

# When Should Political Scientists Use the Self-Confirming Equilibrium Concept? Benefits, Costs, and an Application to Jury Theorems

**Arthur Lupia**

*Department of Political Science, University of Michigan, 4252 Institute for Social Research,  
426 Thompson Street, Ann Arbor, MI 48104-2321  
e-mail: lupia@umich.edu (corresponding author)*

**Adam Seth Levine**

*Department of Political Science, University of Michigan, 4252 Institute for Social Research,  
426 Thompson Street, Ann Arbor, MI 48104-2321  
e-mail: adamseth@umich.edu*

**Natasha Zharinova**

*Risk Advisory Services, ABN AMRO Bank N.V., Gustav Mahlerlaan 10,  
PO Box 283, 1000 EA Amsterdam, The Netherlands  
e-mail: natalia.zharinova@nl.abnamro.com*

Many claims about political behavior are based on implicit assumptions about how people think. One such assumption, that political actors use identical conjectures when assessing others' strategies, is nested within applications of widely used game-theoretic equilibrium concepts. When empirical findings call this assumption into question, the self-confirming equilibrium (SCE) concept provides an alternate criterion for theoretical claims. We examine applications of SCE to political science. Our main example focuses on the claim of Feddersen and Pesendorfer that unanimity rule can lead juries to convict innocent defendants (1998. Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political Science Review* 92:23–35). We show that the claim depends on the assumption that jurors have identical beliefs about one another's types and identical conjectures about one another's strategies. When jurors' beliefs and conjectures vary in ways documented by empirical jury research, fewer false convictions can occur in equilibrium. The SCE concept can confer inferential advantages when actors have different beliefs and conjectures about one another.

## 1 Introduction

In game-theoretic studies of politics, the choice of an equilibrium concept can be equivalent to making assumptions about how people think. Many theorists adopt the Nash equilibrium (NE) concept that, when applied to numerous games, entails the assumption that all players think in a very similar manner when assessing one another's strategies (see, e.g., Turner 2000, 2001). In a NE, all players in a game base their strategies not only on

---

*Authors' note:* We thank Timothy J. Feddersen, Drew Fudenberg, Alexander Von Hagen-Jamar, Mika Lavaque-Manty, Justin Magouirk, William McMillan, Marco Novarese, Scott E. Page, Thomas Palfrey, Alexandra Shankster, Dustin Tingley, and Barry R. Weingast for helpful comments.

© The Author 2009. Published by Oxford University Press on behalf of the Society for Political Methodology. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

knowledge of the game's structure but also on *identical conjectures about what all other players will do* (Aumann and Brandenberger 1995).

The NE criterion pertains to whether each player is choosing a strategy that is a best response to a shared conjecture about the strategies of all players. A set of strategies satisfies the criterion when all player strategies are best responses to the shared conjecture. In many widely used refinements of the NE concept, such as subgame perfection and perfect Bayesian, the inferential criteria also require players to have shared, or at least very similar, conjectures.

How many political actors think about one another in such ways? Clearly, some do not. Citizens who have little interest in politics, such as many people who are called upon to act as voters or jurors, do not appear to base decisions on identical (or even similar) conjectures. Religious conservatives and humanist liberals, and rich and poor, are among pairs of politically relevant groups who think about important aspects of social life in very different ways. As a result, it is plausible that diverse citizens can base political decisions on very different conjectures about one another.

How should these facts affect game-theoretic political science? It depends on the situation. We agree with those who argue that many people do not literally engage in the kind of reasoning that common equilibrium concepts presuppose (see, e.g., Rubinstein 1998). We also agree with those who claim that some political actors make decisions "as if" such reasoning occurs. We agree, for example, with Satz and Ferejohn (1994) who argue that institutions can structure choices in a way that gives people an incentive to think about their options in ways that are consistent with Nash-based assertions.

So it is possible that many citizens in the contexts that political scientists study reason as if they have identical, or at least very similar, conjectures. But what about those who do not? For them, "as if" claims are hard to justify. Consider, for example, jury decision making—a topic to which theoretical political scientists have paid much attention (see, e.g., Feddersen and Pesendorfer 1998, henceforth FP). Should jurors be modeled as having identical conjectures about one another's strategies?

Jurors come to courtrooms with widely differing worldviews. Many have little or no experience in legal settings. Many jurors receive little or no feedback on the quality of their decisions and have little motivation to think about any feedback that they might receive. Empirical research on juries shows that there are significant variations in how jurors think about one another. Such variations lead to important differences in how jurors describe their conjectures about the meaning of evidence, courtroom presentations, and jury room deliberations (see, e.g., Pennington and Hastie 1990). Similar questions can be asked about the shared conjectures of other political actors such as voters (who pay varying amounts of attention to politics and can have very different conjectures about social cause and effect) or diverse peoples who are asked to contribute to novel public goods despite not having interacted with sufficient frequency to share behavioral expectations.

Given the frequency with which political scientists encounter actors who share decision contexts despite having diverse worldviews and experiences, it is reasonable to question whether commonly used equilibrium concepts provide the most effective means for characterizing all kinds of political behavior. A generation of theorists have recognized such challenges and taken steps to meet them. Some, like Harsanyi (1967, 1968) and Kreps and Wilson (1982), have refined the Nash concept to allow players to choose best responses to the strategies of others even though they lack information about specific aspects of the game. Others, such as Aumann (1974) and McKelvey and Palfrey (1995), have diverged farther from the basic Nash concept (Nash 1950).

We argue that political science should consider the benefits and costs of turning some of its theoretical energies to alternate approaches. One such approach entails using the

self-confirming equilibrium (SCE) concept (Fudenberg and Levine 1993, 1998; Dekel, Fudenberg, and Levine 1999; Dekel, Fudenberg, and Levine 2004, henceforth DFL 2004). The key element of a SCE is the correspondence between what a player does and what she observes. If her observations are consistent with her conjectures about other players' strategies, then her rationale for her actions is positively reinforced. If *all* players receive such reinforcement, then their actions are "in equilibrium."

Like Nash-based equilibrium concepts, a SCE characterizes players who are goal oriented (in that they have utility functions) and strategic (in that they seek to maximize utility by basing plans of action on what they believe, conjecture, and observe). Unlike Nash-based concepts, SCE does not require that players know much else. A player can be wrong about important features of the game, including what other players are doing, and yet her strategy can remain "in equilibrium" if what she observes about the game is consistent with her conjectures about it. The benefit to political scientists of using the SCE is that it can provide a rigorous platform for deriving theoretical claims in situations where political actors need not have similar conjectures about one another's strategies. Since the SCE allows us to build a wider range of assumptions about how people reason into our models, it can also expand our abilities to integrate a greater range of empirically supported psychological insights into game-theoretic political science.

There are also costs to using the SCE concept. The main cost is that the SCE concept can generate more equilibria than do more commonly used equilibrium concepts. For many scholars, this fact provides sufficient rationale for ignoring the SCE. But when the "as if" assumption is empirically implausible, discarding the SCE implies a preference for inferences based on untenable assumptions over inferences from a more plausible empirical basis. When should we sacrifice the plausibility of the assumptions for a reduced number of equilibria? The goal of this paper is to support the proposition that this question is at least worth debating—particularly in circumstances where evidence documents political actors thinking very differently about critical elements of their decision contexts.

To support this goal, we proceed as follows. We begin by describing reasoning assumptions that are implicit in the application of common equilibrium concepts. Then, we present the SCE concept. In the process, we offer examples where basing inferences on the SCE concept leads to different, but constructive, insights about important political questions. In each of our examples, the findings are more than a technical curiosity—they come from attempts to reconcile a formal model with empirically defensible assumptions about how political actors think.

In our main example, we use the SCE concept to cultivate a link between psychological and game-theoretic studies of jury decision making. We reexamine the jury model of FP in light of psychological research on how jurors process trial information (e.g., Pennington and Hastie 1993) and on variations in how rigorously people think (Cacioppo and Petty 1982). FP claim that the likelihood that a unanimous jury verdict convicts an innocent defendant is increasing in jury size. Using SCE to characterize behavior and outcomes in a variant of the original model, we show that their claim *depends on the assumption that all jurors have identical conjectures about one another's strategies*. We show that allowing juror conjectures to vary in empirically documented ways is sufficient to reduce the number of false convictions in equilibrium.

Our examples support the proposition that the credibility of game-theoretic political science need not rest on the sometimes-untenable assumptions about human reasoning that are embedded in important applications of common equilibrium concepts. Where evidence shows that all political actors do not share conjectures about one another's strategies, using the SCE allows scholars to derive theoretical conclusions from premises that are easier to defend empirically.

## 2 Ways of Thinking in Game-Theoretic Equilibrium Concepts

For many people, game theory and the NE concept are synonymous. Given the frequency with which the concept is used in game-theoretic political science, the perceived synonymy is understandable. NE, however, is just one of several often-used equilibrium concepts. Although many noncooperative game-theoretic studies in political science do not use NE, almost all use refinements of the Nash concept. Common refinements include the subgame perfect, trembling-hand perfect, Bayesian Nash, perfect Bayesian, and sequential equilibrium concepts. Subgame perfection, for example, is an NE refinement that strengthens the inferential power of game-theoretic treatments in extensive form games—where strengthening implies introducing an additional technical criterion that is appropriate for that class of games. The other attribute of these refinements is that they retain core properties of the original NE concept—in particular, its requirement that player strategies constitute best responses to the strategies of all other players—with the response evaluated along the equilibrium path in games containing sequences of moves.<sup>1</sup>

Many people treat Nash-based concepts as substantively innocuous—as entailing no substantive baggage. This is wrong. Each of these concepts presumes that players reason in a specific manner. To see how, consider Gibbons' (1992: 8–9) definition of a NE, where  $S_i$  denotes the set of possible strategies for player  $i$ ,  $s_i$  denotes an element of that set, and  $u_i(s_1, \dots, s_n)$  denotes player  $i$ 's utility function and refers to the fact that her utility can be a function of other players' strategies as well as her own.

In the  $n$ -player normal-form game  $G = \{S_1, \dots, S_n; u_1, \dots, u_n\}$ , the strategies  $(s_1^*, \dots, s_n^*)$  are a Nash equilibrium if, for each player  $i$ ,  $s_i^*$  is (at least tied for) player  $i$ 's best response to the strategies specified for the  $n - 1$  other players,  $(s_1^*, \dots, s_{i-1}^*, s_{i+1}^*, \dots, s_n^*)$ :  $u_i(s_1^*, \dots, s_{i-1}^*, s_i^*, s_{i+1}^*, \dots, s_n^*) \geq u_i(s_1^*, \dots, s_{i-1}^*, s_i, s_{i+1}^*, \dots, s_n^*)$  for every feasible strategy  $s_i$  in  $S_i$ ; that is,  $s_i^*$  solves  $\max_{s_i \in S_i} u_i(s_1^*, \dots, s_{i-1}^*, s_i, s_{i+1}^*, \dots, s_n^*)$ .

This definition requires *shared conjectures*. As Aumann and Brandenberger (1995: 1163, underline added) describe,

In an  $n$ -player game, suppose that the players have a common prior, that their payoff functions and their rationality are mutually known, and that their conjectures [about the strategies of others] are commonly known. Then for each player  $j$ , all the other players  $i$  agree on the same conjecture  $\sigma_j$  about  $j$ ; and the resulting profile  $(\sigma_1, \dots, \sigma_n)$  of mixed actions is a Nash equilibrium.

Common Nash refinements have similar attributes. Although these refinements differ in what they allow players to know and believe, they continue to require that actors share identical conjectures of other players' strategies (or the actions of specific types of other players) along the equilibrium path.

It is reasonable to ask how many citizens base political decisions on universally shared conjectures. Reasoning requires time, effort, and at least a modicum of cognitive energy. Even for motivated people, information processing is characterized by severe constraints (see, e.g., Kandel, Schwartz, and Jessell 1995: 651–66). Chief among these constraints are the very limited storage capacity and high decay rates of working memory as well as the restrictive rules by which stimuli gain access to long-term memory.<sup>2</sup> One implication of

<sup>1</sup>For simplicity, we use the term “equilibrium path” to characterize paths of any length (including zero), which allows us to use a single term to cover equilibria in all normal and extensive form games.

<sup>2</sup>Bjork and Bjork (1996) and Schacter (1996, 2001) provide entry-level references for properties of memory and their implications for social interaction.

these attributes is that citizens are likely to pay attention to different stimuli and remember different events, which can create and reinforce diverse internal theories of cause and effect and, ultimately, lead people to develop divergent conjectures about what others would do under certain circumstances.

To be sure, some political actors process information in ways that yield identical conjectures about what everyone else is doing. Just as surely, others do not. Fudenberg and Levine (1993, 1998), Fudenberg and Kreps (1995), and Dekel, Fudenberg, and Levine (1999) developed the SCE concept for game-theoretic analyses of the latter case. To date, this concept has had limited application in political science. In the remainder of this section, we offer a brief primer on the concept and then examine benefits and costs of its use to political scientists.

The primer is as follows. Our main reference for it is DFL. Let  $i$  be a player in the game, and let  $I$  be the set of such players. Following DFL, we assume that all parameters of the game, including the number of players, their possible actions, and their types, are finite. Let  $\theta_i \in \Theta_i$  be player  $i$ 's type, and let  $\theta_{-i}$  denote the vector of other players' types. Let  $a_i \in A_i$  denote player  $i$ 's action, and let  $\sigma_i(a_i) \in \Delta(A_i)$ , henceforth  $\sigma_i$ , denote a mixed strategy for player  $i$  in the set of possible actions for her.

The attributes of a game that are assumed to be “common knowledge” are an important difference between the DFL setup and more familiar Nash-based approaches. In many games, even those featuring incomplete information, nearly all attributes of the game are assumed to be common knowledge. In the DFL setup, the common knowledge can be quite limited. Although the common knowledge includes players knowing their own utility functions, it need not include much more. It need not include the full set of strategies available to other players. It need not include knowledge of the distributions from which player types are drawn. As a result, players can have different beliefs about the kind of game they are playing, what actions are available to which players, and they can assign different prior probabilities over the set of types. Players need not even be aware that others have different views of such matters.<sup>3</sup>

Stated mathematically, let  $\mu_i(\theta_i)$  be player  $i$ 's prior belief about her own type and let  $\mu_i(\theta_{-i}|\theta_i)$  be player  $i$ 's beliefs about other player's types, given her own type. Let  $r$  be the true distribution from which player types are drawn, where  $r(\theta_i)$  denotes the true distribution from which player  $i$ 's types are drawn and  $r(\theta_{-i})$  denotes true distributions from which player types other than  $i$ 's are drawn. When  $\mu_i(\theta_i) = r(\theta_i)$ , we say that player  $i$  has correct beliefs about his own type, and when  $\mu_i(\theta_{-i}) = r(\theta_{-i})$ , we say that player  $i$  has correct beliefs about the types of all other players. When  $\forall i, j \in I, \mu_i = \mu_j$ , we say that players have common prior beliefs. With these definitions in hand, it is important to note that in what follows, we need not always assume common or correct prior beliefs.

DFL's setup represents everything else that players know about Nature and their opponents by “private signals.” Let  $y_i = y_i(a, \theta)$  be player  $i$ 's private signal about the play of the game. This signal is what player  $i$  observes in the game. This signal can include any or all

<sup>3</sup>This representation of common knowledge distinguishes the SCE concept from other generalizations of the NE idea such as rationalizability (Pearce 1984; Bernheim 1984). Rationalizability makes strict assumptions about what is common knowledge during the game. It includes each player's entire set of payoffs as well as the range of counterfactuals that other players must be running (i.e., “their rationality”). SCE permits weaker assumptions about both these items. In this framework, the full range of others' payoffs may be unknown, and each player may not be running complete counterfactuals about what all players would do under all possible information sets.

the following: which terminal node is reached, information about other players' moves, and payoffs. It may also include none of the above—an assumption we can make if we want to model a situation where a player either receives no feedback about a game or is unable to pay attention to available feedback.

The term “private signal” when used in an SCE context is *not* equivalent to the term “private information” that often describes game attributes that are known to one player but not another. Although the information contained in a private signal can be private information, it need not be. In other words, in most games with private information, it is common knowledge that private information exists and that the content of the private information is the result of a draw from a common knowledge distribution. Here, by contrast, such knowledge need not be common. In sum, each player observes her own action  $a_i$ , type  $\theta_i$ , and private signal  $y_i(a, \theta)$ .

Let  $\hat{\sigma}_{-i} \in \times_{-i} \Delta(\sigma_{-i})$  be player  $i$ 's conjecture about his opponents' play (specifically, his conjecture about the strategy profile of his opponents), and let  $u_i(a_i, \theta)$  be player  $i$ 's expected utility from playing  $a_i$ . We now have sufficient definitions and notation to present DFL's (p. 286) definition of an SCE.<sup>4</sup>

*Definition:* A strategy profile  $\sigma$  is a SCE with conjectures  $\hat{\sigma}_{-i}$  and beliefs  $\hat{\mu}_i$  if for each player  $i$ , (1)  $\forall \theta_i, r(\theta_i) = \hat{\mu}_i(\theta_i)$ , and for any pair  $\theta_i, \hat{a}_i$  such that  $\hat{\mu}_i(\theta_i) \cdot \sigma_i(\hat{a}_i | \theta_i) > 0$  both the following conditions are satisfied, (2)  $\hat{a}_i \in \arg \max_{a_i} \sum_{\theta_{-i}} \sum_{a_{-i}} u_i(\hat{a}_i, a_{-i}, \theta_i, \theta_{-i}) \hat{\mu}_i(\theta_{-i} | \theta_i) \hat{\sigma}_{-i}(a_{-i} | \theta_{-i})$ , and (3) for every  $\bar{y}_i$  in the range of  $y_i$ :

$$\begin{aligned} & \sum_{\{a_{-i}, \theta_{-i}; y_i(\hat{a}_i, a_{-i}, \theta_i, \theta_{-i}) = \bar{y}_i\}} \hat{\mu}_i(\theta_{-i} | \theta_i) \hat{\sigma}_{-i}(a_{-i} | \theta_{-i}) \\ & = \sum_{\{a_{-i}, \theta_{-i}; y_i(\hat{a}_i, a_{-i}, \theta_i, \theta_{-i}) = \bar{y}_i\}} r(\theta_{-i} | \theta_i) \sigma_{-i}(a_{-i} | \theta_{-i}). \end{aligned}$$

In words, a SCE has three requirements. Condition 1 states that each player has correct beliefs about her own type. Condition 2 states that any action that a player plays with positive probability must maximize her utility, given her beliefs about Nature and her conjectures about other players' strategies. Condition 3 (hereafter *C3*) describes allowable player conjectures in equilibrium. *C3* is the key difference between SCE and common Nash refinements.

Although Condition 2 requires that each player's strategy be a best response to the player's beliefs about Nature and conjectures about opponents' play, *C3* requires that these beliefs and conjectures be consistent only with what the player herself observes. When a player's observations, beliefs, and conjectures are in synch, what she sees confirms her choice and gives her no reason to change. When the same is true for all players, then the strategy profile is in equilibrium.

In a SCE, each player's strategy is a best response to her own beliefs, conjectures, and observations (if any) and not necessarily to the actual strategies of other players. To satisfy *C3*, it is sufficient that player beliefs, conjectures, and observations are consistent. How they become consistent—whether through conjectures that are shared, unshared, simple, or complex—is irrelevant.

Two additional characteristics about SCE are important to note. First, there exist NE that are not SCE (i.e., SCE is not a NE refinement, DFL: 290–3). Second, the SCE concept does

<sup>4</sup>We restrict attention to what DFL (p. 287) call SCE with independent beliefs, which implies that player  $i$ 's beliefs about her opponents' types do not depend on her own type. This independence restriction parallels an assumption made in nearly all games of incomplete information in political science.

not require that players use Bayes's rule to process information. It requires only that actors' beliefs and conjectures, however drawn, are consistent with their observations.<sup>5</sup>

### 3 Benefits and Costs of SCE

For political science, the *SCE* concept has four critical properties: observations must be consistent with beliefs and conjectures, incorrect conjectures are allowed, two players can disagree about a third (or Nature), and more precise observations by players imply greater constraints on what conjectures constitute a SCE. We address the substantive implications of each property in turn.

#### 3.1 *The Relationship between Observations, Beliefs, and Conjectures*

[E]ach player attempts to maximize his own expected utility. How he should go about doing this depends on how he thinks his opponents are playing, and the major issue . . . is how he should form those expectations (Fudenberg and Levine 1998: 14).

The SCE requires that players' expectations are formed by their beliefs and conjectures and confirmed by their observations. A motivation for this move is as follows:

The most natural assumption in many . . . contexts is that agents observe the terminal nodes (outcomes) that are reached in their own plays of the game, but that agents do not observe the parts of their opponents' strategies that specify how the opponents would have played at information sets that were not reached in that play of the game . . . [I]n many settings players will not even observe the realized terminal node, as several different terminal nodes may be consistent with their observation (Fudenberg and Levine 1998: 175).

So unlike in Nash-based concepts, players in a SCE need not justify their strategies as best responses to the anticipated strategies of other players. In a SCE, players just need a theory of cause and effect that keeps them from making mistakes that they can recognize given what they see. If an actor's private signal provides imprecise feedback, or no feedback at all, then she may choose actions in equilibrium that she would view as suboptimal if her private signal were more informative. Nevertheless, if what she sees is consistent with what she believes and conjectures, she has no rationale for changing her strategy.

Of course, we can imagine cases where actors would be hesitant to base their conjectures on partially informative or uninformative private signals. If such actors had opportunities to improve their feedback, they would do so. Fair enough. But many actors that political scientists study lack the willingness or ability to gain such information. SCE can help theorists better represent such actors in formal models.

#### 3.2 *Incorrect Beliefs and Conjectures Are Allowed*

An important difference between the SCE concept and more common equilibrium concepts is that actors in a SCE can maintain incorrect beliefs and conjectures. In many

---

<sup>5</sup>Most noncooperative games of incomplete information use refinements of the NE concept (e.g., perfect Bayesian equilibrium, sequential equilibrium) that presuppose players' use of Bayes's rule to draw inferences. The SCE concept, by contrast, does not require that actors use Bayes's rule. It requires only that actors' beliefs and conjectures, however drawn, are consistent with their observations. In other words, when Bayesian updating is assumed, posterior beliefs are constrained to have a specific functional relationship to prior beliefs. In an SCE, things are different. To the extent that a player's private signal is generated by reality (i.e., the true distribution of Nature's and/or players' types), it is not correct to say that the SCE outcome must be independent of prior beliefs. However, in a SCE, the relationship between priors and posteriors can be far less direct than Bayes's rule posits.

common equilibrium concepts,  $\hat{\mu}_i = r$  and  $\hat{\sigma}^{-i} = \sigma_i$  ( $\forall i$ ). In particular, all players must share correct conjectures about the action that every single type of every single player would choose at every decision node along the equilibrium path. In a SCE, by contrast, variance in the quality of the private signal allows players to maintain incorrect beliefs and conjectures in equilibrium.

A maximizing strategy in a SCE can include actions that are suboptimal so long as the information conveyed by the private signal does not reveal the suboptimality. In other words, if a player does not expect to learn that her conjecture is untrue—and if she never receives the kind of feedback that would expose her to her conjecture’s error—then she has no reason to rethink her strategy. We can certainly imagine political actors who approach political decisions in such ways. If, for example, the evidence and feedback that a voter or juror receives are consistent with whatever simple rules-of-thumb she may be using (i.e., “vote Republican” or “always doubt the testimony of police officers”), why should she think any more about these matters? Her conjectures (which, unbeknownst to her, may be false) and her observations (which, unbeknownst to her, may be limited in their informative value) reinforce one another and that is sufficient to internally justify her strategy.

To some readers, such a statement may seem to be an anathema. Game theory, after all, is often linked with the idea of rationality. The maintenance of incorrect conjectures and potentially suboptimal strategies will strike some readers as anything but rational. To such reactions, one thing is worth pointing out. A problem with many claims about “rationality” is that there are numerous conflicting definitions of the term in circulation (see, e.g., the definitional inventory in Lupia, McCubbins, and Popkin 2000: 3–11). Among the least useful of these definitions for explaining the actions of flesh-and-blood human actors are definitions that equate rationality with omniscience. Alternative definitions hold rationality as the product of human reason, where reason is the ordinary function of the mind. Therefore, it is a reader’s positing of omniscience as a desirable analytic standard, rather than a search for properties of standard human reason, that makes an equilibrium featuring incorrect conjectures appear to be an oxymoron.

### 3.3 *Two Players Can Disagree about Attributes of a Third*

Unlike common Nash-based concepts, two players in a SCE can disagree about the actions or types of a third. For example, in a three-player game where a player’s shirt color affects player payoffs, Player 1 can believe that Player 3 is wearing a blue shirt, Player 2 can believe that Player 3 is wearing a yellow shirt, and as long as the observations of Players 1 and 2 are consistent with these beliefs (which means that private signals could not include Player 3’s shirt color), neither player has an incentive to change her actions or beliefs. To see why this factor matters, consider a simple example (adapted from Fudenberg and Levine 1998) that shows the impact of moving from Nash-based equilibrium concepts to SCE. In Fig. 1, Congress and the President are in a standoff over the budget. If the standoff persists, as it did in the mid 1990s, the government will shut down, which hurts many voters.

If Congress and the President end the standoff, then all players earn a payoff of 1. If either player continues the standoff, the government shuts down and the move goes to a representative voter. The voter, who observes a government shutdown, but not why it occurred, blames either the President or the Congress. The player who is not blamed benefits with a payoff of  $2 + e$ , where  $e > 0$  and can be very small.

The outcome  $(\sim s, \sim s)$  (i.e., Congress and the President agree to end the standoff) is an SCE when Congress conjectures that it is more likely than the president to be blamed for

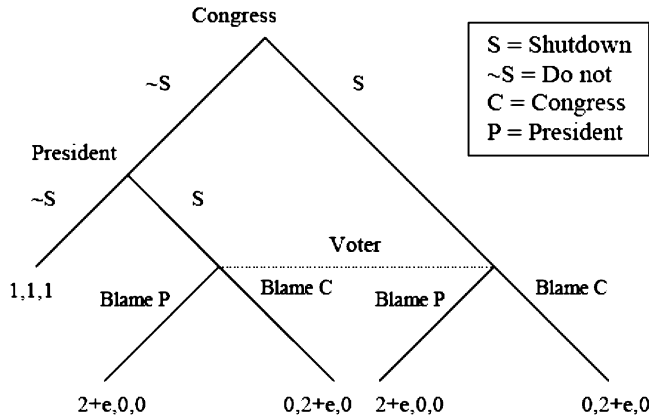


Fig. 1 Congress-president standoff.

the standoff, whereas the President conjectures that he/she is more likely than Congress to be blamed. Since the voter’s decision node is not reached in equilibrium, everything the Congress and the President observe is consistent with their conjectures (i.e., they never observe the other being blamed). Their choices of strategy are confirmed.

The outcome  $(\sim s, \sim s)$ , however, does not occur when the equilibrium concept requires Congress and the President to have identical conjectures about the voter’s move. To see why, note that the standoff continues if the voter blames either player with probability greater than or equal to  $1/2$ , which the voter must do since the two probabilities (the probability of blaming Congress and the probability of blaming the President) must sum to 1. Therefore, any mixed strategy by the voter would induce at least one of the other two players to continue the standoff.

It is worth noting that if  $e$  is sufficiently small, then producing the SCE described above requires only a small difference between presidential and congressional conjectures about voter behavior. Each entity could, for example, conjecture that the likelihood of it being blamed was 51%. This 2% difference is within the margin of error of even the best political polls and, as such, can be smaller than the difference in polls that each entity might commission in reality.

Not only can each player maintain different conjectures about a third in a SCE but also *different types of the same player* can maintain varying conjectures about Nature or other players. For politics, this aspect of the SCE permits greater flexibility in representing the mind-sets of different types of people who can inhabit the same player roles—such as that of a pivotal juror or voter. In such roles, we can imagine lifelong Democrats and lifelong Republicans basing their strategies on very different notions of political cause and effect (e.g., why George W. Bush pursued a war with Iraq) and/or different conjectures about how they themselves would act if they were members of the other party. SCE allows us to derive characterizations of players who share neither common prior beliefs nor identical ways of thinking about any information that they do have.

**3.4** *As the Precision and Range of Observation Increases, SCE and NE Converge*

The correspondence between a game’s NE and SCE depends on what players observe. In general, the more players observe, the closer is the correspondence. The theoretical implications of this correspondence become clear in extreme cases.

At one extreme, suppose that the private signal is completely informative. In this case, NE and SCE coincide. That is, when the play of the game reveals a player's own payoff and other players' beliefs, conjectures, and strategies, then utility maximization implies choosing a best response to other players' (observed) strategies. Moreover, when players' private signals fully reveal Nature's move in a game where at least one player (or Nature) has multiple types, then SCE and common Nash refinements such as Bayesian NE converge as well (i.e., each actor must maximize utility with respect to every player's type that they expect to encounter).

As the private signal becomes less informative, NE and SCE can diverge. At the extreme, when private signals are least informative, the set of SCE allows all profiles of ex ante undominated strategies (DFL: Proposition 1). In other words, players attempt to maximize utility but without any of the feedback we normally think of game-theoretic actors possessing.

### 3.5 *The Costs of SCE: Multiple Equilibria*

Having presented a basic definition of SCE and an overview of the analytic advantages it offers, we now turn to the topic of cost. For some observers, the main cost lies in the number of equilibria produced. The SCE concept will typically yield a larger set of equilibria than will equivalent models characterized using better known equilibrium concepts.

A multiplicity of equilibria is problematic for several reasons. First and foremost, when a researcher has a unique equilibrium, strong statements about cause and effect are easy to derive. When equilibria multiply, initial conditions can produce numerous, and sometimes contradictory, conclusions. Since a primary rationale for pursuing formal logic is to produce clear causal statements, multiplicity of equilibria is seen as problematic.

For these reasons, game theorists often view expansions of the set of equilibria as a bad thing and have spent significant time developing refinements that reduce the number of equilibria (see, e.g., Govindan and Wilson 2008). So it is reasonable to ask whether the extra equilibria that emerge from using the SCE concept merit scholarly attention.

The answer to this question depends on the value of deriving theoretical conclusions from empirically defensible premises. When the set of SCE and NE differ, the difference is the result of loosening the Nash-based concept's reasoning requirements. When empirical evidence demonstrates that the people whose behavior a model is constructed to explain *do not* reason as if they share critical beliefs or conjectures, the change in the set of equilibria caused by moving to SCE is a signal that the set of *Nash-based conclusions were artifacts of psychological assumptions that are difficult to defend*. In other words, the choice between a SCE approach and a Nash-based approach is akin to the choice between a process whose products are based on demonstrably incorrect assumptions about how people think and a process whose product is no less logical but may be more difficult to characterize. In our view, when "as if" assumptions are clearly false, ignoring the SCE or treating its additional equilibria as valueless clutter is akin to sacrificing the argument's soundness.

### 3.6 *The Costs of SCE: Myerson's Critique*

Our view does not imply that Nash-based equilibrium concepts are never appropriate. Far from it. There are many circumstances in which it is reasonable to model political actors as if they have shared conjectures. These circumstances include theoretical examinations of

professional legislators and other political elites who can reasonably be expected to have a large set of shared experiences and, hence, common expectations about one another and their environs. The same will be true of other decision makers who, through habit or custom, are in decision environments where common expectations can be expected to arise.

In this sense, our way of thinking about when SCE is most valuable follows Myerson (2006). Myerson (2006) critiques the use of the SCE concept of de Figueiredo, Rakove, and Weingast (2006) to explain British-American conflict in the Revolutionary War era. de Figueiredo, Rakove, and Weingast (2006) argue that fundamental differences in the beliefs and conjectures of the two sides led to radically different interpretations of key events (i.e., which game they were playing) that, in turn, led to conflict escalation. Myerson (2006: 427) counters that players in this game “are intelligent enough to understand anything that we game theorists can understand about their game” (i.e., historical experience allowed the British and Americans to know a lot about one another).

Myerson’s (2006) argument focuses on cases where political actors have the ability and motivation necessary to form shared conjectures about one another’s strategies. His argument does not apply when historical experience cannot be counted on to provide identical beliefs or common conjectures about critical decision-related phenomena. If, for example, we want to explain the actions of goal-oriented actors who are in unfamiliar surroundings, actors who receive little or no feedback about their actions, and those who may have limited opportunity or motivation to think about any feedback that they do receive, then a modeling approach that allows diverse beliefs about players’ types and conjectures about their strategies can be constructive. In our final example, we use the SCE concept for just this purpose.

#### 4 Thinking Differently in Jury Theorems

In this example, we briefly reexamine an important question about jury decision making. This topic has received great attention from game theorists in recent years. Psychologists have also studied it extensively. The psychological research reveals significant variations in how jurors think. But the theoretical and psychological literatures do not speak to one another. As a result, theoretical consequences of observed variations in how jurors think have not been explored. Our brief SCE-oriented example draws insights from both research traditions in an attempt to clarify these consequences.

##### 4.1 Background

The focus of current jury theorems begins with the Condorcet (1785) jury theorem (henceforth CJT). In it, a jury of  $n$  members chooses one of two alternatives, say  $A$  or  $C$  (i.e., acquit or convict). It is common knowledge that one of these alternatives corresponds to the true state of the world (innocence or guilt) and that everyone prefers the group to choose that alternative. But the true state of the world need not be known. The CJT shows that if the probability of each member choosing the “better alternative” is greater than .5, then the probability that a majority will also choose it goes to 1 as  $n \Rightarrow \infty$ . The result highlights beneficial information aggregation properties of common collective decision rules.

Austen-Smith and Banks (1996) showed that information aggregation need not be so beneficial. Their analysis begins with a question about whether individuals make the same choices when voting as a member of a jury as they do when voting alone. Austen-Smith and

Banks (1996) model each juror as receiving an evidentiary signal, say  $m_j \in \{G, I\}$ , that conveys information about the true state of the world (i.e., guilty or innocent).<sup>6</sup> Substantively, the signal represents a juror's view of trial evidence and deliberation. Technically, each juror's signal is determined by a single, independent draw from a Bernoulli distribution. Although it is assumed that each juror observes only their own signal, two things about the distribution are commonly known. First, the true state of the world is  $G$  with probability  $s \in (0, 1)$ —and is  $I$  with probability  $1 - s$ . Second, each signal conveys the true state of the world with probability  $p \in (.5, 1)$ —and the false state of the world with probability  $1 - p$ .

The work of Austen-Smith and Banks (1996) investigates whether all jurors in this circumstance would vote to convict when  $m_j = G$  and vote to acquit when  $m_j = I$ . If all jurors were to vote in accordance with their evidentiary signals, the CJT's beneficial information aggregation properties would survive. But Austen-Smith and Banks (1996) show that such behavior *need not* be a NE. Their finding comes from seeing a juror as being in one of two situations: "pivotal" or "not pivotal." If a juror is not pivotal, then her vote cannot affect the verdict and what she does with her information has no bearing on whether or not the group chooses the better alternative. By contrast, if the juror is pivotal and majority rule is being used, then the aggregate outcome is a tie without her vote. In this case, if everyone else is voting in accordance with their evidentiary signal, then it must be the case that the other jurors have observed  $G$ 's and  $I$ 's in equal amounts. Austen-Smith and Banks (1996) assume that jurors use this information *as well* when casting a vote. They prove that if a juror's prior beliefs about the true state of the world are sufficiently strong (i.e., if  $s$  is sufficiently close to zero or one) and if the juror uses Bayes's rule and hypothesizes what signals other jurors must have seen to make her vote pivotal, then the juror maximizes her expected utility by ignoring her own evidentiary signal. In other words, her best response to everyone else voting in accordance with their evidentiary signals is *not* to do the same. In equilibrium, the juror's vote is carried not by her observation of the trial evidence but by the weight of her beliefs and conjectures about what others must be thinking and doing if her vote is indeed the tiebreaker.

FP extend this logic to the case of unanimous verdicts. A common rationale for unanimity in juries is that it minimizes the probability of convicting the innocent. If jurors vote in accordance with their evidentiary signals, a kind of voting that FP call "informative voting," then unanimity minimizes the probability of false convictions. But FP identify a NE in which unanimity produces more false convictions than do other decision rules because jurors need not vote informatively.

In their model, a juror is not pivotal if at least one other juror is voting to acquit. Under unanimity rule, only one acquittal vote is sufficient for an acquittal. Hence, if a juror is pivotal under unanimity rule, then she can infer that *every other juror must be voting to convict*.

In other words, the juror can infer that either her vote makes no difference to the outcome or her vote is pivotal. If her vote is pivotal, she can make an inference about how many other jurors received guilty signals that, in turn, can change her beliefs about the likelihood of the defendant's guilt. The authors identify conditions in which the weight of each juror's conjecture about what other jurors are doing leads *all of them* to conclude that they should vote to convict—even if they all received innocent signals. False

<sup>6</sup>We use the term "evidentiary signal" to describe what the jury models call a "private signal" to avoid confusion with the SCE literature's long-standing, but distinct, use of the same term.

convictions come from such calculations and are further fueled by jury size (as  $n$  increases, so does the informational power of the conjecture “If I am pivotal, then it must be the case that every other juror is voting to convict.”). Such results call into question claims about unanimity’s beneficial normative properties.<sup>7</sup>

Driving the difference between the CJT result and newer results is the assumption that all jurors rigorously contemplate other jurors’ strategies. Questions about whether citizens think in such ways prompted clever experiments by Guernaschelli, McKelvey, and Palfrey (2000; henceforth GMP). Using students as subjects, they examined juries of different sizes ( $n = 3$  and  $n = 6$ ). The GMP experiments lend mixed support to the recent claims. Some jurors do vote to convict despite receiving innocent signals, and this behavior can lead to false convictions. But neither behavior happens as frequently as the NE on which FP focus suggests. GMP (p. 416) report that where: “Feddersen and Pesendorfer (1998) imply that large unanimous juries will convict innocent defendants with fairly high probability. . . this did not happen in our experiment.” In fact, and contrary to another conclusion from the 1998 paper, this occurrence happened less frequently as jury size increased.

We will now approach the jury decision problem in a different way. Before presenting our own model of such phenomena, we first review empirical research that motivates our theoretical framework.

There exists a substantial psychological literature on jury decision making. It is grounded in experiments built around mock juries with participants sampled from courthouse jury pools. The literature documents important attributes of how jurors think. Focal citations include a series of papers and books by Nancy Pennington and Reid Hastie. Their research begins with the premise that jurors encounter a massive database of evidence during a trial. The evidence is often presented in a scrambled order. Instead of being strictly chronological, plaintiffs and defendants produce different kinds of evidence at different times. From many jurors’ perspectives, the evidence is piecemeal and leaves many gaps in their attempts to understand what really happened.

How do jurors react? Pennington and Hastie explain their reactions with “story models.” Each juror attempts to make sense of the evidence by assembling it into a narrative format. A narrative comes from three sources: case-specific information acquired during the trial, a juror’s knowledge of similar events, and a juror’s expectations of what constitutes a complete story. Comparing the story model approach to other empirically-based explanations of jury decision making, MacCoun (1989: 1047) finds that it is “the only model in which serious consideration is given to the role of memory processes during the trial,” whereas Devine et al. (2001: 624) concludes that it is “the most widely adopted approach to juror decision making.”

These studies reveal interesting variations in story content. Some jurors use complex narratives to make sense of what they see. Others use simple narratives. For our purpose, just as important is the fact that many jurors are shocked to learn of such variations after the fact. For example, Pennington and Hastie (1990: 94, emphasis added) found not only that “many jurors tended to construct *only one* of the possible stories” but also that “*jurors were surprised to discover that there were other possible stories*” that fit the evidence. Many jurors construct a simple story as a means of understanding the evidence and provide

<sup>7</sup>Later work by Coughlan (2000) and Austen-Smith and Feddersen (2006) examines whether allowing jurors to participate in a straw poll prior to the final vote reduces the pathological effects of information aggregation identified in FP. Coughlan (2000) identifies an equilibrium where it does, but Austen-Smith and Feddersen (2006) find that this result is not robust to the introduction of interjuror uncertainty about whether other jurors are biased for or against conviction.

no evidence of having put any thought at all into the possibility that others drew different conclusions from the same evidence.

That jurors differ in these ways is consistent with other core findings in the psychological study of how people think. Building from studies by Cohen et al. (1955), Cacioppo and Petty (1982) began to document differences in how much people enjoy thinking about—and actually think about—complex matters. Whereas some citizens enjoy dealing with logical abstractions, others strive to minimize the mental effort devoted to such activities. Over the span of several decades, substantial variation in citizens’ “need for cognition” (henceforth NFC) has been observed (Wegener et al. 2000). Such variation explains and reinforces the variations in story quality observed by psychological jury scholars. Story model and NFC studies provide insight into the range of mental constructs on which jurors base their voting decisions.

## 4.2 *The Next Step*

At present, there is little interaction between the psychological and theoretical literatures just described. A recent quote (Hastie and Kameda 2005: 12) suggests both a reason for the isolation and a strategy for more effective interaction.

[GMP’s] empirical study is an antidote to a previous controversial paper that argued, on the basis of a theoretical model (not behavioral data), that unanimity rule without discussion was universally inferior to the majority rule (Feddersen and Pesendorfer 1998).

In the quote, the theory’s logic is unchallenged. But the theory’s relevance is called into question because it is not based on behavioral data.

To be sure, recent theoretical claims presume that jurors efficiently contemplate abstractions such as “what others must be thinking if I am pivotal.” It may be the case that all jurors think in such ways or proceed “as if” they have such thoughts. But what if some do not?

Contrary to the “as if” assumption, story model and NFC studies suggest that many jurors are in unfamiliar surroundings, receive little or no feedback about their actions, and have limited opportunity or motivation to think about how others decide. With such findings in hand, it is reasonable to ask whether integrating stronger psychological premises into a model like FP’s alters what we can conclude about the frequency of false convictions under unanimity rule.

We will now present a model that addresses this question. Like previous models, our model’s jurors are goal oriented, in that they prefer to acquit the innocent, and strategic, in that they plan their actions to maximize their expected utility. Like previous psychological work, the model’s jurors vary in how they think (or do not think) about the information that is presented to them. To leverage the kind of psychological variation in empirical work, we use the SCE concept to derive our conclusions.

Our model’s foundation is FP. It is a game with  $N = \{1, 2, \dots, n\}$  jurors that begins with Nature determining the state of the world. Let  $\Omega = \{G, I\}$ , where  $\Omega = G$  means that the defendant committed the crime in question and  $\Omega = I$  means that he did not.  $G$  and  $I$  occur with equal probability. No juror observes the true state of the world directly. Instead, each juror receives an *evidentiary signal*. As in previous models, each evidentiary signal is an independent Bernoulli random variable,  $m_j \in \{g, i\}$ , which, for each juror  $j$ , reveals the true value of  $\Omega$  with probability  $p \in (.5, 1)$  and the false value of  $\Omega$  with probability  $1 - p$ . After observing  $m_j$ , each juror casts a vote  $X_j \in \{A, C\}$ , where  $X_j = A$  is a vote by juror  $j$  to acquit and  $X_j = C$  is a vote to convict. We focus on unanimity, so if all  $n$  jurors choose  $C$ , then the group decision is  $C$ , otherwise it is  $A$ . All jurors prefer to convict only the guilty and to set only the innocent free:  $u(C, G) = u(A, I) = 0$  and  $u(C, I) = -q$  and  $u(A, G) = -(1 - q)$ ,

**Table 1** Differences between high-NFC and low-NFC jurors

	<i>Low NFC</i>	<i>High NFC</i>
Private signal permits “If I am pivotal . . .” thinking	No	Yes
Beliefs about $p$	$p = 1$ for everyone	They know the value of $p$
Beliefs about jury composition	“Everyone is like me”	They know the number of high- and low-NFC jurors
Conjecture about others’ strategies	“All vote informatively”	Depends on $p$ , $n$ , $q$ , and number of high-NFC jurors

where  $q \in (0, 1)$  is the same for all jurors and “characterizes a juror’s threshold of reasonable doubt” (FP: 24). Juror  $j$ ’s voting behavior is described by the strategy  $\sigma_j: \{g, i\} \Rightarrow [0, 1]$ , which maps evidentiary signals into a probability of voting to convict.

We now break from FP. We assume that the jury contains two kinds of jurors. Some jurors are high in NFC, and others are low NFC. The difference between the jurors is their ability to construct complex stories about what they do not observe and their motivation to imagine that other jurors think differently (about the evidence presented, what other jurors are thinking, etc.).

A low-NFC juror’s private signal contains her evidentiary signal along with the knowledge that unanimity is the decision rule and all jurors have identical utility functions. The private signal does not include the fact that their evidentiary signal was the result of a single draw from the Bernoulli distribution. Instead, they interpret their evidentiary signal as “the truth.” Technically, we assume they believe that every juror believes  $p = 1$ . Low-NFC jurors do not consider the possibility that other jurors may have received different signals. They do not think about what they do not observe. So our low-NFC jurors are like the jurors in the studies of Pennington and Hastie who were shocked to learn that other jurors constructed causal stories different than their own. They also resemble the subset of actors in the deliberation model of Hafer and Landa (2007), who craft strategies to maximize utility but do not process information via Bayesian updating because they “do not know what they do not know.”<sup>8</sup>

High-NFC jurors differ from low-NFC jurors in that their private signals are more informative. A high-NFC juror’s private signal includes their evidentiary signal and everything that was common knowledge in FP. Unlike their low-NFC counterparts, they also know the proportion of high-NFC and low-NFC jurors in the jury. Therefore, they are capable of the kind of information processing assumed in the recent generation of formal models (“My vote matters only when I am pivotal and if I am pivotal, it must be the case that . . .”). Table 1 describes the differences between the two kinds of jurors.

With this framework in hand, we use the model to reexamine the focal question of FP: With what frequency do false convictions occur? We conclude that the problem of false convictions increases with the proportion of high-NFC jurors. When all jurors are low NFC, unanimity rule minimizes the frequency of false convictions.

<sup>8</sup>Also see Tingley (2005). In reviewing work by Byrne et al. (2000), he highlights “actions (as opposed to inactions)” as being likely sources for the kinds of cognitive assessments that are relevant in many games.

To reach this conclusion, we make two additional assumptions. First, we follow the common practice of eliminating weakly dominated strategies from consideration. Second, like FP, we focus on “responsive” and “symmetric” equilibria.<sup>9</sup> *Responsiveness* requires that jurors change their vote as a function of their evidentiary signal with positive probability [i.e.,  $p\sigma_j(g) + (1 - p)\sigma_j(i) \neq (1 - p)\sigma_j(g) + p\sigma_j(i)$ ]. In FP, *symmetry* requires that similarly situated actors take identical actions. In our model, high-NFC and low-NFC jurors are not similarly situated—they receive different private signals. Hence, in our model, symmetry requires that all low-NFC jurors choose identical strategies and that all high-NFC jurors choose identical strategies. But it does not require that high- and low-NFC jurors choose the same strategies.

We begin with the case where all jurors are low NFC. To determine whether a particular set of strategies constitutes an SCE, we must determine whether a juror’s observations are consistent with her conjecture and beliefs.

*Low NFC Proposition:* If all jurors are low NFC, then all jurors voting informatively is the only responsive and symmetric SCE.

*Proof:* Every juror believes that all other jurors see the same signal. If a juror  $j$  observes  $m_j = G$ , then she believes that  $\Omega = G$  with probability 1. Given the knowledge that all jurors have identical utility functions, she conjectures that all other jurors are voting to convict. If  $\sigma_j(g) = 1$  (she votes to convict), then her belief and conjecture lead her to expect utility  $u(C, G) = 0$ . If  $\sigma_j(g) = 1 - z$ ,  $z \in [0, 1]$ , then she conjectures that her vote will preclude a unanimous guilty verdict with probability  $z$ . Given her belief and conjecture, she expects utility  $u(A, G) = -z(1 - q)$ . Since  $q \in (0, 1)$ , her expected utility is maximized at  $z = 0$ . Therefore, if  $m_j = G$ , then any responsive, symmetric SCE must include  $\sigma_j(g) = 1$ ,  $\forall j$ . If  $m_j = I$ , then the juror believes that  $\Omega = I$  and conjectures that all other jurors are voting to acquit. Whether she votes to convict or acquit, the defendant will be acquitted. Given her belief and conjecture, she expects utility  $u(A, I) = 0$  from any strategy  $\sigma_j(i) \in [0, 1]$ ; however, only  $\sigma_j(i) = 0$  survives weak domination. Q.E.D.

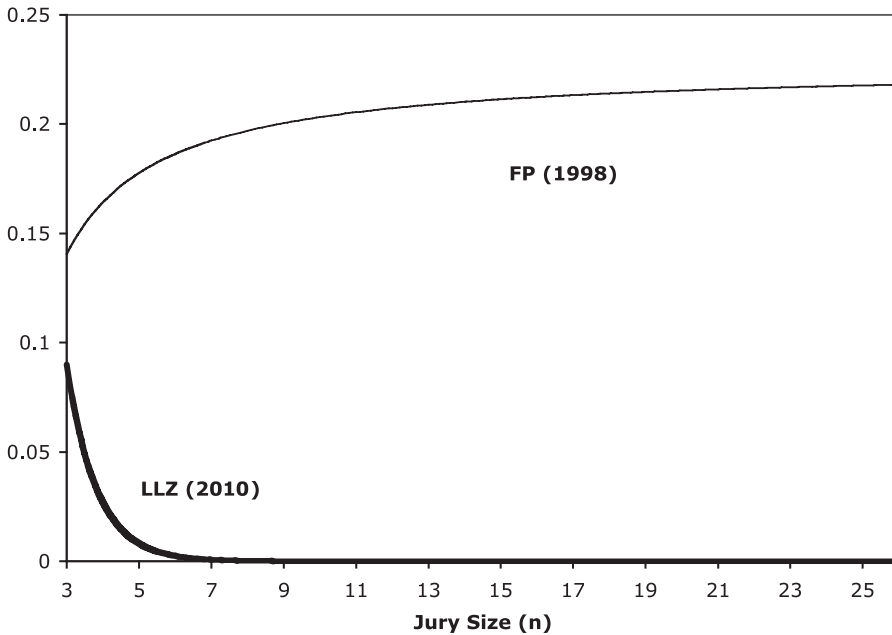
In this case, informative voting constitutes equilibrium behavior. This outcome is unlike FP’s Nash-based conclusion. Here, moreover, the false conviction probability is  $(1 - p)^n$ , as is true in the original Condorcet’s jury theory (CJT). In other words, a false conviction occurs only if *every juror* receives a false “guilty” signal when the true state of the world is innocent. If we take the normal size of the jury ( $n = 12$ ) and use the least-flattering assumption about signal quality in the theoretical papers cited ( $p$  approaches .5 from above), then the probability of a false conviction is roughly 1/4096. As signal quality or jury size increases, the probability of a false conviction goes to zero. We now consider the case where all jurors are high NFC.

*High NFC Proposition:* Under the technical conditions described in Proposition 2 of FP (p. 26), if all jurors are high NFC, then the only responsive and symmetric SCE entails  $\sigma(g) = 1$  and  $\sigma(i) > 0$ .

Here, the unique symmetric and responsive SCE is identical to FP’s unique and responsive NE. The proof follows accordingly and in the case described above ( $p \approx .5$ ) the probability of a false conviction diverges away from zero as jury size increases.

Now compare the two propositions. What the comparison reveals is that it is not strategic voting per se that generates FP’s high rate of false convictions—as low-NFC jurors in our

<sup>9</sup>Other equilibria exist, including all voters choosing to acquit regardless of their signal. This is true for both FP’s NE-based inferences and our SCE-based inferences.



**Fig. 2** The probability that an innocent defendant is convicted as a function of jury size for  $p = .7$  and  $q = .5$ .

model do not generate high false conviction rates. *Driving the increase in false convictions is the assumption that all jurors conjecture that all other jurors are thinking in the same manner as they are.*

To consider what these results imply for the normative qualities of unanimous verdicts with real juries, we recall from the psychological literature that most juries will likely contain a mix of high- and low-NFC jurors. In our model, the two kinds of jurors can be mixed in many different proportions, but a full mathematical treatment of behavior in all such cases is beyond the scope of this example. We can, however, use the results derived above to give some intuition about how the presence of jurors who vary in the kinds of stories they construct affects the probability of false convictions.

Suppose that there exists a jury containing 1 high-NFC juror and  $n - 1$  low-NFC jurors and, as in FP's focal example, let  $p = .7$  and  $q = .5$ . For low-NFC jurors, this case is observationally equivalent to that described in the "Low NFC Proposition." Therefore, any symmetric and responsive SCE must involve all such jurors voting informatively. Moreover, if  $n > 2$ , then this SCE includes the high-NFC juror voting to convict regardless of their evidentiary signal. To see why, note that the high-NFC juror recognizes (as in Austen-Smith and Banks 1996) that he is either pivotal or not pivotal and (as in FP) concludes that if he is pivotal under unanimity rule, then it must be the case that *every other juror is voting to convict*. So, if the high-NFC juror receives an innocent signal, he calculates the probability of guilt as  $Z = [(1 - p)p^{n-1}] / [(1 - p)p^{n-1} + p(1 - p)^{n-1}]$  and votes to convict if  $-q(1 - Z) > -(1 - q)Z$ . When  $p = .7$  and  $q = .5$ , this inequality is satisfied for  $n > 3$ . If he receives a guilty signal, he calculates the probability of guilt as  $Z' = p^n / [p^n + (1 - p)^n]$  and votes to convict if  $-q(1 - Z') > -(1 - q)Z'$ . For  $p = .7$  and  $q = .5$ , this inequality is satisfied for all  $n$ .

What is the probability of false convictions in this case? As Fig. 2 shows, it is far less than that reported in FP.

In our version, the probability of a false conviction is  $(1 - p)^{n-1}$ . This probability is lower than FP's because only a limited number of jurors vote contrary to their evidentiary signals. In FP, symmetry requires that if one juror votes against his evidentiary signal with positive probability, then all other jurors must do the same. This attribute of FP's example is what drives the false conviction probability away from zero as jury size grows. In our version of the example, letting high- and low-NFC jurors have different conjectures about others' strategies drives this same probability to zero as jury size grows. More generally, the extent to which the pathologies of unanimity rule pointed out by FP occur in our model is a function of the ratio of high-NFC to low-NFC jurors. When all jurors are low NFC, unanimity rule retains the beneficial normative properties attributed to it by the CJT. As high-NFC jurors appear, so does the probability of false convictions.

Our results imply that understanding how often unanimity rule convicts the innocent requires knowledge of how jurors think. In particular, we should examine questions such as "Under what conditions are  $w$  of  $N$  jurors likely to act like high NFCs?" For cases where most jurors are like low NFCs, our model suggests that unanimity rule will generate few false convictions. But where evidence suggests that most or all jurors are high NFCs who think in ways that the recent generation of game-theoretic models describes, we would follow FP in questioning the virtues of unanimous verdicts.

### 4.3 *Comparing Our Explanation to That of GMP*

Viewing jury decision making through the SCE's conceptual lens complements the approach adopted by GMP, whose empirical work we referenced earlier. Their work is based on the notion of a quantal response equilibrium (QRE). Like SCE, QRE addresses empirical challenges caused by the gap between actual human reasoning and that posited in Nash-based concepts—but SCE and QRE do this in different ways.

In a QRE, Nash-based behavior (which leads to a probabilistic distribution over actions) is assumed. In GMP, statistical procedures are used to estimate the shape of that distribution with respect to the data in hand. So, in the GMP paper, the QRE does not provide an *ex ante* prediction about behavior that is superior to FP's NE prediction, but it does provide the basis for a statistical analysis of the data from which a stochastic error term is derived *ex post*. Once the error term is fed back into the theoretical analysis, GMP's improved explanation emerges.

SCE, by contrast, encourages scholars to think about how actors think about one another (including probabilistic distributions of such actions). In our example, we relied on the psychological jury literature to inform assumptions about a range of possible juror beliefs and conjectures. This linkage led us to derive theoretical conclusions not from an initial assumption of Nash-based best responses to the strategies of others but from observed behavioral variations in psychology-based jury studies.

The SCE and QRE concepts challenge researchers to increase the transparency and rigor with which they deal with the psychological underpinnings of strategic behavior. Whether SCE, QRE, or a Nash-based equilibrium concept is most appropriate for political contexts is an interesting question.<sup>10</sup> We contend that such questions are, at least in part, empirical.

<sup>10</sup>Both QRE and SCE can explain GMP's observation of a widening gap between the theoretical predictions and the experimental observations as jury size grows. GMP treat the gap as a result of respondents making errors in their attempts to implement NE strategies. Our SCE-based explanation is that as jury size grows, the cognitive effort required to act like a high-NFC voter (If I am pivotal, . . .) grows. Faced with a harder "math problem," and holding motivation constant, jurors are more likely to seek simple stories of cause and effect—they are more likely to act like low-NFC jurors. Therefore, the gap between the probability of false convictions and the observed rate of false convictions should grow with jury size.

In situations where empirical research or other theory suggests that political actors are unlikely to share conjectures about one another's strategies and beliefs about their types, an SCE-based approach provides analytic advantages. Where evidence suggests cultural norms or institutions lead people to have probabilistically convergent conjectures and beliefs, a QRE-based approach will provide advantages. When evidence suggests that people reason as if they share conjectures and beliefs, then Nash-based concepts make sense.

## 5 Conclusions

Common game-theoretic equilibrium concepts used by political scientists entail implicit assumptions about how people reason. One assumption is that political actors share conjectures about one another's strategies. But evidence from psychology and related fields make it unlikely that all political actors in important decision contexts share such thoughts. This paper responds to that evidence. We contend that attempts to reconcile equilibrium concepts with observed psychological phenomena can allow scholars to derive theoretical conclusions from sound empirical foundations.

To be sure, implementing SCE poses new challenges. On the one hand, it allows us to expand the empirically defensible range of conjectures that can be integrated into models. On the other hand, if we want to reduce the number of focal equilibria, then the SCE approach induces us to provide a more detailed psychological account than is true for many Nash-based approaches. For some, such psychological accounts will represent Pandora's boxes—questions that are best unopened. We disagree. The SCE concept does not create tensions between key theoretical assumptions and psychological factors, it only makes apparent the logical consequences of ignoring these tensions.

Another challenge of using SCE is multiplicity of equilibria. In many applications, using SCE will increase the range of strategy profiles that are in equilibria. One way to reduce the number of SCE is to restrict the range of the private signals—as we have done in the jury example. We understand that some scholars sometimes see such restrictions as arbitrary, or at least unusual. Since an infinite number of such restrictions are possible, researchers need to have a very strong rationale for basing conclusions on any particular restriction. Our view, which is reflective of our desire to develop “applied models,” is that paying close attention to empirical work that documents phenomena relevant to actors' abilities to form conjectures is one way to justify such a restriction.

In general, scholars can benefit from asking informed, direct, and concrete questions about how the actors they model view their environs and those around them. Psychology is producing a growing range of findings about the kinds of information to which political actors attend and remember (see, e.g., Kuklinski 2001, 2002). Such information can play an important role in clarifying the conditions under which key political actors share beliefs or conjectures. If conditions are such that political actors are unlikely to see one another—or important elements of their decision context—in similar ways, then the SCE concept can be a constructive means for developing logically rigorous explanations of important political phenomena.

## References

- Aumann, Robert. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1:67–96.
- Aumann, Robert, and Adam Brandenberger. 1995. Epistemic conditions for Nash equilibrium. *Econometrica* 63:1161–80.
- Austen-Smith, David, and Jeffrey S. Banks. 1996. Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review* 90:34–45.

- Austen-Smith, David, and Timothy J. Feddersen. 2006. Deliberation, preference uncertainty, and voting rules. *American Political Science Review* 100:209–17.
- Bernheim, B. Douglas. 1984. Rationalizable strategic behavior. *Econometrica* 52:1007–28.
- Bjork, Elizabeth Ligon, and Robert Bjork. 1996. Continuing influences of to-be-forgotten information. *Consciousness and Cognition* 5:176–96.
- Byrne, Ruth M. J., Susana Segura, Ronan Culhane, Alessandra Tasso, and Pablo Berrocal. 2000. The temporality effect in counterfactual thinking about what might have been. *Memory and Cognition* 28:264–81.
- Cacioppo, John T., and Richard E. Petty. 1982. The need for cognition. *Journal of Personality and Social Psychology* 42:116–31.
- Cohen, Arthur R., Ezra Stotland, and Donald M. Wolfe. 1955. An experimental investigation of need for cognition. *The Journal of Abnormal and Social Psychology* 51:291–94.
- Condorcet, Marquis de [1785] 1994. *Essai sur application de l'analyse a la probabilité des décisions rendues a la pluralité des voix*. Paris: Translation by Iain McLean and Fiona Hewitt.
- Coughlan, Peter J. 2000. In defense of unanimous jury verdicts: Mistrials, communication, and strategic voting. *American Political Science Review* 94:375–93.
- de Figueiredo, Rui, Jack Rakove, and Barry R. Weingast. 2006. Rationality, inaccurate mental models and self-confirming equilibrium: A new understanding of the American Revolution. *Journal of Theoretical Politics* 18:384–415.
- Dekel, Eddie, Drew Fudenberg, and David K. Levine. 1999. Payoff information and self-confirming equilibrium. *Journal of Economic Theory* 89:165–85.
- . 2004. Learning to play Bayesian games. *Games and Economic Behavior* 46:282–303.
- Devine, Dennis J., Laura D. Clayton, Benjamin B. Dunford, Rasmy Seying, and Jennifer Pryce. 2001. Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy, and Law* 7:622–727.
- Feddersen, Timothy, and Wolfgang Pesendorfer. 1998. Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political Science Review* 92:23–35.
- Fudenberg, Drew, and David M. Kreps. 1995. Learning in extensive-form games I: Self-confirming equilibrium. *Games and Economic Behavior* 8:20–55.
- Fudenberg, Drew, and David K. Levine. 1993. Self-confirming equilibrium. *Econometrica* 61:523–45.
- . 1998. *The theory of learning in games*. Cambridge, MA: MIT Press.
- Gibbons, Robert. 1992. *Game Theory for Applied Economists*. Princeton, NJ: Princeton University Press.
- Govindan, Srihari, and Robert B. Wilson. 2008. Refinements of Nash equilibrium. In *The new Palgrave dictionary of economics*. 2nd ed., eds. Steven N. Durlauf and Lawrence E. Blume. Hampshire, UK: Palgrave Macmillan. Available at [http://www.dictionaryofeconomics.com/article?id=pde2008\\_R000242](http://www.dictionaryofeconomics.com/article?id=pde2008_R000242)> doi:10.1057/9780230226203.1155
- Guernaschelli, Serena, Richard D. McKelvey, and Thomas R. Palfrey. 2000. An experimental study of jury decision rules. *American Political Science Review* 94:407–23.
- Hafer, Catherine, and Dimitri Landa. 2007. Deliberation as self-discovery and institutions for political speech. *Journal of Theoretical Politics* 19:329–60.
- Harsanyi, John. 1967. Games with incomplete information played by 'Bayesian' players I: The basic model. *Management Science* 14:159–82.
- . 1968. Games with incomplete information played by 'Bayesian' players II: Bayesian equilibrium points. *Management Science* 14:320–34.
- Hastie, Reid, and Tatsuya Kameda. 2005. The robust beauty of majority rules in group decisions. *Psychological Review* 112:494–508.
- Kandel, Eric R., James H. Schwartz, and Thomas M. Jessell. 1995. *Essentials of neural science and behavior*. Norwalk, CT: Appleton and Lange.
- Kreps, David, and Robert Wilson. 1982. Sequential equilibria. *Econometrica* 50:863–94.
- Kuklinski, James H. ed. 2001. *Citizens and Politics: Perspectives from Political Psychology*. New York: Cambridge University Press.
- , ed. 2002. *Thinking about political psychology*. New York: Cambridge University Press.
- Lupia, Arthur, Mathew D. McCubbins, and Samuel L. Popkin. 2000. Beyond rationality: Reason and the study of politics. In *Elements of reason: Cognition, choice, and the bounds of rationality*, eds. Lupia Arthur, Mathew D. McCubbins, and Samuel L. Popkin, 1–20. New York: Cambridge University Press.
- MacCoun, Robert J. 1989. Experimental research on jury decision making. *Science* 244:1046–9.
- McKelvey, Richard D., and Thomas R. Palfrey. 1995. Quantal response equilibria in normal form games. *Games and Economic Behavior* 10:6–38.

- Myerson, Roger B. 2006. Game-theoretic consistency and international relations. *Journal of Theoretical Politics* 18:416–33.
- Nash, John. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences* 36:48–9.
- Pearce, David G. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52: 1029–50.
- Pennington, Nancy, and Reid Hastie. 1990. Practical implications of psychological research on juror and jury decision making. *Personality and Social Psychology Bulletin* 16:90–105.
- . 1993. Reasoning in explanation-based decision making. *Cognition* 49:123–63.
- Rubinstein, Ariel. 1998. *Modeling bounded rationality*. Cambridge, MA: MIT Press.
- Satz, Debra, and John Ferejohn. 1994. Rational choice and social theory. *Journal of Philosophy* 91:71–87.
- Schacter, Daniel L. 1996. *Searching for memory: The brain, the mind, and the past*. New York: Basic Books.
- . 2001. *The seven sins of memory: How the mind forgets and remembers*. Boston, MA: Houghton-Mifflin.
- Tingley, Dustin. 2005. Self-confirming equilibria in political science: Cognitive foundations and conceptual issues. Philadelphia, PAPaper presented at the 2005 Annual Meeting of the American Political Science Association.
- Turner, Mark. 2000. Backstage cognition in reason and choice. In *Elements of reason: Cognition, choice and the bounds of rationality*, eds. Lupia Arthur, Mathew D. McCubbins, and Samuel L. Popkin, 264–86. New York: Cambridge University Press.
- . 2001. *Cognitive dimensions of social science*. Oxford: Oxford University Press.
- Wegener, Duane T., Norbert L. Kerr, Monique A. Fleming, and Richard E. Petty. 2000. Flexible corrections of juror judgments: Implications for jury instructions. *Psychology, Public Policy, and Law* 6:629–54.