opti-*

[1] Other factors affect cell variability. For instance, even when exposed to the same stimuli, neurons do not always fire in the same way. This can be understood as internal noise and may vary across individuals and across experiences. We will neglect this type of variability to concentrate on the effect of external variability on decision-making. For a detailed discussion, see Brocas (2012).

mized to maximize expected utility. This presupposes that the information is interpreted taking into account information regarding prior beliefs and the intensity of preferences. This also presupposes that the way this information is taken into account is compatible with expected utility theory. Assuming neurons can perform this type of optimization is consistent with the work by Platt and Glimcher (1999). The authors show that the brain represents the magnitude of the possible payoffs (G_A^i and G_B^i in our model) as well as their likelihood (prior beliefs and signals). They also show that neural activity correlates with decisions and fits an expected utility maximization model. In typical oculomotor tasks, rewards and information components modulate the activation of the lateral intraparietal area where the decision is computed given projections from the reward system (which evaluates payoffs) and the sensory system (which interprets the evidence).

Overall, decision-making in the brain can be represented by an 'as if' model with the following timing of events. First, a threshold is set. Second, evidence is encoded in the sensory system. Third, evidence is compared to the threshold. And fourth, an action is implemented. An optimal threshold is one that takes into account all the relevant information and realizes that it is filtered out when making a decision. Said differently, it maximizes the expected utility of the individual given the likelihood of the events, that is, given the evidence about the state that is retained.

Recall that in the threshold mechanism, a decision is based on a coarse partition of information: either there is reasonable evidence in favor of A ($c \geqslant x$) or reasonable evidence in favor of B ($c < x$). Therefore, depending on these two possibilities, i can hold ex post either of two posterior beliefs, $\Pr(A \mid c \geqslant x) \equiv \overline{p}_i(x)$ or $\Pr(A \mid c < x) \equiv \underline{p}_i(x)$, depending on whether $c \geqslant x$ or $c < x$. These beliefs are to be contrasted with those obtained in a pure Bayesian framework which would be based on the actual evidence collected c. Formally, in the standard Bayesian model, the posterior belief based on the evidence c would be $\Pr(A \mid c) \equiv \pi_i(c)$. Our first result is as follows.

Theorem 1 **[Brocas and Carrillo, 2012a]** *In the optimal threshold mechanism, decision-making is efficient but posterior beliefs are biased.*

It is efficient to take action a if the probability that the true state is A is high enough. In the standard Bayesian model, it means that there exists a certain amount of evidence x_i^* such that it is efficient to take action a for all $c \geqslant x_i^*$ and to take action b for all $c < x_i^*$. This means in particular that, for the purpose of decision-making, it is irrelevant whether the information obtained is slightly below x_i^* or far below it: the same decision is efficient either way (action b). Therefore, basing a decision on the coarse information partition $c < x_i^*$ or $c \geqslant x_i^*$ does not affect the ultimate decision. However, the posterior emerging from the threshold mechanism *is biased* because the intensity of the signal is emphasized or de-emphasized to keep just the information necessary to take the optimal action. Therefore, the individual will report an opinion that does not take into account the actual information he observed but reflects instead his (biased) interpretation.

We will exploit this result to analyze belief formation and decision-making in legal environments in which a decision must be made regarding a potential offense. We have in mind situations in which a judge must interpret the evidence available in order to make a decision. We will analyze such situations in section 3. We are also interested in situations in which a jury is delegated the decision and the verdict is subject to institutional rules (e.g., unanimous or majority voting). In such situations, each member of the jury will interpret the evidence provided and recommend the decision he/she thinks is best. Those situations will be studied in section 4.

3 Decisions rendered by a judge

The motivating example in this section is a judge (he) who must review evidence and make a decision based on his own judgement. The judge wants to take the correct action given his limited information and is endowed with the preferences described earlier.

3.1 Priors, preferences and biases in beliefs

We first address the question of whether posterior beliefs are biased in a systematic way. In particular, we are interested in determining if particular prior beliefs or particular attitudes towards outcomes may affect the interpretation of the evidence.

Proposition 1 *The interpretation of the evidence affects the judge's information processing in a way that (i) his prior beliefs tend to be confirmed, and (ii) his posterior beliefs depend on the payoffs of the different actions.*

This result is a special case of Brocas and Carrillo (2012a). Even though the judge is impartial, in the sense that he does not favor a decision and rather tries to take the one that fits best the state, he is subject to two biases. First, he tends to reinforce his prior: the higher his confidence in state B, the higher the threshold x_i^*. This means that he is more prone to interpret information as evidence of B (technically, $c < x_i^*$ is more likely) and update his prior towards that state.

The intuition for the confirmatory bias effect is as follows. Suppose the judge believes that the person is very likely to be innocent ($\Pr(B)$ is high) and is leaning towards acquitting him given the current evidence. He will require a lot of evidence of the person being guilty in order to change his mind. This is rational as the judge will need to be confident in his change of mind to compensate for his current belief. This means that the threshold should be high. It is therefore very likely the evidence will fall below the threshold. As a consequence, the judge is likely to interpret the evidence in favor of his initial prior and acquit the person eventually.

Second, the judge tends to favor the state in which highest payoffs can potentially be obtained: the higher the benefit of taking the correct action in state B, the higher the threshold. Again, he is likely to interpret the information as evidence of B and update his prior towards B. The logic of the argument is the same as before. Suppose the judge is most interested in releasing innocent persons (and only to a lesser degree, convict guilty persons). He will change his mind only if very strong evidence that the person is guilty is produced. Therefore, it is optimal to set a high threshold which, again, makes a change of mind very unlikely.

As mentioned, decisions are not biased if the threshold is set optimally. That is, the judge will take the same action as in the standard Bayesian framework where the signal is perfectly processed rather than interpreted. The main consequence, however, is that the judge will hold degrees of confidence in his decision that do not reflect the exact evidence produced. This would have an effect on future decisions if he were to re-evaluate past choices or obtain new information.

3.2 Opinions and impartial judgements

The previous section established that biases in beliefs can be linked to primitives: posterior beliefs are shaped by prior beliefs but also by the magnitude of the state-contingent payoffs. Even though a judge tries to be impartial, he may end up holding a belief that reflects his own aversion to a particular crime and report an opinion that does not seem in line with the evidence from the perspective of an outside observer. The next result makes this statement more transparent.

Proposition 2 *Two judges exposed to the same evidence may report different opinions. In particular, two judges sharing the same prior may end up having different posteriors, and two judges with different priors may interpret identical evidence in opposite ways.*

This proposition is a consequence of the results obtained before (Theorem 1 and Proposition 1). Two judges holding the same prior and exposed to the same evidence would hold the same posterior in the standard Bayesian model. However, in the threshold model, they may end up with different posteriors. This result can be due exclusively to a difference in preferences, and not to a difference in prior beliefs, a leading assumption to explain this type of phenomenon. Consider two judges, 1 and 2, who hold the same prior belief but different preferences, namely $G^1_A > G^2_A$. By Proposition 1, $x^*_1 < x^*_2$. Assume also that the evidence is 'mixed' in that it falls between the two thresholds, $c \in (x^*_1, x^*_2)$. In the standard model, the posterior beliefs are identical: $p_1(c) = p_2(c)$. In the threshold model, however, judge 1 revises his belief upwards (the threshold is surpassed) while judge 2 revises it downwards (the threshold is not reached). In other words, they disagree even though they share the same prior and are exposed to the same evidence (see Figure 1).[2]

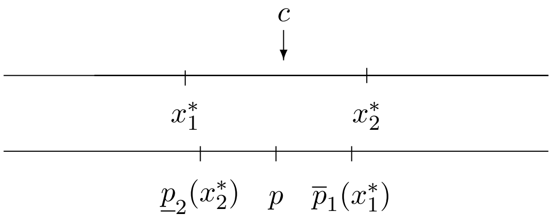

Figure 1: Judges having the same prior p but different preferences exposed to the same mixed evidence c.

Also, two judges having the same preferences but whose prior opinion differs, $p_1 \neq p_2$, may also interpret identical evidence in opposite ways. Consider for example two judges, 1 and 2, with priors $p_1 > p_2$. Again following Proposition 1, $x^*_1 < x^*_2$. Assume now that evidence is mixed, $c \in (x^*_1, x^*_2)$. In the threshold model, judge 1 will update his belief upwards as his threshold is surpassed while judge 2 will update his belief downwards as his threshold is not reached (see Figure 2), thereby leading to a polarization of opinions. The fact that people holding different priors may disagree after receiving the same evidence is, again, not novel. However, in the Bayesian model, subjects exposed to the same evidence would either update in the same direction (both upward or both downward) or else converge (the subject with the initially lowest belief increases it and the subject with the initially highest belief decreases it). Either way, they could never exhibit divergent beliefs.

The main lesson so far is that identical evidence will be interpreted differently by individuals with different preference intensities and/or different priors. Also, posterior beliefs are shaped by preferences. Individuals tend to confirm their priors and they tend to develop strong confidence in their decisions to avoid the least beneficial outcomes.

[2] Notice that they also take different actions (judge 1 chooses action a whereas judge 2 chooses action b). However, this is not surprising: it is also true in the standard Bayesian setting that two individuals with different preferences facing the same information may choose different actions.

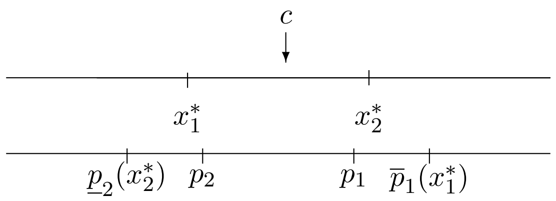

Figure 2: Judges having different priors p_1 and p_2 but the same preferences exposed to the same mixed evidence c.

3.3 Decision-making and beliefs over time

Recall that decision-making is efficient conditional on prior and preferences. Said differently, if at time t the judge holds a belief μ_t, the threshold is set in such a way that the optimal decision is taken after the release of date t evidence. However, in many applications, a judge is receiving several pieces of evidence before making a decision. Given posteriors are built vía a biasing filtering mechanism, the belief used when the last piece of evidence is interpreted reflects the previous filtering. Because such belief is biased at that point, the decision will itself be also biased from an earlier perspective.

Proposition 3 *Beliefs and decisions are path dependent. In particular, the order in which the evidence is received affects choices over time.*

To understand this proposition, consider a judge who initially believes the person is equally likely to be guilty or innocent and therefore sets an average threshold. He receives two pieces of evidence one (weakly) in favor of A and the other (weakly) in favor of B.

If he is first exposed to the evidence in favor of B, the average threshold is not reached. His confidence on B increases, and so does his threshold. The second piece of evidence (weakly in favor of A) does not reach this higher threshold either, so it does not reverse the judge's belief that b should be implemented. Conversely, if the judge is first exposed to the evidence in favor of A, the average threshold is surpassed. The threshold is lowered so the second piece of evidence (weakly in favor of B) also surpasses it. The judge ends up having two pieces of evidence above the thresholds and chooses action a.

Overall, the result shows that, in sharp contrast to a standard Bayesian framework, in the threshold model first impressions matter. Indeed, the thresholds generate endogenously an anchoring effect: the judge may 'adopt' the interpretation of the first piece of evidence and reinforce this interpretation as new evidence comes in. Under some specifications, we can show that thresholds also become more extreme over time (see Brocas and Carrillo (2012a) for details).

Proposition 3 has many implications. We discuss two that are particularly relevant for our application.

- Sequence of information revelation in trials. A judge is never exposed to one single piece of evidence before making a recommendation nor to a number of arguments presented simultaneously. The decision comes after a sequence of signals have been disclosed and interpreted. The previous result suggests that the order in which the evidence is produced may tilt the verdict. In particular, early disclosure of evidence that the person is innocent is more likely to be followed by a release. These anchoring effects are reinforced by the intensity of the preferences. A judge who values releasing innocent people relatively more than convicting guilty people and who is exposed first to favorable evidence on the person will be most likely to release the person. Of course, this is true only if the decision is taken under a certain amount of doubt (a common scenario). If unambiguous evidence is produced, any judge in our model would take the "correct" decision. We would also like to make a comment concerning the presumption of innocence principle. It reduces roughly to working under the assumption that the prior $\Pr(A)$ is low enough so that in the absence of evidence, the person should be acquitted. Given our results, the principle should affect the way a judge forms an opinion based on evidence, because thresholds are such that priors tend to be reinforced. A verdict rendered under the presumption of innocence and where the very first evidence produced suggests innocence is most likely to be favorable to the person.

- Wisdom or stubbornness. Judges are often appointed for long periods of time. The evidence produced in a given trial sometimes contains information pertinent for other trials. The judge may therefore learn about the underlying state of the world throughout his mandate. His prior belief in the first case of his career will therefore be different from his prior belief in a later case. The biasing effects emphasized here suggests that judges may not only develop opinions that reinforce earlier pieces of evidence but also take path-dependent biased decisions. A judge who convicts the first few persons is more likely to convict offenders also in the future. This may have implications on the optimal length of judges' mandates from the perspective of a planner.

4 Decisions rendered by juries

We now turn to study the verdict of juries. This problem has been analyzed theoretically elsewhere. Notably, Feddersen and Pesendorfer (1998) proposed a model of strategic voting by jurors to assess the merits of unanimous verdicts. Contrary to earlier literature in which it is assumed each juror behaves as if her vote alone determines the outcome, the authors assume that each juror possesses private information about the state (e.g., technical knowledge) and makes a strategic vote.[3] Our perspective is different. Instead of assuming jurors differ in their private information, we assume they differ in either their taste (preferences over outcomes) or initial opinion (prior beliefs). Also, rather than modeling the information produced during the trial as a private signal that is processed in a pure Bayesian manner, we model it as a common signal which is processed through a threshold mechanism. Our perspective is related however because we do presuppose a strategic interpretation of information. In other words, each juror takes into account that other jurors interpret the information to make a recommendation and that recommendations are aggregated given the rule.

The analysis in this section borrows results from Brocas and Carrillo (2012b). With respect to the previous section, we extend the model in the following way. The jury is composed of n jurors indexed by i. When evidence is produced, each juror interprets the information to form an updated belief. Each juror makes a recommendation r_i and the judge makes a decision based on the reports of all jurors. Jurors are heterogenous. To keep the analysis simple and tractable, we assume that there are two possible types of jurors indexed by k. The proportion of type-1 and type-2 jurors is ν and $1-\nu$. Each type-k juror ($k=1,2$) has a prior belief p_k that the person is guilty. Preferences across groups are given by:

$$U_k(a;A) = G_A^k > 0, \quad U_k(b;B) = G_B^k > 0, \quad U_k(b;A) = 0, \quad U_k(a;B) = 0$$

All jurors are exposed to the same evidence and the encoded information is c for all of them, so there are no differences in private signals. When making a recommendation, each juror anticipates how the verdict will be rendered given the recommendations made in his group and the recommendations made in the other group. Therefore, recommendations are strategic.

In the next subsections, we assume that given the primitives of the model (G_A^1, G_B^1, p_1, G_A^2, G_B^2, p_2), were jurors delegated the decision, the thresholds x_1^* and x_2^* would be

[3] Many other studies in political economy have analyzed the effect of strategic voting. See for instance Austen-Smith and Banks (1996), Feddersen and Pesendorfer (1996) and Myerson (1998).

such that $x_1^* < x_2^*$.[4] That is, preferences are such that type-1 jurors are more willing to convict the person than type-2 jurors.

4.1 Polarization

In this section we restrict the attention to the case where the judge requires an unanimous vote. Whenever jurors do not agree, a default option o is implemented. We assume that the default option is a 'compromise' option. To capture this idea, we assume that $U_k(o|A) = U(o|B) = G^k$ and, to simplify the analysis, we consider the case illustrated in Figure 3 where $G^k = G_A^k G_B^k/(G_A^k + G_B^k)$. In our construction, juror-k possessing a posterior belief below $G_B^k/(G_A^k+G_B^k)$ prefers b to o and o to a. Similarly, juror-k possessing a posterior belief above $G_B^k/(G_A^k + G_B^k)$ prefers a to o and o to b.

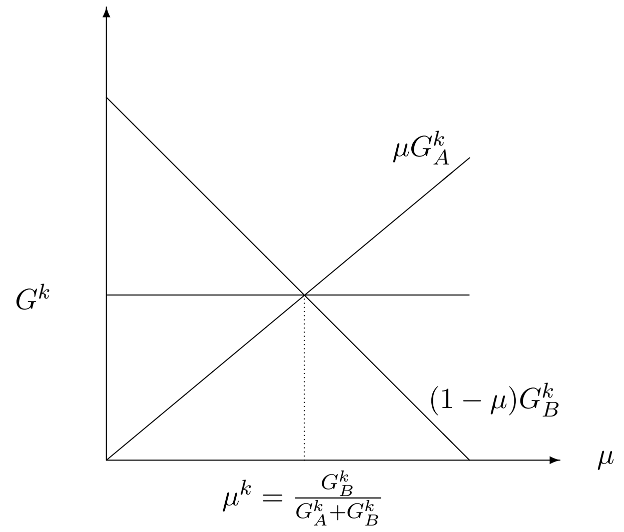

Figure 3: Expected payoffs as a function of the posterior belief

Consider now the decision of a type-k juror facing evidence c. Evidence c below the threshold x_k triggers recommendation b while evidence c above the threshold triggers recommendation a. Denote by x_2 the threshold used by type-2 jurors, and consider thresholds x_1 such that $x_1 < x_2$. If $c \leqslant x_1$, all jurors agree on recommendation b and if $c \geqslant x_2$, all jurors agree on recommendation a. When $c \in (x_1, x_2)$, type-1 jurors recommend action a while type-2 jurors recommend action b, in which case o is implemented. From the perspective of a type-1 juror, the payoff obtained if $c \leqslant x_1$ is the expected payoff of taking

[4] For example, $G_A^1 > G_B^1$, $G_A^2 < G_B^2$ and $p_1 = p_2$.

action a given the posterior belief $\underline{p}_1(x_1)$. The payoff obtained if $c > x_1$ however depends on whether the evidence was sufficient or not to trigger the same recommendation from type-2 jurors, that is on whether $c \gtrless x_2$. We have the following result.

Proposition 4 *Type-k juror set the optimal threshold at x_k^*. They recommend $r_k = b$ if $c < x_k^*$ and $r_k = a$ if $c \geqslant x_k^*$.*

The verdict is unanimous in favor of b when $c < x_1^*$ and unanimous in favor of a when $c > x_2^*$. The vote is split when $c \in (x_1^*, x_2^*)$ and the default option o is implemented. In that case, jurors polarize.

If type-1 jurors are a priori more in favor of convicting, their threshold will be lower than that of type-2 jurors. As a consequence, there exists a region of mixed evidence (x_1, x_2) where their best choice would be to take action a but action o will be taken instead. Given the best action is always either action a or action b, the optimal thresholds need only to discriminate between those two actions and they coincide with the thresholds obtained in the previous section.

The fact that thresholds differ implies that jurors will polarize in case of disagreement. This result is consistent with evidence from various experimental studies. It is known that individuals who exhibit confirmatory biases may interpret the same information in opposite ways. This *polarization effect* occurs when mixed evidence is given to subjects whose existing views lie on both sides of the evidence. Their beliefs may then move farther apart. In an early work, Lord et al. (1979) presented a set of typical arguments for and against the death penalty to a pool of subjects. When asked about the merits of death penalty, people who were initially in favor of (respectively against) capital punishment were more in favor (respectively against) after reading the studies (see also Walker and Main (1973), Darley and Gross (1983) and Plous (1991)). The literature explains this effect in terms of cognitive biases and non-Bayesian information processing. In particular, it has been argued that individuals focus attention on the elements that support their original beliefs and (consciously or unconsciously) neglect the elements that contradict them. Some researchers attribute polarization to heterogenous prior beliefs (e.g. Dixit and Weibull (2007)), or to non-Bayesian updating (e.g. Rabin and Schrag (1999)). Our analysis suggests that attentional deficits, multiple priors or biased information processing need not be at the origin of this result. Instead, our threshold mechanism can fully account for this behavior. Indeed, even if $p_1 = p_2$, jurors will polarize as long as they have different preferences (hence, different thresholds) and the evidence received is mixed. Moreover, note that in the threshold mechanism, the information that is retained is still processed in a (constrained) Bayesian way. Only, some information is filtered out. The result is illustrated in Figure 4.

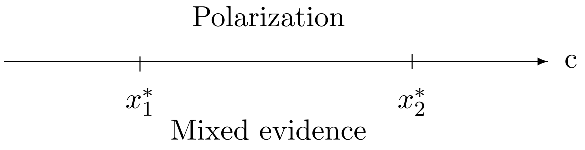

Figure 4: Optimal thresholds in the jury model

Interestingly, the polarization effect depends crucially on the intensity of preferences. Indeed, suppose that the preferences of both types of jurors become stronger in opposite directions. For example, both G_A^1 and G_B^2 increase. Type-1 and type-2 jurors have now higher incentives to recommend actions a and b, respectively. Therefore, type-1 jurors optimally decrease their threshold x_1^* whereas type-2 jurors optimally increase their threshold x_2^*, resulting in a increased gap between x_1^* and x_2^*. However, it is precisely when evidence falls in this region (what we called mixed evidence) that polarization occurs. Overall, stronger differences in preferences will result in higher levels of polarization following the release of evidence. This, again, could never occur in the traditional Bayesian framework.

4.2 Decision rules and biases in beliefs [xxx OUT? xxx]

In this section, we study decision-making under majority voting. To simplify the analysis, we assume that different types of jurors have identical preferences but different beliefs: $G_A^1 = G_B^1 = G$, $G_B^1 = G_B^2 = G$ and $p_1 \neq p_2$. To avoid cases of indifference, we assume that $\nu \neq 1/2$. The judge follows the recommendation of the majority (convict or acquit). In this simple model, a juror prefers b to a if his posterior belief is below 1/2 and a to b if his posterior belief is above 1/2. Type-k jurors set a threshold x_k and recommend a if $c \geqslant x_k$ and b if $c < x_k$.

Proposition 5 *The interpretation of the evidence depends on the rule.*

Assume $\nu < 1/2$. Then, type-2 jurors always obtain their preferred choice. The problem is equivalent to the case in which type-2 jurors are delegated the decision. Their threshold is therefore x_2^*. When type-1 jurors disagree with type-2 jurors, they are outnumbered. The degree of disagreement is also irrelevant to them as it does not affect the decision. If type-1 jurors are a priori relatively more in favor of action b, i.e. $x_1^* < x_2^*$, then any threshold $\check{x}_1 \leqslant x_2^*$ is an equilibrium. Overall, type-1 jurors do not have any clear strategy as their recommendation will never be followed and we may observe either little or large polarization.

This example illustrates the effect of the rule on the likelihood of polarization in juries. The rule affects the payoffs obtained in case of disagreement, and the optimal cutoff internalizes this externality. Therefore, jurors will process the information as a function of the rule, end up recommending different verdicts and hold different beliefs.

4.3 Jury selection

The previous results demonstrate that the interpretation of the evidence produced in front of a jury varies as a function of the prior beliefs and the intensity of the preferences of the jurors. As such, jury selection affects the overall outcome of a trial. Some experts believe that a large share of cases litigated are won or lost in the jury selection phase (see Farhinger (1993-1994)). During this process, attorneys will select jurors and they compete to secure opposite verdicts.

Informally, suppose there are two attorneys α and β. Attorney j's preferences can be represented by a utility function over decisions $H_j(\cdot)$ that does not depend on the state realized. Attorney's α utility is such that $H_\alpha(a) > H_\alpha(b) = 0$ while attorney β's utility is such that $H_\beta(b) > H_\beta(a) = 0$ independently of the state. These utilities induce an indirect preference over juries. For instance, the defense attorney (here β) would like to secure a release (action b) and is willing to avoid jurors likely to interpret the evidence in favor of a conviction. His objective is to maximize the number of jurors with high thresholds and to minimize those with low thresholds. The prosecutor has the opposite preferences and incentives.

Consider now the case with a default option, and let us assume that $H_\alpha(o) = H_\beta(o) = 0$ to reflect the fact that attorneys care exclusively about winning. From the perspective of α for example, it is optimal to eliminate jurors with a strong bias against A and to keep jurors with a strong bias in favor of A. As attorneys have opposite incentives, they will eliminate the jurors with the strongest views on both sides of the spectrum. Overall, the adversarial jury selection process will keep jurors with moderate views. As a consequence and compared to the initial draw of jurors, the jurors selected in this way will (i) vote unanimously more often and (ii) polarize their opinion less often after the release of evidence.

5 Conclusion

In this article we have built on neuroscience evidence to model belief formation and analyze the behavior of judges and juries. We have shown that physiological constraints generate posterior beliefs with qualitatively different properties from Bayesian posterior beliefs. In particular, a decision-maker will tend to reinforce his prior beliefs and to hold posteriors shaped by his preferences over outcomes. Also, the well-known polarization effect is a direct consequence of the model and should be observed when evidence is mixed, as behavioral evidence suggests.

There are a series of implications for decisions rendered by judges and juries. We have shown that cases analyzed in the early career of a judge may affect future decisions on cases that are a priori independent. Also, early evidence produced in a trial may matter more than late evidence. For the case of juries, the distribution of preferences in a jury affects the way information is interpreted by individual jurors. Finally, we argue that the endogenous selection of jury members reduces the likelihood of polarization and split opinions.

This analysis could be extended in various ways. For instance, the results obtained in section 4 presuppose that all information is public knowledge. In the presence of uncertainty about the preferences and prior beliefs of other members in the jury, the thresholds should reflect the distribution of preferences and beliefs rather than the exact values. We have also restricted ourselves to situations in which decision-makers want to take the "correct" action in each state. This provides the most likely scenario for avoiding biases and therefore constitutes the most natural benchmark. However, the results obtained may not fit all the available data on the polarization effect for instance. If jurors have other types of preferences, or the rule in case of disagreement is different, the group interaction may have different effects on posterior beliefs. This is one interesting alley for future research.

Appendix

We offer a brief description of the mathematical model and a sketch of some important results. We refer the reader to Brocas and Carrillo (2012a) and Brocas and Carrillo (2012b) for additional details and long proofs.

In Brocas and Carrillo (2012a), information is modeled as follows. When the state is S, the likelihood of c is $f(c|S)$ with $F(c|S) = \int_0^c f(y|S)dy$ representing the probability of a cell firing activity below c. A high cell firing is more likely to occur when $S = A$ and a low cell-firing is more likely to occur when $S = B$, which is captured by the Monotone Likelihood Ratio Property $\frac{\partial}{\partial c}\left(\frac{f(c|B)}{f(c|A)}\right) < 0$ for all c.

Efficient decision-making.

In the standard Bayesian framework, information is encoded and interpreted to its full extent. Namely, c represents the correct intensity of the signal and the decision is based on the realization of c. Consider the decision of individual i who holds a posterior belief μ_i that the state is A after receiving evidence . Given the posterior, i's expected payoffs of taking action $\gamma_i = a$ and $\gamma_i = b$ respectively are $V_i(a;\mu_i) = \mu_i G_A^i$ and $V_i(b;\mu_i) = (1-\mu_i)G_B^i$. The optimal decision is therefore:

$$\gamma_i^* = \begin{cases} a & \text{if } \mu_i \geqslant \mu_i^* \equiv \dfrac{G_B^i}{G_A^i + G_A^i} \\ \\ b & \text{if } \mu < \mu_i^* \end{cases} \tag{1}$$

Suppose the i is exposed to a signal of intensity c. Conditional on interpreting the signal at its exact value, i's posterior belief is therefore

$$\pi_i(c) = \frac{f(c|A)p_i}{f(c|A)p_i + f(c|B)(1-p_i)} \tag{2}$$

The higher c and the higher the posterior, and therefore there exists x_i^* such that $\pi_i(c) < \mu_i^*$ when $c < x_i^*$ and $\pi_i(c) \geqslant \mu_i^*$ when $c \geqslant x_i^*$. Formally x_i^* satisfies:

$$\frac{f(x_i^*|B)}{f(x_i^*|A)} = \frac{p_i}{1-p_i} \cdot \frac{G_A^i}{G_B^i} \tag{3}$$

Decision-making in the threshold mechanism

Given this mechanism, i can hold either of two posterior beliefs, $p(A|c \geqslant x) \equiv \overline{p}_i(x)$ or $p(A|c < x) \equiv \underline{p}_i(x)$, depending on whether $c \geqslant x$ or $c < x$. Formally,

$$\overline{p}_i(x) = \frac{[1-F(x|A)]p_i}{[1-F(x|A)]p_i + [1-F(x|B)](1-p_i)}; \qquad \underline{p}_i(x) = \frac{F(x|A)p_i}{F(x|A)p_i + F(x|B)(1-p_i)}$$

The expected payoff associated with the threshold mechanism is therefore:

$$W(x) = Pr(c \geqslant x)V_i(a; \overline{p}_i(x)) + Pr(c < x)V_i(b; \underline{p}_i(x))$$

$$W(x) = p_i G_A^i (1 - F(x|A)) + (1 - p_i) G_B^i F(x|B)$$

The optimal threshold solves $W'(x) = 0$ and satisfies

$$\frac{f(x_i^*|B)}{f(x_i^*|A)} = \frac{p_i}{1 - p_i} \frac{G_A^i}{G_B^i} \tag{4}$$

It is easy to see that actions a and b are taken under the same circumstances. However, the posterior beliefs are different. This establishes Result 1 The optimal cutoff is a function of p_i and $\frac{G_A^i}{G_B^i}$. It is sufficient to differentiate the first order condition with respect to p_i and $\frac{G_A^i}{G_B^i}$ to see that x_i^* is decreasing in the prior belief p_i and in the relative payoff $G = \frac{G_A^i}{G_B^i}$ (proposition 1). We obtain Proposition 2 by direct inspection of $\overline{p}_i(x_i^*)$ and $\underline{p}_i(x_i^*)$ given equation (4). Now, given some information is filtered out, the prior belief used for a subsequent episode of information interpretation is a biased posterior. Given this, decision-making in subsequent periods will not be efficient from a purely Bayesian perspective. Furthermore, the evidence used to compute the biased posterior will be reinforced in a subsequent period and the order in which evidence is produced matter. The details of the argument and proof of Proposition 3 can be found in Brocas and Carrillo (2012a), Proposition 3.

Juries in the threshold model

Fix type-2 jurors' threshold x_2. The expected payoff of type 1 choosing $x_1 < x_2$ is

$$\begin{aligned} W_1(x_1, x_2) &= Pr(c < x_1)(1 - \underline{p}_1(x_1))G_B^1 + Pr(c \in (x_1, x_2))G + Pr(c > x_2)\overline{p}_1(x_2)G_A^1 \\ &= F(x_1|B)(1 - p_1)G_B^1 + p_1\Big[F(x_2|A) - F(x_1|A)\Big]G \\ &\quad +(1 - p_1)\Big[F(x_2|B) - F(x_1|B)\Big]G + [1 - F(x_2|A)]p_1 G_1^A \end{aligned}$$

The optimal threshold for type-1 jurors satisfies the first order condition $\frac{\partial W_1}{\partial x_1} = 0$. The first order condition writes as:

$$f(x_1|B)(1 - p_1)[G_B^1 - G] - p_1 f(x_1|A)G = 0$$

The solution to this equation is x_1^*. Similar calculations apply for type-2 optimal threshold and we find it is x_2^*. These arguments prove Proposition 4. A more detailed argument can be found in Brocas and Carrillo (2012b).

Majoritarian rules.

Type k agents set a threshold x_k and recommends a if $c \geqslant x_k$ and b if $c < x_k$. If $\nu < 1/2$, type 2 are delegated the decision and set threshold x_2^*. Let us assume that $x_1 < x_2^*$ and let $p(x_1, x_2^*)$ be the posterior belief that the true state is A conditional on the evidence being in (x_1, x_2^*). The expected payoff of a type-1 juror is

$$\begin{aligned}\check{W}_1(x_1, x_2^*) &= Pr(c > x_2)G\overline{p}_1(x_2^*) + Pr(c \in (x_1, x_2^*))G(1 - p(x_1, x_2^*)) + Pr(c < x_1)G(1 - \underline{p}_1(x_1)) \\ &= (1 - p_1)F(x_2^*|B)G + p_1(1 - F(x_2^*|A))G\end{aligned}$$

and therefore the solution is any $\check{x}_1 \in (0, x_2^*)$. This proves Proposition 5.

References

[1] Austen-Smith D. and J.S. Banks (1996), “Information aggregation, rationality and the condorcet jury theorem”, American political Science Review 90(1): 34-45.

[2] Brocas, I. (2012), “Information Processing and Decision-making: Evidence from the Brain Sciences and Implications for Economics”, *mimeo (USC)* available at http://www.neuroeconomictheory.org/PDF/infoproc-survey.pdf

[3] Brocas I. and J.D. Carrillo (2012a), “From perception to action: an economic model of brain processes”, *Games and Economic Behavior*, 75(1), pp.81-103.

[4] Brocas I. and J.D. Carrillo (2012b), “The neuroeconomics of strategic decision-making”, *mimeo, USC.*

[5] Darley J. and P. Gross (1983),“A Hypothesis-confirming Bias in Labeling Effects”, *Journal of Personality and Social Psychology*, XLIV, 20-33.

[6] Ditterich, J., Mazurek, M., and M. Shadlen (2003), “Microstimulation of Visual Cortex Affects the Speed of Perceptual Decisions”, *Nature Neuroscience*, 6(8), 891-898.

[7] Dixit A.K. and J.W. Weibull (2007), “Political Polarization” *Proceedings of the National Academy of Science* 104(18), 7351-7356.

[8] Fahringer, H.P. (1993-1994), Mirror, Mirror on the Wall ...: Body Language, Intuition, and the Art of Jury Selection, 17, Am. J. Trial Advoc., pp. 197

[9] Feddersen T.J. and W. Pesendorfer (1996), “The swing voter’s curse.” American Economic Review 86(3), 408-424.

[10] Feddersen T.J. and W. Pesendorfer (1998), “Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting”, *The American Political Science Review*, 92(1), 23-35.

[11] Gold, J.I and M.N. Shadlen (2001), “Neural Computations that Underlie Decisions about Sensory Stimuli”, *Trends in Cognitive Sciences*, 5(1), 10-16.

[12] Hanes, D.P. and J.D. Schall (1996), “Neural Control of Voluntary Movement Initiation”, *Science*, 247, 427-430.

[13] Lord C.G, Ross L., and Lepper M.R. (1979), “Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence” Journal of Personality and Social Psychology, 37(11), 2098-2109.

[14] Myerson R. (1998) "Extended Poisson Games and the Condorcet Jury Theorem," Games and Economic Behavior, 25(1), 111-131.

[15] Platt, M.L. and P.W. Glimcher (1999), "Neural Correlates of Decision Variables in Parietal Cortex", *Nature*, 400, 233-238.

[16] Plous S. (1991), "Biases in the Assimilation of Technological Breakdowns: do Accidents Make us Safer", *Journal of Applied Psychology*, 21, 1058-1082.

[17] Rabin M. and J.L. Schrag (1999) "First impressions matter: a model of confirmatory bias", The Quarterly Journal of Economics, 114(1), 37-82.

[18] Shadlen, M.N., Britten, K.H., Newsome, W.T., and J.A. Movshon (1996), "A Computational Analysis of the Relationship between Neuronal and Behavioral Responses to Visual Motion", *Journal of Neuroscience*, 16, 1486-1510.

[19] Walker, T. G. and Main E. C. (1973), "Choice shifts and extreme behavior: Judicial review in the federal courts", *The Journal of Social Psychology*, 91(2): 215221.