A Theory of Prejudice and Why it Persists (*or* To Whom is Obama Still Black?)

Arthur Lupia, University of Michigan

Version 1.1. Comments appreciated.

A Theory of Prejudice and Why it Persists (*or* To Whom is Obama Still Black?)

Abstract

Anti-black prejudice affects how citizens evaluate black candidates. Can continued contact with such candidates change prejudice's role in subsequent evaluations? Scholars disagree. To clarify when continued contact reduces evaluative prejudice, I integrate psychological insights into a formal model.

I find that widely-cited factors such as exposure frequency and motivation to avoid appearing prejudiced are not sufficient to reduce evaluative prejudice. Instead, citizens must associate negative personal consequences with their prejudice. Even then, prejudice's role in subsequent evaluations declines *only if* a citizen's goals, context, and beliefs about the future combine to make them rethink prejudiced beliefs. This result implies that past empirical studies about prejudice reach different conclusions because only some supply the necessary conditions. The result also explains how contextual variations can cause huge evaluative differences to emerge amongst similarly prejudiced people.

In sum, I find that prejudice change is possible, but only for certain people in special circumstances.

Introduction

Barack Obama has a Kenyan father and a Caucasian mother. These facts of his life are constant. They are, in principle, knowable to all.

At the same time, racial identity is a social construction. Social implications of a person's racial identity, such as his suitability for political office, need not be constant. These implications of identity are potentially malleable. When can they change? I will address this question in a political context.

Many candidates for office, including Barack Obama, are characterized by negative black stereotypes and evaluated unfavorably as a result. Piston (2010) finds such effects after examining how citizens' endorsements of negative stereotypes correspond to their support for Democratic presidential candidates from 1996 to 2008. He finds that "negative stereotypes about blacks significantly eroded white support for Barack Obama" but did not affect white voter support for any white Democrat. He concludes "that white voters punished Obama for his race" (also see Pasek et. al. 2009).

For whom will continued contact with a black candidate change prejudice's role in subsequent evaluations? For insight, we can turn to research on prejudice change. Scholars working in this field disagree about the malleability of racial stereotypes and associated prejudices. Bargh (1999: 378), for example, famously argued that "[o]nce a stereotype is so entrenched that it becomes activated automatically, there is really little that can be done to control its influence."

A more common claim is that contact reduces prejudice. From this perspective, individuals learn new things about other races through contact with them and become less prejudiced as a result. Here, contact refers to various means of obtaining new information (e.g., in person, electronic media, other individuals' accounts) about people of another race.

In a meta-analysis of hundreds of psychological experiments on intergroup contact and prejudice, for example, Pettigrew and Tropp (2006) find that almost all such contacts (94%) reduce prejudice. Political scientists also claim that contact with black officeholders reduces prejudice. Hajnal (2007), for example, finds that as white voters become familiar with black mayors, prejudice plays a diminished role in subsequent evaluations. He concludes (2007:3) that a black mayor "essentially changes the way that many white Americans think about the black community and therefore subtly alters the nature of racial politics and race relations in this country." Such findings resonate with Kurzban, Cosmides, and Tooby's (2001: 15391) claim that "race is a volatile, dynamically updated cognitive variable, easily overwritten by new circumstances." About such claims, however, Bargh (1999: 362) counters that "conclusions drawn from the data have overestimated the degree to which…stereotypes can be controlled through good intentions and effortful thought-and thereby have underestimated the extent to which stereotypes continue today to cause problems in social relations."

It is logically impossible for all of these claims to be correct simultaneously. A stereotype cannot be rigid, unresponsive and entrenched at the same time that it is easily overwritten.

In developing a constructive response to the disagreement, it is worth noting that laboratory experiments are the evidentiary source of most of this literature's competing claims. Experiments have important advantages and disadvantages in this context. An advantage is the ability to create observations that speak directly to the truth-value of a focal causal or existential proposition. An experiment that produces new, direct, and definitive observations of a focal proposition's truth-value can provide great inferential power. Experiments' disadvantages are apparent when the propositions we want to evaluate refer to dynamic and complex conditional relationships between multiple, continuous factors. Most experiments on stereotypes and prejudice vary a single value of a single factor (Pettigrew 2008). More dynamic multi-value or multi-factor research designs are rare in the literature on contact and prejudice.

Scholars are increasingly interested in understanding how dynamic and complex interactions amongst psychological factors (particularly factors that differentiate individuals) and contextual variables affect prejudice. As Pettigrew and Tropp (2006: 768) argue, "more elaborate models are needed to integrate and account for these varied intergroup effects…multilevel models that consider both positive and negative factors in the contact situation, along with individual, structural, and normative antecedents of the contact, will greatly enhance researchers' understanding of the nature of intergroup contact effects." Since individual differences and contextual factors can interact in multiple and dynamic ways, and since such interactions can have complex and conditional relationships to how contact affects prejudice, a model that complements the advantages of experiments by offering clarity about which interactions and relationships are – and are not – logically possible can expand our inferential power.

In what follows, I examine questions of contact and prejudice in a political context. Specifically, I examine how new information about a black candidate for office affects prejudice's role in subsequent evaluations of the candidate. To make progress on this matter, I develop a model that uses psychology insights as inputs and offers their joint logical implications as outputs. These outputs clarify which of an important set of more complex propositions about contact and prejudice can and cannot be true given the inputs.

This work's main substantive implication is that widely-discussed factors such as frequency of contact with out-group members and external motivations to avoid appearing prejudiced are neither necessary nor sufficient to reduce prejudice in evaluations. Instead, a citizen must associate a negative personal consequence with their prejudicial evaluation (i.e., the citizen must realize that the prejudice inhibits achievement of desired goals). A citizen's ability to have such realizations, in turn, depends on the content of the information to which they are exposed and how their circumstances affect their motivation to process such information.

These findings imply that evaluative prejudice declines only if a person's goals, environment, and feelings of efficacy combine in ways that make them rethink their current beliefs about blacks. In particular, I prove that for contexts that are common to many citizens, the following conditions *are required* for contact to reduce prejudice in candidate evaluations:

- A person must be in a situation where they can *associate their application of a prejudice with a negative personal consequence*. If they cannot relate the negative consequence to the prejudice, then they have no incentive to consider changing their use of it in candidate evaluations.
- If a person associates a negative outcome with the prejudice, changing prejudice's role in subsequent evaluations does not occur automatically. The person must be *willing to rethink their beliefs* to make that change happen.
- Willingness to rethink beliefs, in turn, requires *sufficient motivation* and *expectation of efficacy*.
- *Sufficient motivation* refers to the idea that continuing to act on the prejudice will cause additional negative personal consequences in the future. If, by contrast, the person is convinced that the negative outcome they observed was a one-time-thing, their motivation to change is nil.
- The *expectation of efficacy* comes from the belief that an attempt to change a prejudice will mitigate future negative consequences and that the benefit of such mitigation outweighs the costs inherent in trying to rethink a prejudice and change one's actions accordingly. People who reside in contexts where

prejudices can be expressed without negative consequences are less likely to expect efficacy than those who reside in contexts where such expressions are taboo.

Put another way, if even one of these conditions fails to hold, evaluations can remain prejudiced regardless of how much evidence contradicting relevant anti-Black stereotypes is available.

Two concepts, *evaluative strategy* and *therapeutic strategy*, are important in this explanation. An evaluative strategy converts psychological phenomena and contextual information into an evaluation. A therapeutic strategy converts similar inputs into a decision about how a person reacts when confronted with a contradiction to a prejudicial belief. The advantage of analyzing evaluation and therapy as strategic is that it focuses attention on whether seemingly plausible explanations of when people will exert effort to change a prejudice – or are not -- logically consistent with more fundamental psychological and contextual phenomena. Studying the problem in this way also offers an explanation of why previous empirical studies produce such different conclusions about prejudice change. In short, some research designs supply the necessary conditions for prejudice change identified above. Others do not.

In sum, changing prejudice's role in candidate evaluations requires more than external motivations to appear non-prejudiced. Even exposure to massive amounts of counter-stereotypical information can be insufficient to reduce evaluative prejudice. Instead, reducing evaluative prejudice requires particular combinations of perception, incentive, and context – combinations that, for many citizens, are unlikely to ever occur.

A consequence of this work is that hopes for a post-racial society in which racial prejudice does not influence candidate evaluations requires much more than a black president. The role of prejudice in a citizen's evaluations will not change unless they have opportunities to see, and are motivated to consider, that their prejudice is a cause of their own pain. So, a post-racial America is possible. But the path from this present to that future is unlikely to be short, nor is it likely to be traveled by everyone. Long after his presidency, racial prejudice will remain salient in many citizens' evaluations of Barack Obama and other black candidates for office.

The paper continues as follows. In Section 2, I describe the models' psychological foundations. In Section 3, I use these inputs to create a model. I use the model to clarify how continued contact affects evaluative prejudice. In Section 4, I discuss implications of the findings including whether, and for whom, racial prejudice will continue to affect evaluations of President Obama.

2. Foundations

In this section, I provide a brief overview of research on prejudice change. I begin by describing competing positions taken in scholarly debates about prejudice change. I then describe how models can complement extant research by clarifying when important empirical claims can and cannot be true.

2.a. Malleability

When we encounter a new person with whom we are likely to have subsequent relations, we evaluate that person through a combination of conscious and unconscious means (Fiske and Taylor 1984). Individuating features that are easily detectable and that are plausibly correlated with likely outcomes of interacting with the person become highly relevant.

When citizens evaluate political candidates, race-related attributes can come to mind. One is skin color, a focal marker of racial identity. While it is a fact that the skins of various people have differing pigments, how pigment affects evaluations is not solely a function of the pigment itself. The dependence of evaluation on context suggests the potential cognitive malleability of color-related phenomena that we often perceive as a constant, such as another person's racial identity. As Churchland and Sejnowski (1992: 211) describe:

> "[c]olor perception depends on wavelength, but it is not identical to wavelength. For example, note that a "red wagon may look red under a wide range of illuminations, including broad daylight, dusk, candle light, fluorescent light, and so forth, where the physical wavelengths actually impinging on the retina vary dramatically."

In other words, while we often think of an evaluation as the product of a relationship between attributes of an object and attributes of the perceiver, evaluation is more accurately described as the product of an interaction between the perceiver, the object, and the context in which the two meet. What types of interactions can cause prejudice to change in the context of political evaluations?

Allport's (1954) *contact hypothesis* provides a common framework for answering such questions. Many people interpret the hypothesis as "contact decreases prejudice." But Allport argued that contact reduces prejudice only in the presence of: equal status amongst groups, common goals, intergroup cooperation, and the support of authority. Pettigrew used meta-analyses to evaluate whether Allport's conditions were necessary for intergroup contact to reduce prejudice (see, e.g., Pettigrew 1998, Pettigrew and Tropp 2006, Pettigrew 2008). He found that they were not. Contact reduces prejudice even when one or more of Allport's conditions are not satisfied. Pettigrew and Tropp (2006) subsequently report that "contact typically reduces intergroup prejudice" (751). They find that "94% of the samples show an inverse relationship between contact and prejudice" (757).

Work in evolutionary psychology reaches a similar conclusion. Kurzban, Tooby, and Cosmides (2001: 15391) claim that "race is a volatile, dynamically updated cognitive variable, easily overwritten by new circumstances." Their argument begins with the premise that "categorizing others by their race is a precondition for treating them differently according to race" (2001: 15387). They then argue that "categorizing individuals by race is not inevitable…. encoding by race is instead a reversible byproduct of cognitive machinery that evolved to detect coalitional alliances." (2001: 15387). Their evidence comes from an experiment where subjects answer questions about members of two different teams. Each team has black and white members but race is not relevant to which team players are on. The authors expect that "[t]he strength of race encoding will be diminished by creating a social context in which (i) race is no longer a valid cue…and (ii) there are alternate cues that do reliably indicate coalitional affiliation" (2001: 15388). They find that "subjects bring the tendency to categorize by race with them into the experiment, but then begin to lose it as the circuitry detects that it no longer predicts relevant coalitions within this context." (15390). They conclude that "What is most striking about these results is just how easy it was to diminish the importance of race by manipulating coalition – especially given the repeated failure over decades to find other means to influence racial encoding" (15391).

Other research finds less malleability. For example, Gaertner and Dovidio (1986) find that prejudice does not disappear as much as it changes form, a phenomenon they call aversive racism. As Fiske (1998:360) describes, when people's "behavior can be explained away by other factors (i.e., when they have a non-racial excuse), or when situational norms are weak, ambiguous, or confusing…then aversive racists are more likely to discriminate overtly because they can express their racist attitudes without damage to their nonracist self-concept."

Another phenomenon that allows prejudice to persist is called subtyping. As Allport (1954:23) defined it, "There is a common mental device that permits people to hold prejudgments even in the face of much contradictory evidence. It is the device of admitting exceptions...By excluding a few favored cases, the negative rubric is kept intact for all other cases." In other words, people who observe blacks whose behaviors contradict their stereotype, can find it comforting to conclude that the person in question is "exceptional" and "not really black." Recent decades have seen increased interest in when subtyping occurs (see. e.g., Weber and Crocker 1983, Hewstone 1994, Fiedler 1996). Consider, for example, Sinclair and Kunda (1999). Their subjects were evaluated by an authority figure that was introduced to them during the experiment. The authority figure was either black or white.[1] The main finding is that when the figure praises subjects, they are motivated to see him in a positive light and inhibit negative stereotypes. By contrast, when the figure criticizes subjects they become motivated to disparage the authority figure and activate negative stereotypes. Put bluntly, "a Black professor who delivers praise may be categorized and viewed as a professor, whereas a Black professor who delivers criticism may be categorized and viewed as a Black person" (Sinclair and Kunda 1999: 885).

Other studies reveal that persistence in prejudice has important consequences. Hochschild (2001: 324) shows that whites hold a wide range of attitudes about blacks including some that are quite mistaken. She then demonstrates that these differences in white attitudes play a significant role in how whites evaluate black-related political phenomena – often to the detriment of policies that tend to benefit blacks. So despite ostensibly having access to the same historical information about blacks, many whites sustain false beliefs about blacks that affect a broader set of political and policy opinions.

2.b. Models as Complements

The literature on prejudice contains many contradictory claims about change. Some portray change as easy. Others disagree. It is logically impossible for all of these claims to be correct simultaneously. A stereotype cannot be simultaneously have all of the attributes described above (e.g., rigid, unresponsive, entrenched, and easily overwritten).

[1] The authority figure is presented as nearby but in a different location. He communicates with subjects via video. A common video is used across subjects within an experimental group, which allows an identical presentation of the stimulus within experimental groups.

In beginning to think about whether and how the empirical claims listed above relate to prejudice change in a candidate evaluation context, it may be useful to review the bases for these claims. Recall, for example, Pettigrew and Tropp's (2006) claim "94% of the samples show an inverse relationship between contact and prejudice" (757). Regarding the applicability of that claim to politics, a question to ask is "*94% of what?*" The underlying meta-analysis compiles findings from hundreds of individual studies. But we have very little information about the extent to which this set of observations is representative of intergroup contact effects on evaluative prejudice in political contexts or in general. In fact, given that so many of the observations come from experiments that involve college undergraduates and that vary a single value of a single factor, we can infer that the observations do not provide a representative view of more dynamic and logically conditional relationships between contact and prejudice.

When thinking about the applicability of the Sinclair and Kunda experiments, it is worth noting that the authority figure's race, and his praise or criticism, constitute just about all of the information that experimenters give subjects. In other circumstances, particularly political circumstances, the informational environment is more diffuse. Moreover, the experiment is designed so that the exposure to the black authority figure impinges directly on subjects' self-esteem. By contrast, people can choose to ignore what a political candidate does. So the question becomes, under what conditions would we expect the presence of a black president to affect prejudice's role in subsequent evaluations?

In what follows, I want to pursue a complementary approach. I want to characterize a broader set of conditions in which contact does and does not reduce prejudice. In so doing, I will derive a set of conditions that sorts claims about prejudice change by the conditions in which they can and cannot occur. For example, I will argue that the circumstance that "erases race" in studies such as Kurzban, Tooby, and Cosmides (2001), and others reviewed above, are special cases of a more dynamic set of conditions that further clarify when contact can and cannot reduce evaluative prejudice.

3. A Theory of Prejudice Change

I now introduce a theory whose purpose is to clarify how exposure to a black officeholder affects prejudice's role in subsequent evaluations. I draw the model's inputs from the psychological studies described above and from mathematical approaches to belief change. I then derive joint logical implications of these inputs and use these outputs to address questions about prejudice change.

The theory consists of two formal models. Model 1 is a game-theoretic representation of how prejudice affects evaluations when people vary in their opportunities and motivations to process racial information. Model 2 is a decision theoretic treatment of what happens after a person in Model 1 associates their prejudice with a negative personal consequence. To simplify the presentation, I develop each model separately and then use their joint logic to identify necessary conditions for prejudice change in the context specified by the model. I show that under a wide range of conditions, including conditions under which one would not normally expect racial prejudice to persist, even frequent exposure to someone like Barack Obama is far from sufficient to alter how prejudice affects evaluations.

3.A. Model 1 Initial Premises

Model 1 features two players, called Citizen 1 (C1) and Citizen 2 (C2). While C1 and C2 are similar in many ways, they differ in one important way. C1 can know things about blacks that C2 does not know. I draw conclusions about changes in evaluative prejudice by identifying conditions under which C2 can – and cannot – learn from C1. So, if we think of C2 as a bigoted citizen, we may think of C1 as a person or entity that C2 observes. Perhaps C1 is a neighbor or a television personality. Or C1 may be a collective entity such as an interest group or a political party that makes statements about a specific black candidate or about blacks in general.

The citizens play an N-period "evaluation game." N, the number of periods, is finite but can be arbitrarily large. In each of the game's $n \in N$ periods, C1 and C2 must say which of two candidates they prefer. The two candidates are not players in the game – they exist only for citizens to evaluate. Figure 1 depicts the sequence of moves in a single period of Model 1. Except where noted, and there will be important exceptions, I assume that all aspects of the game are known to both players. In particular, I assume that C1 and C2 can remember their own past actions, as well as anything that they observed in previous periods.

[FIGURE 1 ABOUT HERE.]

In each period, C1 and C2 evaluate candidates with respect to an objective. The objective can represent a material aspiration (such as getting a certain policy passed) or a non-material aspiration (such as living in accordance with a particular moral or ethical standard). I assume that a citizen benefits (receives higher utility) when he favors the candidate that is best able to help him achieve his objective in that period. I represent this difference by saying that in each period, one candidate is more *skilled* than the other.

Skill represents a candidate's ability to complete their constitutional duties, to work effectively with other branches of government, to maintain public support of important policies, or to act in accordance with a particular moral or ethical standard. When evaluating a president, the evaluation's point of comparison can be someone else who the citizen imagines can do the job. The game's N periods represent N such evaluations.

I focus on the case where both citizens have the same objective in each period and where C2 may be ignorant of this fact (in ways that I shall describe below). This case is helpful because it is simple and stark. For example, it would seem that C2 can increase his utility by simply mimicking C1. Given that my main result is that prejudice change is difficult, the fact that such a circumstance exists makes the main result more difficult to achieve and potentially more informative as a result.

As just mentioned, C2 may lack information not only about the candidates' skills, but also about C1's motives. Model 1 includes cases where C2 falsely believes that he and C1 have different objectives. This lack of information may lead C2 to discount or ignore his observations of C1. These actions, in turn, can allow racial prejudice to affect C2's evaluations despite the fact that there is no objective rationale for doing so. Which brings us to how prejudice enters the model.

Besides skill levels, candidates have one other potentially salient attribute – their races. In each period, one of the candidates is black and the other is white. I denote Citizen i's, $i\in\{1,2\}$, choice to favor the white candidate or the black candidate in period n as $f_{in}\in\{0,1\}$, where $f_{in}=1$ denotes citizen i favoring the black candidate in period n and $f_{in}=0$ denotes favoring the white candidate.

An important element of the model is what citizens know about the relationship between race and skill. In some periods, the black candidate is more skilled than the white candidate. In other periods, the opposite is true (i.e., which candidate is better able to help citizens can vary from objective to objective).

For simplicity, I model this variation by holding the white skill level constant across periods and allowing black skill levels to vary. Let $\theta_n\in\{0,1\}$ denote the black candidate's skill level in period n, where $\theta_n=1$ means that the black candidate is more skilled with respect to that period's objective and $\theta_n=0$ means that the black candidate is less skilled. I assume that skill levels are independently determined for each period. In other words, $\Theta\in[0, 1]$ is the (exogenous) probability that the black candidate is more skilled $(\theta_n=1)$ in any given period.

To represent the idea that citizens benefit from favoring skilled candidates, I define their motivation as follows. If Citizen *i* favors the white candidate in period *n*, then he earns a utility of .5 for that period ($U_{in}(f_{1n}=0)=.5$). A citizen's utility from favoring a black candidate depends on θ_n. In periods where the black candidate is more skilled, favoring him yields a utility of 1 ($U_{in}(f_{1n}=\theta_n=1)=1$). When the black candidate is less skilled, favoring him yields a zero payoff ($U_{in}(f_{1n}=0,\theta_n=0)=0$). Across periods, the utilities for each player, denoted U_i, are simply summed ($U_i= \Sigma^n{}_1 U_{in}$).[2]

Model 1's key premise is that C2 varies in his ability to observe candidate skill levels. While C1 always knows which candidate is more skilled (the true value of θ_n), C2 may lack such knowledge and instead base his evaluation on racial prejudice. Variance in how C2's processes his observations of C1's actions, and in how this processing affects C2's evaluative prejudice, is my principal analytic focus.

In an important sense, Model 1 is similar in approach to Kurzban, Tooby, and Cosmides (2001). In both cases, the cognitive status of race is viewed as potentially flexible and as a product of political objectives. In both cases, individuals' abilities to achieve basic objectives are affected by whom they choose to support. The principal difference between my approach and theirs is scope. The conclusion of their argument is *an existence claim*. They argue that race can be erased and develop an experimental circumstance that leads to this outcome. By contrast, I will to characterize a broader set of conditions in which contact does and does not reduce evaluative prejudice. Below, I will argue that the circumstance that "erases race" in their study is a special case of the set of conditions that I identify as capable of reducing prejudice's role in political evaluations.

3.B. Citizen 2's Private Signal

In many game-theoretic models, nearly all attributes of the game, such as the underlying structure of beliefs and allocation of information, are assumed known to all players. The same is not true here. To tell a more realistic story about prejudice, I design Model 1 to allow greater variance in what players know. Since C1 knows all relevant

[2] This specification also implies risk-neutral citizens. This assumption is conservative relative to my conclusion that prejudice persists (i.e., the white candidate is favored) even when objective rationales for such prejudice are absent. Were I to assume risk-averse citizens who see black candidates as more risky than white candidates, the conditions in which prejudice persists would expand even further.

information, the model's knowledge variations pertain to C2. I also assume that Citizen 2 may be oblivious to his ignorance.

Following Fudenberg and Levine (1993, 1999), I represent knowledge variations as "private signals."[3] C1's private signal is completely informative about all aspects of the game. As it is a constant in this model, there is no value in denoting it further. The content of C2's private signal varies and is the model's focal element. A technical definition of C2's private signal is in the appendix. In words, it: either reveals the true value of Θ or it reveals nothing about whether blacks or whites are more skilled on average; either reveals the true value of θ_n or it reveals nothing about whether the black or white candidate is more skilled in period *n*; and it either reveals C1's utility function or it does not.

I analyze the game under three kinds of private signal for C2. These three kinds are sufficient to clarify how exposure to new information can change prejudice's role in C2's evaluations. I call the three private signals "completely informative", "minimally informative", and "able to observe some contradictions." They are defined as follows.

- A *completely informative* signal for period *n* gives C2 all information about the game. This signal is akin to having access to friends or a news program that perfectly reveals the correspondence between a candidate's race and skill.
- A *minimally informative* signal for period *n* reveals to Citizen 2 neither Θ, θ_n, nor C1's utility function. Hence, even though C2 can observe C1's evaluations, he may lack information about what the evaluations imply about the black candidate's skill level. This kind of signal is akin to a situation where citizens have no means for accessing definitive information about race and skill.
- The *able to observe some contradictions* signal differs from a minimally informative private signal only in that it allows C2 to observe the black candidate's skill level in any given period with a non-zero probability. To simplify the presentation, it is not necessary to specify what this probability is beyond stating that it is non-zero and is not correlated with the true value of θ_n.

Since C1 has complete information and since both citizens have identical motivations to favor the period's higher skilled candidate, I have set up a circumstance in

[3] The term "private signal" used here is not equivalent to the term "private information" that game theorists use to describe game attributes that are known to one player but not another. See Fudenberg and Levine (1999) for more information on the distinction.

which C2 can maximize his utility by simply mimicking C1's evaluations. However, C2 may be ignorant about what C1 wants. For example, in the minimally informative private signal case, C2 can represent a conservative citizen who sees a liberal's (C1's) evaluations but does not consider them relevant to his own evaluation strategy because he conjectures that liberals have fundamentally different values than he does (i.e., different utility functions; "he cares more about diversity than performance"). So, even though C2 can observe the actions of an informed peer, questions remain about whether this observation affects prejudice's role in subsequent evaluations.

3.C. Model 1 Sequence of Events and Equilibrium Concept

Each period's sequence of events is as follows. First, Nature (i.e., factors outside of the citizen's control; history) determines which candidate is higher skilled. Recall that Θ is the probability that the black candidate is more skilled in any given period. Next, C1 observes θ_n, the black candidate's skill level for period *n*, and evaluates the candidates ($f_{1n} \in \{0,1\}$). Next, C2 observes his private signal. C2 uses the private signal to develop beliefs about how race and skill relate. C2 then renders his evaluation, $f_{2n} \in \{0,1\}$.

I characterize strategies and outcomes in this model using the self-confirming equilibrium (SCE) concept (Fudenberg and Levine 1993, 1998). This concept is unlike the Nash equilibrium concept that is commonly associated with game theory. In a Nash equilibrium, a player's strategy is evaluated for whether or not it is a best response to the strategies of other players. In a SCE, by contrast, each player rationalizes their actions only with respect to their own beliefs and conjectures.

In Model 1, the key element of a SCE is the correspondence between what C2's private signal reveals, what C2 believes about the black candidate's skill level, and what C2 conjectures about how C1 evaluates candidates. If C2's beliefs and conjectures are consistent with his observations, then he has no incentive to reconsider his strategy.

In Model 1, the relevant strategy pertains to evaluation. An *evaluation strategy* is a function that converts a citizen's knowledge of the game, beliefs about aspects of the game that he does not know, and conjectures about other citizens' evaluation strategies into a rule for favoring one candidate over the other in each period. Intuitively, an evaluation strategy is a goal-oriented plan of action that converts a citizen's personal circumstance into an evaluation.

When no player has a rationale for changing their evaluation strategy, then their actions are "in equilibrium." The substantive implication of this equilibrium notion for our present purposes is as follows: When a set of beliefs and strategies are "in equilibrium," the logical implication is that if none of the inputs change, then the output will not change and, as a result, that such circumstances better characterize strategies and behaviors better that circumstances that are not in equilibrium.

This notion of equilibrium, while not grounded in the psychological literature on prejudice, is also not unknown to it. Consider, for example, the literature on motivated correction referenced above. In describing its main tenets, Fiske (1998: 363) who says that people "normally engage in cognitive shortcuts, unless motivated to go beyond them."

An appendix contains a technical definition of an SCE for this game. The most pertinent attributes of that definition are as follows. First, to be part of a SCE, an evaluative strategy must maximize a citizen's expected utility given what he knows, believes, and conjectures about the game. Second, to survive as part of an SCE, a citizen cannot observe anything during the game that contradicts his beliefs, or conjectures.

Unlike common equilibrium concepts such as Nash Equilibrium and Perfect Bayesian Equilibrium, we do not ask whether strategies are best responses to the strategies of other players. Instead, we ask whether each citizen's strategy is a best response *to her own* beliefs, conjectures, and observations. So if C2 has a minimally informative private signal, he can have mistaken beliefs about the black candidate's skill level and false conjectures about C1's utility function. However, if C2's beliefs and conjectures are never contradicted by what he sees (i.e., the information conveyed by his private signal), he has no reason to change anything – he has a rationale for maintaining his evaluative strategy. When the same is true for both citizens, then the strategies are "in equilibrium."

An implication of this way of thinking about prejudice's role in candidate evaluation is that we need not assume that citizens have the same belief about blacks and whites. Hence, C1 can know that the black candidate is better able to help the citizens achieve their objectives -- at the same time that C2 believes the opposite. Hence, then examining Model 1's SCE can clarify conditions under which C2 can maintain prejudiced evaluations even though contradictory evidence exists. An advantage of SCE is that it allows us to represent citizens as starting with very different beliefs about race -- and as using new information about a black candidate in very different ways.[4]

[4] For other uses of the self-confirming equilibrium in political contexts see Lupia, Levine, and Zharinova (2010). Two additional characteristics about SCE are important to note.

3.D. Model 1 Results

I now present conclusions that are logical implications of the premises named above. An appendix contains proofs of conclusions whose logical relationship to the premises is not obvious.

Proposition 1 refers to the case where C2's private signal is completely informative.

Proposition 1. The unique SCE for the *completely informative* case entails both citizens favoring the most skilled candidate in every period.

Here, the citizens' knowledge of candidate skill levels makes race irrelevant in their evaluations.

The outcome changes when C2 knows less. Table 1 describes a SCE for each of the three kinds of private signal. For the two remaining private signals, *minimally informative* and *able to observe some contradictions*, I focus on the case $\Theta = .5$ even though the results I derive for this case can be derived in many other circumstances. I focus on this case because it is the one in which black and white skill levels are equal in expectation. This case is advantageous rhetorically, because it is the case where, if C2 knew everything, he would never benefit from evaluating candidates based on their race.[5] Moreover, I focus further on conditions under which a focal equilibrium can, and cannot, be sustained. In the focal equilibrium, C2 begins the game with an anti-black prejudice: C2 believes that$\Theta = 0$ (i.e., the black candidate is always less skilled). My analysis clarifies when new information changes C2's prejudice.

[TABLE 1 ABOUT HERE.]

First, SCE is a generalization of Nash Equilibrium rather than a refinement. Second, the SCE concept does not require that players use Bayes' Rule to process information. It requires only that actors' beliefs and conjectures, however drawn, are consistent with their observations. I choose to solve the model using SCE instead of Perfect Bayesian Equilibrium as it provides a ready means for describing goal-oriented information processing in the presence of a wide range of potential information-processing inefficiencies.

[5] This focus parallels that of Coate and Loury (1993). In a model designed to clarify affirmative action's effect on negative stereotypes, they focused on the case where identifiable groups are equally skilled. They proved that affirmative action did not eliminate the stereotypes in many such instances.

The following set of strategies, beliefs, and conjectures constitutes a SCE when C2's private signal is *minimally informative*.

Proposition 2. For the *minimally informative* case and Θ=.5, the following is a SCE: $\phi^*_{1n}(f_{1n}=1|\theta_n=1)=1$, $\phi^*_{1n}(f_{1n}=0|\theta_n=0)=1$, $\phi^*_{2n}(f_{2n}=0|\beta^*_{2n},\phi^*_{-2n})=1$,$\beta^*_{2n}(\theta_n=1|f_{1n},(\varnothing,\varnothing,\varnothing))=0$, $\phi^*_{-2n}(f_{1n}|\beta^*_{2n})=\sum_1^n \left|\frac{|f_{1n}=1|}{|n|}-.5\right|\rightarrow 0$, as $N\rightarrow\infty$.

In words, Proposition 2 says:

- C1 has a dominant strategy: Favor the black candidate when he is more skilled. Otherwise, do not
- C2 has a belief, conjecture, and a strategy that are mutually reinforcing. About skill, he believes $\Theta=0$ (blacks are inferior). About C1, he conjectures that C1's utility function is increasing in diversity (favoring the black and white candidate equally over time) rather than in actual candidate skill levels. Given his beliefs and conjectures, *the only evaluation strategy that C2 can rationalize is to always favor the white candidate, regardless of what C1 does.* In each period, and over *N* periods, C2's utility of .5 per period from favoring whites is always higher than the utility he expects to receive from ever favoring a black candidate in any period, *0*. Moreover, as $N\rightarrow\infty$, C2 observes C1 adding $(|\theta_n=1|)*N$ black applicants. Since $\Theta=.5$, this observation is not inconsistent with C2's false conjecture about Citizen 1's desire for diversity (i.e., as $N\rightarrow\infty$, $\sum_1^n \left|\frac{|f_{1n}=1|}{|n|}-.5\right|\rightarrow 0$). Hence, C2 continuously regards C1's evaluations as uninformative about candidates' true skill levels. Therefore, C2 can sustain his mistaken beliefs, false conjectures, and prejudicial evaluations indefinitely.

In this SCE, race is irrelevant to C1. So, if C1 were evaluating President Obama, then we can describe his strategy by saying "Obama is not stereotypically black" for the purpose of the evaluation. For C2, by contrast, race is all that matters. So, if C2 were evaluating President Obama, then we could describe C2's strategy by saying that for the purpose of evaluation, Obama remains stereotypically black.

What allows prejudice to persist in affecting C2's evaluations in the *minimally informative* case? The answer is his *persistent inability to relate his anti-black prejudice to its negative impact on his utility.* Since C2 incorrectly believes that C1 bases his evaluation on an interest in diversity rather than skill, C2's observation that C1 sometimes favors the black candidate does nothing to challenge C2's black inferiority belief. Hence, C2 is never directly confronted with evidence of his errors. This is why C2's prejudice persists. This situation is consistent with Hochschild's (2001) finding that despite ostensibly having access to the same historical information about blacks, many citizens are able to sustain beliefs about blacks that not only mistaken but politically consequential.

What happens we change what C2 can observe? Proposition 3 describes the consequence of moving from a *minimally informative* private signal to a private signal that can *reveal at least some contradictions* amongst C2's beliefs, conjectures, and observations.

Proposition 3. For the *able to observe some contradictions* case and Θ=.5, the SCE named in Proposition 2 cannot be sustained indefinitely as $N\rightarrow\infty$.

Here, C2 begins with the same beliefs and conjectures as in the minimally informative case (i.e., black inferiority). Now, however, there is now a non-zero probability that C2 will observe the black candidate's true skill level, θ_n, in a given period. Since the black candidate is less skilled than the white candidate in some periods, C2's ability to observe the candidates' true skill levels need not be initially sufficient to cause C2 to observe a contradiction between his racial beliefs and reality. He can maintain his initial beliefs and his prejudicial evaluation for as many consecutive periods, starting from period 1, that $\theta_n=0$ when he observes θ_n. But as $N\rightarrow\infty$, there will be a period where C2 observes $\theta_n=1$ and realizes that he is not always better off favoring the white candidate. This observation will contradict his black inferiority belief, $\Theta=0$. It will create the first moment in the game at which C2 realizes that his prejudice may harm his future utility (i.e., a material objective such as getting a certain policy passed or a non-material aspiration such as living in accordance with a particular moral or ethical standard). At this moment, his beliefs, conjectures and strategies are no longer in equilibrium. C2 can no longer rationalize his actions as he once did. Something has to give. I represent the mechanics of such moments as Model 2.

3.E. Model 2 Premises and Conclusions

Model 2 is a decision theoretic model that explains C2's therapeutic strategy. A therapeutic strategy converts attributes of C2's situation into a decision about how to react when he observes a contradiction to his prior belief. I represent such a process as strategic to reflect the idea that for a non-prejudiced response to follow a prejudiced response requires "intentional inhibition of the automatically activated stereotype and activation of the newer personal belief structure. In other words, prejudice is the result of an automatic process but can be controlled under certain conditions" (Devine 1989:5). Treating inhibition as a strategic decision to invest cognitive effort follows Lupia and Menning (2009).

I assume that C2 thinks through the issues specified in Model 2 if and only if he observes something in Model 1 that is inconsistent with his beliefs and conjectures. In other words, absent any motivation to question his evaluative strategy, C2 devotes no effort to doing so. This representation of belief change follows that of Holland et al (1986: 80), who state that: "triggering conditions are the failure of a prediction and the occurrence of some unusual event" (also see Leahey and Harris 2001).

Given such a trigger, C2 must decide whether to invest time and effort in attempting to update his racial beliefs. I denote this choice as $I \in \{0, 1\}$, where $I=1$ denotes a decision to pay cost $k>0$ to try to change his thinking (e.g., therapy -- self-administered or professional -- or taking the time to find relevant information and attempt to use it to change one's beliefs) and $I=0$ denotes a choice not to do so.

When C2 chooses $I=1$, he gains access to a completely informative private signal with probability $z \in [0,1]$, where z is exogenous. In other words, with this probability the consequence of therapy is an inhibitory connection that extinguishes the black inferiority belief and replaces it with the ability to observe the black candidate's true skill level. This sequence is analogous to Kurzban, Tooby, and Cosmides (2001) claim that race is a cognitive variable that can be "overwritten by new circumstances." My treatment differs from theirs, however, in that I do not assume that race is "easily overwritten." Instead, I treat the difficulty of such overwrites as a variable. With probability $1-z$, C2's inhibition attempt fails and he will again act on the basis of his initial belief, $\Theta=0$. So, a high value of z represents the circumstance described by Kurzban, Tooby, and Cosmides. Low values of z, by contrast, reflect the point of view of Kandel, et. al. (1995: 651-666) who argue that even for motivated people, information processing is characterized by severe constraints including the very limited storage capacity and high decay rates of working memory as well as the restrictive rules by which stimuli alter long-term memory (Schacter 2001).

I denote therapy's potential benefit as $x \in \Re$, which represents C2's (exogenously determined) belief about the expected utility of playing future periods of Model 1 as he would if his private signal were completely informative minus the expected utility of continuing to base his Model 1 evaluation strategy on his prejudice. Small values of *x* represent cases where C2 imagines little or no negative consequences from continuing his prejudice. Large values of *x* represent cases where C2's imagines substantial negative personal consequences (i.e., bad policy outcomes, not living in accordance with a desired more or ethical code, realizing that a mistaken belief about black intelligence may cause them to make errors in non-political domains, etc.) from continuing to evaluate blacks as he did before. This representation of belief change's antecedents follows from many empirical observations. As Pham, Cohen, Pracejus, and Hughes (2001: 170) explain, "the initial affective response will prompt subsequent thought generation through both automatic and controlled processes. The initial affective response can automatically cue affect-congruent materials in memory. In addition, knowledge may be actively recruited to more fully assess the affect-eliciting stimulus and to transform the initial affective response into a motivationally relevant response."

To complete the definition of this stage, I state a tie-breaking rule for cases in which investing in "therapy" and not doing so provide equal expected utility: C2 seeks to change his ways only if he expects a positive net benefit from doing so. If C2 believes that pursing or not pursuing therapy provide equal expected utility, then he chooses $I=0$.

Proposition 4 describes the only *therapeutic strategy* that C2 can rationalize in Model 2.

Proposition 4. C2 seeks therapy if and only if $x>k/z$.

Proof. The expected utility of an inhibition attempt is $zx-k$. By contrast, the expected utility of making no such attempt is 0. Therefore, C2 attempts inhibition only when $x>k/z$. *QED*.

Proposition 4 implies that new information about a black candidate will not reduce evaluative prejudice *unless a person associates their prejudice with a negative personal consequence*. In other words, if a person does not feel some kind of negative affect and recognize that their prejudice causes that pain, then they will have no incentive to reconsider their prejudice. So, if an observed contradiction in Model 1 leads C2 to realize that persisting in the belief $\Theta=0$ will lead him to substantial errors in the future (high *x*) and he believes that he can change his prejudice with a reasonable amount of effort, then he will make an effort to do so.

By contrast, if C2 believes that the contradiction he observed is a "one-time thing" and that as a general manner he can continue to evaluate blacks as he did before without suffering any negative consequences (low *x*), or C2 he believes that he is incapable of change (high *k* or low *z*), then he will not make the attempt. The "one-time" observation will be mentally stored as an "exception" or as a "subtype" and Citizen 2 will continue evaluating black candidates as he did before. In referring to C2's choice not to rethink his initial view of blacks as "subtyping," I follow Richards and Hewstone (2001:51), who say, "Subtyping occurs when perceivers respond to members of a target group who disconfirm their stereotypes by seeing them as exceptions to the rule and placing them in a separate subcategory apart from members who confirm the stereotype." I use "subtyping" to describe this outcome, because I assume that C2 cannot pretend that he did not see a contradiction. When C2 subtypes, he maintains his previous beliefs about blacks in general by inferring that the "contradictory" candidate was "not really black."[6]

3.F. Joint Logical Implications of Models 1 and 2

Figure 2 depicts the joint logical implications of Models 1 and 2 for prejudice change. The three rows pertain to C2's evaluative strategy as presented in Propositions 1-3. The two columns refer to the therapeutic strategy described in Proposition 4.

[6] Model 2 shares important similarities with the "self-regulation" model (Monteith and Mark 2005). It assumes (116) that "when people become aware that they have responded in prejudiced ways and such beliefs are inconsistent with their beliefs about how they should respond, negative affect is experienced." A difference between the two models is that I assume that such changes occur only if C2 believes that continuing his prejudice will reduce his future utility. If, instead, C2 believes that his prejudice is unrelated to his future objectives, then I assume that he experiences no negative affect.

Also, in their model (140), "behavioral inhibition in a situation where a prejudiced response may occur is necessary to disrupt an ongoing, automatic behavior and facilitate prospective reflection." They claim (134) that this system "will result in behavioral inhibition that allows one to engage in … a more careful consideration of how to respond so that biased responses can be avoided." In other words, if a discrepancy is discovered, a behavioral inhibition system ends biased responses. I assume that inhibition is attempted only if C2 finds it worthwhile to invest in therapy and that an attempt may not succeed. The modeling differences are consequential. They conclude (142) that "people can learn to bring relatively automatic reactions to members of stereotypes groups under control so as to respond without bias." My conclusion (Table 4) states additional requirements for such an outcome.

The top row represents one extreme: citizens know candidates' true abilities and their evaluations do not depend on race. For the purpose of evaluation, these people are color-blind. Continued contact with a black president does not alter prejudice's role in their evaluations.

The bottom row represents another extreme. These are people whose lives allow them never to realize negative personal consequences from prejudiced evaluations. They have no reason to question their beliefs. So, if they begin with a prejudice, it will persist – even if contradictory evidence exists.

The middle row represents the conditions for prejudicial change. Here, citizens may start out with a mistaken prejudice about race and skill. Because these they can observe contradictions, they can realize that their prejudice will cause them future pain. The question then becomes whether these citizens will put any effort into changing their minds.

On the left side of the middle row, C2 cannot bring himself to try to change his ways. He needs another way to reconcile the contradiction he observed. Hence, C2 concludes that the candidate is "not inferior in the way that I thought a black candidate would be." He subtypes the candidate. From C2's perspective, this candidate is no longer stereotypically black and, in the end, his initial prejudice remains available for evaluations of other blacks.

The right side of the middle row describes conditions for prejudice change. Here, C2's observation leads him to question whether his prejudice will cause him future pain. Moreover, he believes that "therapy" can help him update his beliefs and that his expected utility will rise as a result. This is the situation where contact can alter prejudice.

Now consider Table 4 as a whole. Think about how unlikely it is that some people will find themselves in the small part of the table where contact reduces evaluative prejudice. This view suggests that the conditions required for contact to reduce evaluative prejudice will not be easily satisfied for many citizens. Hence, a key implication of the theory can be viewed in contrast to Hajnal's (2007:3) conclusion from his study of black mayors. He finds that

> "Experience with black incumbents has real consequences for many members of the white community because it imparts critical information about black preferences that reduces whites' uncertainty and fear about black leadership; this information essentially changes the way that many white Americans think about

> the black community and therefore subtly alters the nature of racial politics and race relations in this country."

While this study agrees that such changes are possible, it disagrees that citizens necessarily attend to, or process, performance-based information about black incumbents in prejudice-reducing ways. Interactions amongst, and conditional relationships between, the quality of a person's information (i.e., their private signal), their motivation to process such information (which is increasing in *x*), and their beliefs about the likelihood that trying to change their ways will be worthwhile (low k, high z) are essential parts of the equation. Proceeding as if these conditions and interactions are irrelevant can yield incorrect conclusions about how new information affects prejudice's evaluative implications.

These conditions also explain differences in the empirical claims made above. For example, in Kurzban, Tooby, and Cosmides (2001) experiments, subjects have access to credible and informative information about race's irrelevance to their political objectives. Hence, they have set up a circumstance where contradictions are easy to observe and incentives to update beliefs are strong. Indeed, many claims of the form "contact reduces prejudice" are based (explicitly or implicitly) in the assumption that contradictions to prejudices are easily observed and trivially acted upon and with little or no attention to underlying causal mechanisms (see also Paluck and Green 2009). Sinclair and Kunda's subjects, by contrast, are never confronted with a situation that helps them see negative personal consequences following from their racially prejudiced responses. In such cases, and in cases where people have no reason to believe that they can change their ways or no reason to believe that they could avoid future personal negative consequences by doing so, we should expect prejudice to persist.

In sum, intergroup contact is far from sufficient to reduce prejudice's role in evaluations of his performance in this model. Instead, the following conditions are also necessary:

- A person must be in a situation where they can *associate their application of a prejudice with a negative personal consequence*. If they have no way to relate the negative consequence to the prejudice, then they have no incentive to consider rethinking the prejudice or its evaluative role.
- If a person associates a negative outcome with the prejudice, changing prejudice's role in subsequent evaluations does not occur automatically. The person must be *willing to rethink their beliefs* to make that change happen.

- Willingness to rethink beliefs, in turn, requires *sufficient motivation* and *expectation of efficacy*.
- *Sufficient motivation* refers to the idea that continuing to act on the prejudice will cause additional negative personal consequences in the future. If, by contrast, the person is convinced that the negative outcome they observed was a one-time-thing, their motivation to change is nil.
- The *expectation of efficacy* comes from the belief that an attempt to change a prejudice will mitigate future negative consequences and that the benefit of such mitigation outweighs the costs inherent in trying to rethink a prejudice and change one's actions accordingly. People who reside in contexts where prejudices can be expressed without negative consequences are less likely to expect efficacy than those who reside in contexts where such expressions are taboo.

Hence, the role of racial prejudice in evaluation is best understood as a joint product of psychological, contextual, and strategic factors. In other words, if having an African-American president leads people to realize that a previous belief about black competence is not only mistaken, but also personally costly (i.e., a citizen realizes that some of the decisions he makes causes outcomes that are bad for him), and if the same people are in a situation where they are willing and able to change their views, then the presence of a black president can reduce prejudice in subsequent evaluations. Without this combination of circumstances, prejudice will remain.

4. Discussion

As noted above, scholars are seeking more dynamic models of prejudice. The work in this article is a step in that direction. It clarifies dynamic, conditional, and simultaneous logical relationships between multiple continuous variables. In this closing section, I discuss several additional implications.

A key assumption in the model pertains to what citizens know about the sources of racial information that is available to them. When C2 is sufficiently knowledgeable about C1's motives, C2 can simply mimic C1's evaluations and never mistakenly evaluate less skilled candidates more favorably. What allows C2's prejudice to persist alongside contradictory evidence is C2's isolation from the information or C2 being in contexts where he has little incentive to think about contradictory observations.

I believe that this circumstance reflects an explanation for why racial prejudice persists in many evaluative contexts. Consider, for example, that the increasingly partisan televised media environment (e.g., the rise of Fox News as a conservative-leaning cable news outlet and the evolution of MSNBC as a cable news outlet that leans farther to the left) and the proliferation of narrowly-targeted political websites on the internet allow people to attend only to information that is offered by people who share their beliefs and values (Iyengar and Hahn 2009). The same changes in media also make it easier for people to ignore new information about black candidates. People who at one time might have been exposed to information about a black candidate by turning on the television and seeing that the only viewing option was the nightly news, now have many options for electronic communication that avoid such topics altogether (Prior 2007). These people have more discretion over what racial information they do – and do not -- observe than in the pre-cable era. At the same time continuing segregation in neighborhoods and at the workplace can also prevent people from being exposed to diverse views about race and politics even when they are not in front of a screen.

For such people, persistent prejudice in evaluations can be a self-confirming equilibrium. As Swain (2002:35) suggests, communicative networks of white racists are driven not so much underground, but to private communication channels where they such ideas can circulate without the filtering potential of political correctness.

> "There is real danger, I believe, when like-minded people get together and discuss only among themselves issues about which they care deeply that cannot be discussed in open forums. Such discussions are certain to lead to one sided and distorted conversations that in the context of race will inevitably enhance racial polarization and political extremism."

So many people appear to be in situations where the model can explain the persistence of anti-black prejudice even in the presence of what would otherwise seem to be contradictory evidence. In short, they never see the evidence or are in situations where they have no incentive to process it rigorously.

I continue this discussion by saying a few words about pro-black prejudice. The model's logic can also clarify Obama's effect on those who hold pro-black stereotypes. Consider, for example, citizens who find a special source of pride in the idea that a person with substantial African heritage leads the world's most powerful nation. There are households where the racial aspect of Obama's presidency is emphasized as a means to encourage young people, and others who have choices to make, that great things are possible. Consider, for example, the story told by Ashley Merriman (2009), who tutors minority children in Los Angeles:

> "[O]ne of my tutoring kids came bounding into Tutoring ─ he's a Hispanic seventh-grader who really struggles at home and school, "Ashley! Ashley! I have to tell you something!" he yelled. I'd never seen him this exuberant before.
>
> "OK, what?"
>
> "A couple days ago, there were these three boys at school. And they were all like, 'C'mon, let's ditch. Let's go do something fun. This is boring.'…So I tol' em, 'No,' and they said I was actin' white, and so I said, 'No, I'm actin' Obama' ─ just like you told me to….And they were all like, 'Well, OK, then, 'and left me alone. So I didn't ditch that day….And the other boys didn't ditch neither! We all stayed."
>
> "I am so proud of you! Of all four of you."

Continuing the logic derived above, people who begin with a pro-black stereotype may not rigorously process new information about Obama that counters their prejudice. Unless confronted with evidence that leads them to associate a negative personal outcome with their pro-black prejudice, and in a situation where investing in prejudice-adjusting therapy seems worthwhile, we should not expect their evaluative prejudice to change.

In sum, prejudiced persons may be exposed to positive information about black candidates that counter their initial stereotypes. If acute events force them to pay attention to this information, they will have to find some way to reconcile their observations with their prior beliefs. To handle the inconsistency, people can reevaluate their prejudice or they can maintain it by concluding that Barack Obama is not black in the sense of their initial stereotype. What they do at such moments is not only a function of their prior beliefs about race, but also of the context that affects what they believe about the costs and benefits of attempting to change their ways.

To achieve a future in which racial prejudice neither persists in the mind of the polity nor influences its actions – will require much more than one, or even a series of, black presidents. It also requires that people not only have a concrete means of associating their prejudice with a negative personal consequence but also the willingness and ability to process relevant information in prejudice-changing ways. Therefore, prejudice change is possible, but only for certain people in special circumstances.

APPENDIX. Technical Definition of Model 1 SCEs and Required Proofs

Let $\theta_n \in \{0,1\}$ be C1's type in period *n*, where $\theta_n=1$ denotes the type that knows of higher black skill levels in period *n* and $\theta_n=0$ knows the opposite. I assume that players know their own types.

Let $f_{in} \in \{0,1\}$ be Citizen *i*'s period *n* evaluation, where $f_{in}=1$ denotes Citizen *i* favoring the black candidate in period *n* and $f_{in}=0$ denotes the opposite. I denote C1's evaluation as $f_{1n}(\theta_n)$ to account for different strategies for C1's types. I denote C2's evaluation in a way that accounts for its possible dependence on C1's evaluation $f_{2n}(f_{1n})$. Let ϕ_{in} denote a mixed strategy for Citizen *i* in the set of possible actions for him. Then, ϕ_n is a strategy profile for period *n*.

Let $y_{2n} \in \{\emptyset,\Theta\} \times \{\emptyset,\theta_n\} \times \{\emptyset,U_{1n}\}$ be C2's period *n* private signal. Let $y_{2n}=(\Theta, \theta_{n,} U_{1n})$ be completely informative and $y_{2n}=(\emptyset,\emptyset,\emptyset)$ be minimally informative.

Let $\beta_i(\theta_n)$ represent Citizen *i*'s belief about black skill levels. By definition, $\beta_1(\theta_n)= \theta_n$. When $\beta_1(\theta_n)=\beta_2(\theta_n)$, the citizens have shared beliefs about blacks. I use $\beta_2(\theta_n)=\beta_2(\theta_n| f_{1n}, y_{2n})$ to denote the possible dependency of C2's beliefs about black skill levels on his observation of C1's evaluation and the private signal. Let Θ represent the true probability that $\theta_n=1$ in any period *n*.

Let $\phi_{-in} \in [0,1]$ denote Citizen i's conjecture about the other citizen's period n strategy.

Finally, let $\phi_{1n}(f_{1n})$ denote C1's conditional probability of taking action f_{1n} upon observing θ_n in the mixed strategy ϕ_{1n} and let $\phi_{2n}(f_{2n}|\beta_{2n}(\theta_n| f_{1n}, y_{2n}), \phi_{-2n}(f_{1n}|\beta_{2n}))$ denote C2's conditional probability of taking action f_{2n} upon observing f_{1n} and y_{2n} and then forming belief β_{2n} and conjecture ϕ_{-2n} in the mixed strategy ϕ_{2n}.

Definition. An evaluation strategy profile ϕ^*_n is a SCE with conjectures ϕ^*_{-in} and beliefs β^*_{in}, if the following conditions are satisfied:

- For any θ_n, $U_{1n}(\phi^*_{1n}, \theta_n) \geq U_{1n}(\phi_{1n}, \theta_n)$, $\forall \phi^*_{1n} \neq \phi_{1n}$
- For any θ_n and f^*_{2n} such that $\beta^*_{2n}(\theta_n| f_{1n}, y_{2n}) \cdot \phi^*_{-2n}(f^*_{2n}|\beta^*_{2n}(\theta_n| f_{1n}, y_{2n}))>0$, $U_{2n}(\phi^*_{2n},\theta_n)\beta^*_{2n}(\theta_n| f_{1n}, y_{2n})\phi^*_{-2n}(f_{1n}|\beta^*_{2n}) \geq U_{2n}(\phi_{2n},\theta_n)\beta^*_{2n}(\theta_n| f_{1n}, y_{2n})\phi^*_{-2n}(f_{1n}|\beta^*_{2n})$, $\forall \phi^*_{2n} \neq \phi_{2n}$
- For every possible value of y'_{2n} of y_{2n}, $\Sigma\{f_{1n},\theta_n\text{: } y'_{2n}\}\ \beta^*_{2n}(\theta_n| f_{1n}, y_{2n})\phi^*_{-2n}(f_{1n}|\beta^*_{2n})= \Sigma\{f_{1n},\theta_n\text{: } y'_{2n}\}\ ((1-\Theta)\phi^*_{1n}(f_{1n}|\theta_n=0)) + (\Theta\phi^*_{1n}(f_{1n}|\theta_n=1))$

Proposition 1. The unique SCE for the *completely informative* private signal case is, for $i \in \{1,2\}$,

- $U_{in}(\phi^*_{in}(f_{in}=1|\theta_n=1)=1) = 1 > \phi_{in}(f_{in}=1|\theta_n=1) = U_{in}(\phi_{in}(f_{in}=1|\theta_n=1))$, $\forall \phi_{in}(f_{in}=1|\theta_n=1)<1$
- $U_{in}(\phi^*_{in}(f_{in}=0|\theta_n=0)=1) = .5 > .5\phi_{in}(f_{1n}=0|\theta_n=0) = U_{in}(\phi_{in}(f_{1n}=0|\theta_n=0)<1)$, $\forall \phi_{in}(f_{in}=0|\theta_n=0)<1$

The validity of this claim follows directly from the SCE definition. This solution is also the unique Nash Equilibrium for this case.

Proposition 2. For the *minimally informative* private signal case and $\Theta=.5$, the following is a SCE: $\phi^*_{1n}(f_{1n}=1|\theta_n=1)=1$, $\phi^*_{1n}(f_{1n}=0|\theta_n=0)=1$, $\phi^*_{2n}(f_{2n}=0|\beta^*_{2n},\phi^*_{-2n})=1$, $\beta^*_{2n}(\theta_n=1|\, f_{1n}, (\varnothing,\varnothing,\varnothing))=0$, $\phi^*_{-2n}(f_{1n}|\beta^*_{2n})= \sum_{1}^{n} |\frac{|f_{1n}=1|}{|n|} - .5| \to 0$, as $N \to \infty$.

Proof: To show that these strategies, beliefs, and conjectures are a SCE, it is sufficient to demonstrate that

- $U_{1n}(\phi^*_{1n}(f_{1n}=1|\theta_n=1)=1) = 1 > \phi_{1n}(f_{1n}=1|\theta_n=1) = U_{1n}(\phi_{1n}(f_{1n}=1|\theta_n=1))$, $\forall \phi_{1n}(f_{1n}=1|\theta_n=1)<1$
- $U_{1n}(\phi^*_{1n}(f_{1n}=0|\theta_n=0)=1) = .5 > .5\phi_{1n}(f_{1n}=0|\theta_n=0) = U_{1n}(\phi_{1n}(f_{1n}=0|\theta_n=0)<1)$, $\forall \phi_{1n}(f_{1n}=0|\theta_n=0)<1$
- $U_{2n}(\phi^*_{2n}(f_{2n}=0)=1)=.5 \geq \phi_{2n}U_{2n}(f_{2n}=0) + (1-\phi_{2n})U_{2n}(f_{2n}=1|)\beta^*_{2n}(\theta_n=1|\, f_{1n}, y_{2n})\phi^*_{-2n}(f_{1n}|\beta^*_{2n})$, $\forall \phi^*_{-2n}(f_{1n}|\beta^*_{2n})$, and all $\phi_{2n}(f_{2n}=0)<1$, where $\beta^*_{2n}(\theta_n=1)=0$

The first two inequalities follow from C1's complete information. It remains to validate the third inequality. The private signal in this case is $y_{2n}=(\varnothing,\varnothing,\varnothing)$. Since $\Theta=.5$ and C2 conjectures about C1

$\phi^*_{-2n}(f_{1n}|\beta^*_{2n})= \sum_{1}^{n} |\frac{|f_{1n}=1|}{|n|} - .5| \to 0$, then as $N \to \infty$, C1’s choices will not contradict the conjecture even in the limit. Since the private signal is uninformative, C2 will never observe the personal consequence of favoring a highly skilled black. Hence, C2’s observations are insufficient to contradict his conjecture or belief and the third inequality holds for all *n*. *QED*.

Proposition 3. For the *able to observe some contradictions* private signal and $\Theta=.5$, the SCE named in Proposition 2 cannot be sustained indefinitely as $N \to \infty$.

Proof. The difference between the focal SCE for the minimally informative private signal and the focal SCE for this case is that the signal $y_{2n}=(\varnothing,\theta_n=1,\varnothing)$ contradicts $\beta^*_{2n}(\theta_n=1)=0$. This contradiction implies that "$U_{2n}(\phi^*_{2n}\ (f_{2n}=0)=1)=.5\geq \phi_{2n}U_{2n}(\ f_{2n}=0) +(1-\phi_{2n})U_{2n}(f_{2n}=1)\beta^*_{2n}(\theta_n=1|\ f_{1n},\ y_{2n})\phi^*_{-2n}(f_{1n}|\beta^*_{2n}),\ \forall\ \phi^*_{-2n}(f_{1n}|\beta^*_{2n})$, *for all* $\phi_{2n}(f_{2n}=0)<1$" is false. In other words, C2 realizes that there exists a mixed strategy in which sometimes favoring a black candidate provides greater expected utility than the strategy of always favoring the white candidate. After such a revelation, the Proposition 2 strategy-belief-conjecture triple for C2 cannot be sustained as a SCE. *QED*.

References

Allport, Gordon W. 1954. *The Nature of Prejudice*. Reading MA: Addison-Wesley.

Bargh, John A. 1999. "The Cognitive Monster: The Case Against the Controllability of Automatic Stereotype Effects." In Shelly Chaiken and Yaacov Trope (eds.) *Dual-Process Theories in Social Psychology,* pp. 361-382. New York: Guilford Press.

Coate, Stephen, and Glenn C. Loury. 1993. "Will Affirmative Action Policies Eliminate Negative Stereotypes." *The American Economic Review* 83: 1220-1440.

Churchland, Patricia S., and Terrence J. Sejnowski. 1992. *The Computational Brain*. Cambridge MA: MIT Press.

Devine, Patricia G. 1989. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology* 56:5-18.

Fiedler, Klaus. 1996. "Processing Social Information." In Miles Hewstone, Wolfgang Stroebe, Geoffrey Stephenson (eds.) *Introduction to Social Psychology*. Oxford UK: Blackwell Publishers Ltd.

Fiske, Susan T. 1998. "Stereotyping, Prejudice, and Discrimination." In Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey (eds.) *The Handbook of Social Psychology*, 4th ed. Boston: McGraw-Hill , pp. 357-411.

Fiske, Susan T., and Shelley E. Taylor. 1984. *Social Cognition*. New York: McGraw-Hill.

Fudenberg, Drew, and David K. Levine. 1993. "Self-Confirming Equilibrium." *Econometrica* 61: 523-545.

Fudenberg, Drew, and David K. Levine. 1998. *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Gaertner, Samuel L., and John F. Dovidio. 1986. "The Aversive Form of Racism." In John F. Dovidio and Samuel L. Gaertner (eds.) *Prejudice, Discrimination, and Racism*, pp. 61-89. San Diego: Academic Press.

Hajnal, Zoltan L. 2007. *Changing Attitudes toward Black Political Leadership*. New York: Cambridge University Press.

Hewstone, Miles. 1994. "Revision and Change of Stereotypic Beliefs: In Search of the Elusive Subtyping Model." In Wolfgang Stroebe and Miles Hewstone (eds.), *European Review of Social Psychology*, *Volume 5*: 69-109. Chichester, England: Wiley.

Hochschild, Jennifer. 2001. "Where You Should Stand Depends on What You See: Connections Among Values, Perceptions of Fact, and Political Prescriptions." In James H. Kuklinski (ed.), *Citizens and Politics: Perspectives from Political Psychology*. New York: Cambridge University Press.

Iyengar, Shanto, and Kyu H. Hahn. 2009. “Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use.” *Journal of Communication* 59: 19-39.

Kawakami, Kerry, Elizabeth Dunn, Francine Karmali, and John F. Dovidio. 2009. "Mispredicting Affective and Behavioral Responses to Racism." *Science* 323: 276-278.

Kandel, Eric R., James H. Schwartz, and Thomas M. Jessell. 1995. *Essentials of Neural Science and Behavior*. Norwalk, CT: Appleton and Lange.

Kurzban, Robert, John Tooby, and Leda Cosmides. 2001. “Can Race Be Erased? Coalitional Computation and Social Categorization.” *Proceedings of the National Academy of Science* 98:15387-15392.

Leahey, Thomas H. and Richard J. Harris. 2001. *Learning and Cognition*. Upper Saddle River, NJ: Prentice-Hall.

Lupia, Arthur, and Jesse O. Menning. 2009. “When Can Politicians Scare Citizens Into Making Bad Choices?” *American Journal of Political Science* 53: 90-106.

Lupia, Arthur, Adam Seth Levine, and Natasha Zharinova. 2010. “Should Political Scientists Use the Self-Confirming Equilibrium Concept? Benefits, Costs and an Application to Jury Theorems.” *Political Analysis* 18: 103-123.

Merryman, Ashley. 2009. "Tell Them You're Acting Obama (A Personal Essay)." http://blog.newsweek.com/blogs/nurtureshock/archive/2009/09/08/tell-them-you_2700_re-acting-obama-_2800_a-personal-essay_2900_.aspx. Posted on September 8, 2009. Downloaded 11/19/2009.

Montieth, Margo J., and Aimee Y. Mark. 2005. “Changing One’s Prejudiced Ways: Awareness, Affect, and Self-Regulation.” *European Review of Social Psychology* 16: 113-154.

Paluck, Betsy Levy, and Donald P. Green. 2009. "Prejudice Reduction: What Works? A Review of Research and Practice." *Annual Review of Psychology* 60: 339-367.

Pasek, Josh, Alexander Tahk, Yphtach Lelkes, Jon A. Krosnick, B. Keith Payne, Omair Akhtar and Trevor Tompson. 2009. “Determinants of Turnout and Candidate Choice in the 2008 Presidential Election Illuminating the Impact of Racial Prejudice and Other Considerations.” *Public Opinion Quarterly* 73: 943–994

Pettigrew, Thomas F. 2008. “Future Directions for Intergroup Contact Theory and Research.” *International Journal of Intercultural Relations* 32: 187-199.

Pettigrew, Thomas F., and Linda R. Tropp. 2006. “A Meta-Analytic Test of Intergroup Contact Theory.” *Journal of Personality and Social Psychology* 90: 751-783.

Pham, Michel Tuan, Joel B. Cohen, John W. Pracejus, G. David Hughes. 2001. "Affect Monitoring and the Primacy of Feelings in Judgment." *The Journal of Consumer Research* 28: 167-188.

Piston, Spencer. 2010. "How Explicit Racial Prejudice Hurt Obama in the 2008 Election." *Political Behavior* 32: 431-451.

Plant, E. Ashby, Patricia G. Devine, William T.L. Cox, Corey Columb, Saul L. Miller, Joanna Goplen, and B. Michelle Peruche. 2009. "The Obama Effect: Decreasing Implicit Prejudice and Stereotyping." *Journal of Experimental Social Psychology* 45: 961-964.

Prior, Markus. 2007. *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections*. New York: Cambridge University Press.

Richards, Zoëe, and Miles Hewstone. 2001. "Subtyping and Subgrouping: Processes for the Prevention and Promotion of Stereotype Change." *Personality and Social Psychology Review* 5: 52-73.

Schacter, Daniel L. 2001. *The Seven Sins of Memory: How the Mind Forgets and Remembers.* Boston: Houghton-Mifflin.

Sinclair, Lisa, and Ziva Kunda. 1999. "Reactions to a Black Professional: Motivated Inhibition and Activation of Conlficting Stereotypes." *Journal of Personality and Social Psychology* 77: 885-904.

Swain, Carol M. 2002. *The New White Nationalism in America: Its Challenge to Integration*. New York: Cambridge University Press.

Weber, Renee, and Jennifer Crocker. 1983. "Cognitive Processes in the Revision of Stereotypic Beliefs." *Journal of Personality and Social Psychology* 45: 961-977.

Figure 1. Extensive Form Representation of a Single Period in Model 1

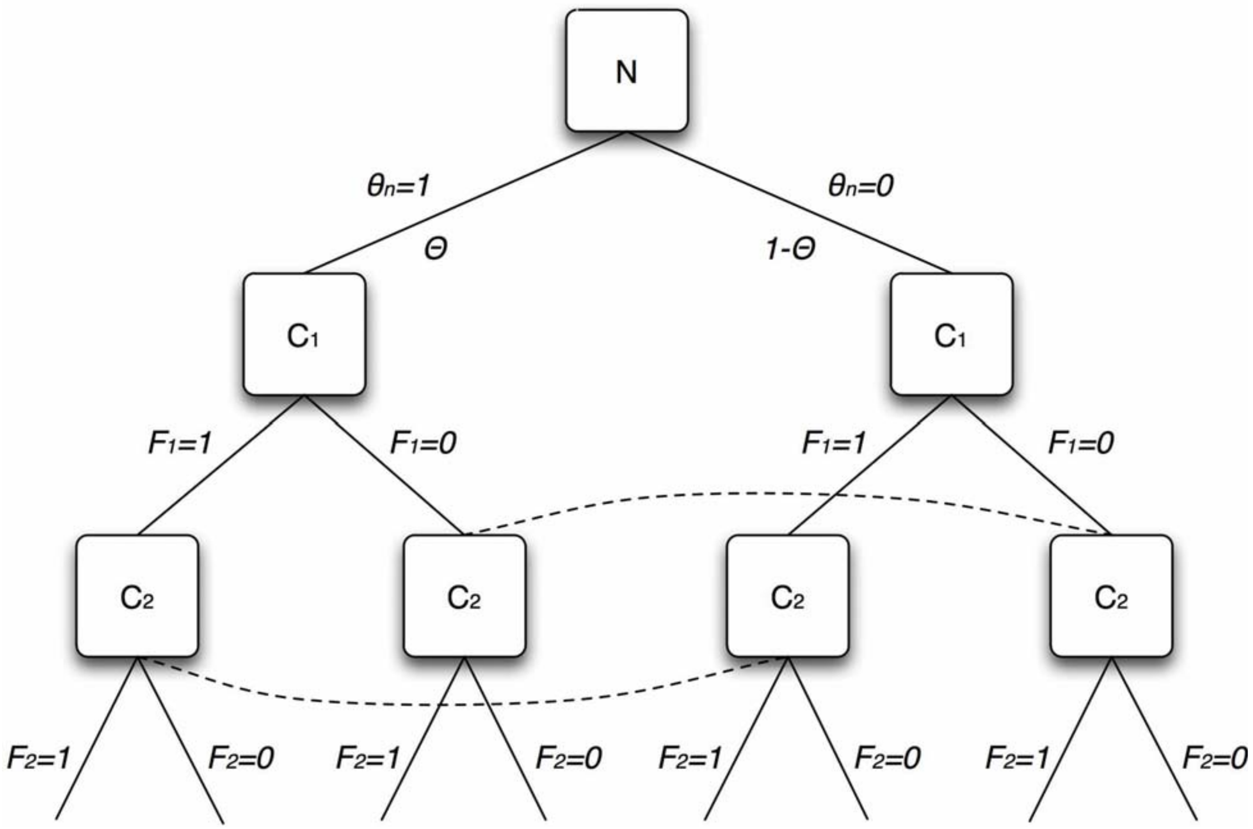

Table 1. Characteristics of Focal Equilibria for Model 1 in Period *n*.

Citizen 2's Private Signal	Completely Informative	Minimally Informative	Able to Observe Some Contradictions
Claim	Proposition 1: uniqueness	Proposition 2: sufficient for prejudice persistence	Proposition 3: sufficient for prejudice change
Domain	$\Theta \in [0,1]$	$\Theta = .5$	$\Theta = .5$
Citizen 1's strategy	$f_{1n} = \theta_n,\ \forall n \in N$	$f_{1n} = \theta_n,\ \forall n \in N$	$f_{1n} = \theta_n,\ \forall n \in N$
Citizen 2's strategy	$f_{2n} = f_{1n},\ \forall n \in N$	$f_{2n} = 0,\ \forall n \in N$	$f_{2n} = 0,\ \forall n \in N$ this strategy ceases to be viable after a period in which $\theta_n = 1$ is revealed
Citizen 2's initial belief about the black candidate	θ_n	$\theta_n = 0$, because $\Theta = 0$*	$\theta_n = 0$, because $\Theta = 0$*
Citizen 2's initial conjecture about Citizen 1's strategy	Utility maximization	*Diversity maximization*:* As $N \rightarrow \infty$, $\sum_{1}^{n} \left\| \frac{\|f_{1n} = 1\|}{\|n\|} - .5 \right\| \rightarrow 0$	*Diversity maximization*:* As $N \rightarrow \infty$, $\sum_{1}^{n} \left\| \frac{\|f_{1n} = 1\|}{\|n\|} - .5 \right\| \rightarrow 0$

***=a false belief or conjecture**

Figure 2. Depiction of Prejudice's Role in Evaluations for Three Focal Cases

<table>
<tr><th></th><th>Citizen 2 believes that inhibition is likely to fail or is not worth trying $x \leq k/z$</th><th>Citizen 2 believes that inhibition is likely to succeed and is worth trying $x > k/z$</th><th></th></tr>
<tr><td>Citizen 2's private signal is *completely informative*</td><td colspan="2">No change:
Evaluation independent of prejudice</td><td rowspan="3">⇐MODEL 1: Evaluation⇒</td></tr>
<tr><td>Citizen 2 is *able to observe some contradictions*</td><td>Subtyping occurs. Initial prejudice otherwise preserved</td><td>**Change is possible**</td></tr>
<tr><td>Citizen 2's private signal is *minimally informative*</td><td colspan="2">No change:
Initial prejudice persists</td></tr>
</table>

Advance Access publication October 6, 2009 *Political Analysis* (2010) 18:103–123
doi:10.1093/pan/mpp026

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.

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) conjectives. 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.[1]

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 $u_i(s_1, \ldots, s_n)$ 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 $G = \{S_1, \ldots S_n; u_1, \ldots u_n\}$, the strategies $(s_1^*, \ldots s_n^*)$ 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, $(s_1^*, \ldots s_{i-1}^*, s_{i+1}^*, \ldots s_n^*)$: $u_i(s_1^*, \ldots s_{i-1}^*, s_i^*, s_{i+1}^*, \ldots s_n^*) \geqslant u_i(s_1^*, \ldots s_{i-1}^*, s_i, s_{i+1}^*, \ldots s_n^*)$ for every feasible strategy s_i in S_i; that is, s_i^* solves $max_{s_i \in S_i} u_i\left(s_1^*, \ldots s_{i-1}^*, s_i, s_{i+1}^*, \ldots s_n^*\right)$.

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 σ_j about j; and the resulting profile $(\sigma_1, \ldots, \sigma_n)$ 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.[2] One implication of

[1] 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.

[2] 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 θ_{-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 σ_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.[3]

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

[3] 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 θ_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.[4]

Definition: A strategy profile σ is a SCE with conjectures $\hat{\sigma}_{-i}$ and beliefs $\hat{\mu}_{\mathrm{i}}$ if for each player i, (1) $\forall\theta_i$, $r(\theta_i) = \hat{\mu}_i(\theta_i)$, and for any pair θ_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:

$$\sum\nolimits_{\{a_{-i},\theta_{-i}:\, y_i(\hat{a}_i,a_{-i},\theta_i,\theta_{-i}) = \bar{y}_i\}} \hat{\mu}_i(\theta_{-i}|\theta_i)\hat{\sigma}_{-i}(a_{-i}|\theta_{-i})$$
$$= \sum\nolimits_{\{a_{-i},\theta_{-i}:y_i(\hat{a}_i,a_{-i},\theta_i,\theta_{-i}) = \bar{y}_i\}} r(\theta_{-i}|\theta_i)\sigma_{-i}(a_{-i}|\theta_{-i}).$$

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

[4] 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.[5]

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

[5] 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 that 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

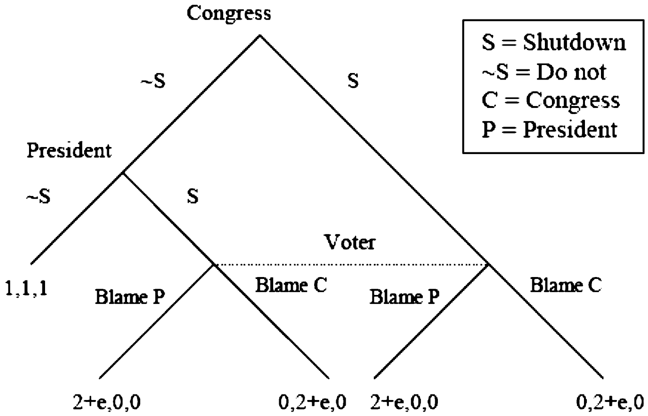

Fig. 1 Congress-president standoff.

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).[6] 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

[6] 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.[7]

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

[7] 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 $N = \{1, 2, ..., n\}$ 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 Ω with probability $p \in (.5, 1)$ and the false value of Ω 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

	Low NFC	*High NFC*
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 σ_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."[8]

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.

[8] 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.[9] *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

[9] 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.

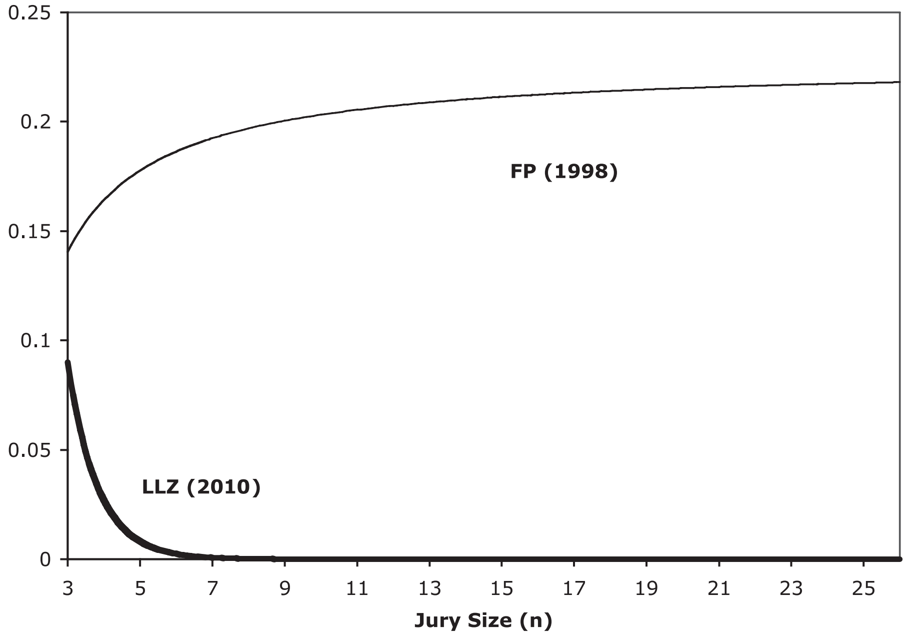

Fig. 2 The probability that an innocent defendant is convicted as a function of jury size for $p = .7$ and $q = .5$.

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.[10] We contend that such questions are, at least in part, empirical.

[10] 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 temporaility 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 probability des decisions rendues a la plurality des voix*. Paris: Translation by lain 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> 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.