ARE JURIES MORE CONSISTENT? COMPARING INDIVIDUAL AND GROUP DECISION-MAKING VIA ONLINE DELIBERATION

Xinlan Emily Hu

Stanford University
May 2020

An honors thesis submitted to the department of
Computer Science
in partial fulfillment of the requirements for the undergraduate
honors program

Advisor: Michael S. Bernstein

_____ Date: _____

Michael S. Bernstein (Thesis Advisor)
Associate Professor
Computer Science

_____ Date: _____

Mark E. Whiting (Thesis Advisor)
Postdoctoral Scholar
Computer Science (University of Pennsylvania)

Abstract

Democratic governance relies upon the assumption that group deliberation achieves more than individuals deciding alone, taking the form of governing boards, committees, panels, and task forces. Even in online communities, recent movements have shifted power from the lone content moderator to a deliberating collective. However, a special form of deliberating group stands out: juries. There, 6–12 average individuals can determine the difference between one's life and death.

Jury decisions are only legitimate if they are also consistent: that is, in a parallel universe, would the same jury have come to the same conclusion? If juries are inconsistent in this manner, it suggests that the adjudication is influenced by arbitrary social factors within the discussion, rather than being grounded in the facts as presented.

Yet it is unclear whether groups outperform individuals on this metric. On the one hand, social influence has been found to increase the accuracy of beliefs; on the other, groups are vulnerable to information signals and reputation pressures, amplifying errors and inconsistency. Thus, group decisions may be far less consistent than an individual deciding alone.

However, consistency in groups has never been directly measured, since the same group cannot convene again without reactivating prior social context. Further, even if the same group reconvenes, its prior experience on a case would likely influence its subsequent decisions, thereby creating a learning effect.

This thesis directly compares the consistency of group and individual decision-makers. To do this, we draw upon the affordances of a pseudonymous online deliberation platform. Using one-way pseudonym masking, we manipulate the perceived identities of one's past collaborators, enabling each repeated group deliberation to begin anew. Reconvened groups adjudicate paired cases that are known to have aligned outcomes. We then use this system to compare the decision consistency of juries to that of individual decision-makers.

Ultimately, we find that groups and individuals are equally consistent; participating in a group also does not affect an individual's own decision consistency. We also find that minority voices are more influential in deliberation than previously expected. These results are especially interesting in light of the fact that participants greatly underestimated the consistency of the teams they participated in. Jury decisions are consistent despite a widespread perception to the contrary.

Acknowledgements

When I arrived at Stanford four years ago, I never thought that I would major in Computer Science, much less write a thesis in this subject. Professor Michael Bernstein, you changed all of that. Thank you for believing in me — from the moment I first joined your lab as a freshman, doe-eyed and inexperienced. You guided me through the trials and tribulations of research; through two different projects; many, many, iterations; tons of growth. Thank you for making me feel at home as an undergraduate in the lab. Thank you for advocating for me constantly; thank you for teaching me almost everything that I know about HCI. I'll never forget the importance of velocity, and that prototypes are questions.

To Dr. Mark Whiting, thank you for being such a supportive research mentor. I could have asked for no one better. You're a champion of remote work (we worked across 2 countries! 4 states!), and you are effective in a way that I can only aspire to be. Thank you for the calls that helped me sort out the research questions, debug *Bang* again and again, and piece a story from the data. Thank you for tolerating me every time I asked a silly question. I am so grateful for you.

Thank you to my family. To my grandpa, *Ya-ya*, this one is for you. You were the first engineer I had ever met, the first one to inspire me. And in a way, this thesis represents you, too: you always said that family decisions should be made by deliberation and consensus; now, we'll see why.

To *Niang-niang, Ye-ye,* and *Nai-nai*, thank you for your endless love. To Mom and Dad, thank you for supporting me in every endeavor I try. I'm so grateful for your inspiration, your willingness to read bad drafts, your home-cooked meals (a benefit of writing this thesis at home, during COVID-19), and your deep love. I promise, Mom and Dad, I do sleep.

To Michael Cooper, my boyfriend. Thank you for being the one I'd Rubber Duck to whenever I got stuck; for getting me tea; for supportive phone calls. Thank you for (nearly) three years of love, for encouraging each other even as our interests co-evolve.

To my best friends and this quarter's project partners: Logan Pearce, Angela Luo, Matthew Trost. Thank you for cheering me on. For picking up some slack I dropped on projects during the busiest moments of writing this thesis. For believing in me and sending me gifs and encouraging messages. I'm so grateful to have each of you in my life.

To Anjini Karthik, thank you for the "thesis parties." For lifting my spirits and — ever since

Oxford — writing with me till daybreak. Thank you for being such a wonderful friend.

To the people who have inspired me to work on (and love doing) research: Jenny Han, Allison Chi, Gobi Dasu, Ali Alkhatib, my first research team. I still have the chocolate you gave me after that first summer of CURIS. I never ate it because the messages you wrote on the packaging were so precious to me. Phoebe Yao, thank you for co-working with me, co-debugging with me, and for keeping me accountable. Thank you for teaching me your entrepreneurial spirit. Tonya Nguyen, thank you for becoming such a wonderful research teammate and friend, even though we've never met in person. To the Digital Civil Society Lab — Dr. Lucy Bernholz, Dr. Argyri Panezi, Dr. Toussaint Nothias, Dr. Jonathan Pace, and Dr. Cadence Wilse — thank you for encouraging me to pursue research. Working with you has been an immense honor and privilege, and I have learned so much. I'm so glad I took COMM 230.

Above all, glory to God. I am so grateful. The above list is only a fraction of all the wonderful people at Stanford and beyond who have encouraged me, influenced me, and inspired me.

Contents

Chapter 1

Introduction

> *There should be little variance in verdicts for a single case, in the hypothetical situation where the same case might be tried repeatedly by similar juries.*
>
> (Hastie, Penrod, & Pennington, 1983)

In 2019, just two weeks apart, the Supreme Court heard two cases. The first involved a Muslim man on death row, who requested that an imam be present in the execution chamber. The second involved a Buddhist death row inmate, who requested the presence of a Buddhist minister.

The court ruled that the Muslim man's execution could proceed, but the Buddhist man's could not (Totenberg, 2019).

Legal outrage ensued; how could the court provide such starkly different judgments on two near-identical cases?

As it turns out, judgment disparities like these are far from uncommon. A similar sort of indignation erupts when African-Americans are handed harsher sentences than Whites for the same crime (Glater, 2007), or when men are given longer sentences than women (Starr, 2015).

The notion that analogous cases are treated unequally stirs a sense of deep unease: it seems to violate an unspoken standard of justice. To be just requires that similar cases have similar outcomes — that verdicts be consistent. Yet there seems to be no manner of knowing precisely how consistent the justice system is.

This thesis capitalizes on the affordances of an online jury to turn Hastie et al.'s (1983) thought experiment — a "hypothetical situation in which the same case might be tried repeatedly by similar juries" — into reality. In doing so, this work positions itself as a commentary on and evaluation of governance practices.

More generally, democratic governance relies upon the assumption that disputes are best resolved via deliberation, one that juries within the United States exemplify, and that our online jury system seeks to recreate. Over one million Americans serve on juries each year (Burghardt, Rand, & Girvan,

2019), deciding the fates of 154,000 cases. Nearly a third of Americans have participated in jury duty (Margolis, Huckaby, Odessey, & Hug, 2009).

Thus, much more than a mere procedural token, the jury is also a source of civic education and democratic legitimacy. The act of granting decision power to a panel of ordinary citizens softens rigid laws with accepted community values (Flango, 2016). A trial by jury is considered so vital that it is enshrined as a right *thrice*: first in the Declaration of Independence (1776), again in the Constitution (1787), and finally in the 6th Amendment (1791) (Alschuler & Deiss, 1994).

Little wonder that scholars have described the deliberation of juries as "more than just talk...[it] is a special form of speech structured according to democratic principles and designed to transform private prejudice into considered public opinion and to produce more legitimate solutions" (Noveck, 2004). Furthermore, as citizen jurors evaluate cases, the process serves a sort of participatory classroom. Tocqueville famously believed that jurors would "learn to think more judiciously" by imagining themselves in the shoes of defendants (Marder, 2005).

In addition to playing a key role in our justice system, juries stand poised to play a rising role in online governance. Digital courts have already begun to adjudicate private disputes (Ast, 2017), and juries have been shown to lend legitimacy to resolving content moderation challenges (Fan & Zhang, 2020).

Juries are, in fact, a special case of using group collaboration to solve complex problems. However, group collaboration is by no means a silver bullet, nor is it a necessarily positive force. Decades of social psychology research on group behavior point to key vulnerabilities: compared to individuals, groups are more prone to information signals and reputation pressures that amplify errors (Sunstein, Hastie, et al., 2014) and cause opinions to become polarized (Sunstein, 2000). There is perhaps nothing more special about deliberation than a tyranny of the majority: when asked to prescribe a solution to an ethical dilemma, groups overwhelmingly settle on the majority opinion (Thorne, Massey, & Jones, 2004) and ignore information that only a few members know about, even when it is important (Stasser & Titus, 1985). Indeed, the size of the majority faction is among the strongest predictors of a jury's final verdict (Hastie et al., 1983). In a far cry from dramatizations such as *Twelve Angry Men*, the lone moral voice rarely speaks up when faced with overwhelming opposition. Instead, social influence can cause jurors to vote against their predispositions, and to punish defendants more harshly than they would have if voting alone (Son, Bhandari, & FeldmanHall, 2019).

As a result, there are as many reasons to distrust groups as there are to depend on them. Group deliberation forms a cornerstone of democracy, yet their decisions could be even more arbitrary than an individual deciding alone.

A jury's verdict can be measured by two core metrics: *accuracy* and *consistency*. Most prior work has focused on the former, both for juries and for groups as a whole. Several influential studies of group work are designed around an estimation task (for example, of the number of jellybeans inside

of a jar), and judged by an accuracy metric. Examples include the "wisdom of crowds" (Surowiecki, 2005; Becker, Guilbeault, & Smith, 2019) and the "survival task," a scenario in which the group's answers to a hypothetical scenario are compared to the "ground truth" responses of expert survivalists (Miner Jr, 1984). Even in mathematical models of jury design, a key assumption is that each vote has some probability of being correct (Kaniovski & Zaigraev, 2011).

However, the literature has woefully under-explored the presumption of consistency: that the same jury, judging two similar cases, should lead to the same outcome. A crucial premise of making decisions via juries requires that the jury's decisions not be random — that they be determined more by the case as presented than by arbitrary factors of the group's discussion. Consistency underlies the sense of outrage and protest when seemingly analogous crimes receive disparate sentences; a justice system perceived as inconsistent contributes to heightened tensions, institutional distrust, and an unwillingness to cooperate between groups (van Prooijen, Gallucci, & Toeset, 2008). Because of consistency's key role in establishing the legitimacy of institutions, our core research question asks: *are group decisions more consistent than individual decisions in a jury task?*

Thus far, exploration of consistency has been unable to do better than comparing a group's predispositions to its final verdict (Roper, 1980). However, this methodology is only an approximation, at best, for the true question of interest: that the same jury's performance on two similar cases should lead to the same outcome. This sort of repeated deliberation is not possible with in-person juries. As the jury deliberates, they build attributions of each others' opinions that would simply be reactivated if the jury were to re-convene, such that deliberation would not truly begin from scratch the second time.

While traditional juries cannot easily forget social context, virtual juries *are* able to do so. This thesis draws on the unique affordances of internet-based collaboration tools to directly investigate group consistency. We implemented a platform that convenes juries online. The system temporarily masks juries' interaction history by assigning new pseudonyms to each member, while holding members' own pseudonyms constant in their own view. As a result, the same jury can deliberate repeatedly without realizing that group members have remained unchanged. Prior work has shown that this process of one-way pseudonym masking leads online groups to effectively begin deliberation from scratch (Whiting, Blaising, et al., 2019; Whiting et al., 2020), thus creating "parallel worlds" with which we can evaluate verdict consistency. We confirmed that this manipulation was effective: after removing teams in which the number of participants who failed the manipulation check was more than two standard deviations (1.098) above the mean (1.014), the accuracy rate of remaining participants was statistically indistinguishable from random guessing ($p = 0.510$).

Using this tool, we compared multi-person groups (emulating a jury) to individuals (emulating a single judge or moderator) in the consistency of their decision-making. We performed field experiments with juries consisting of 5–9 Amazon Mechanical Turk (MTurk) workers. Each experiment involved four randomly-ordered rounds, with two individual rounds and two group deliberation

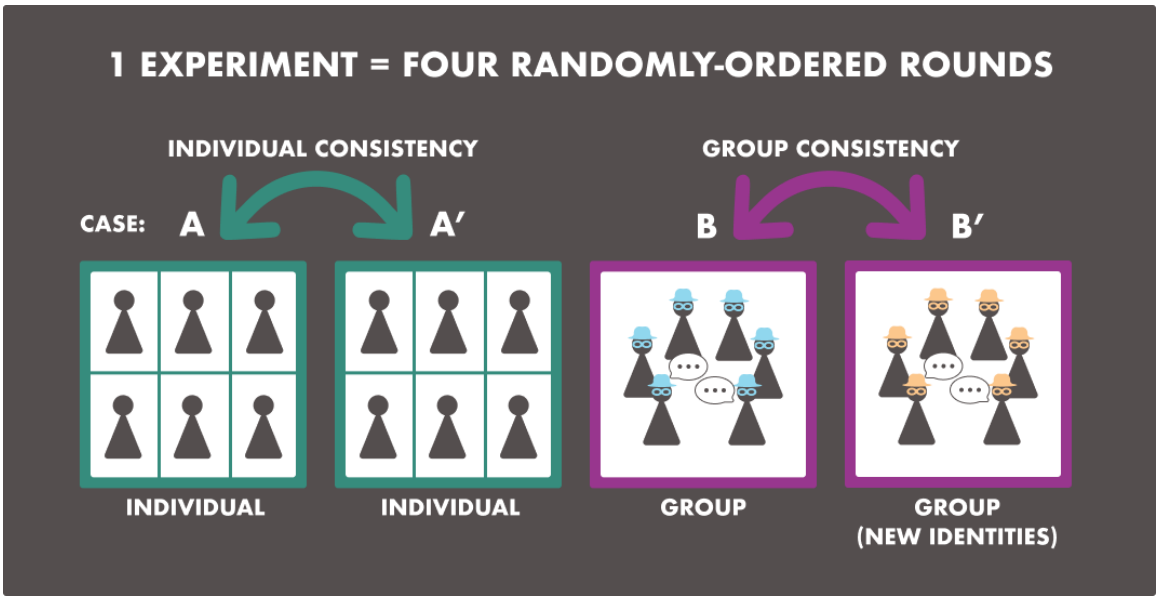

Figure 1.1: A single experiment consists of four randomized rounds: two individual tasks and two group tasks. Participants remain the same throughout all 4 rounds. In the second group round, the changed mask color indicates a new pseudonym identity. The tasks for each round are cases drawn from two pairs, A/ A′ and B/B′, which are known to have correlated outcomes. We measure consistency by comparing the rates at which individuals and groups consistently adjudicate the pairs.

rounds (Figure 1.1). Participants in the two group rounds adjudicated a pair of correlated cases, known to have related outcomes based on a pre-testing process. The use of paired cases, rather than identical ones, ensured that participants' decisions on one case would not be influenced by their experience deciding another.

We then measured consistency, defined as coming to the same judgement for the pair of correlated cases. On the group level, consistency is calculated by aggregating the final votes of all jury members and taking a simple majority (*aggregate consistency*), and on the individual level, by examining the votes of individual jurors (*participant-level consistency*). Furthermore, we used pre-deliberation and post-deliberation surveys to measure if participants changed their minds throughout the process.

Prior work establishes that groups make decisions by integrating social information (Becker et al., 2019); this, we hypothesized, made groups vulnerable to arbitrary features within deliberation, from the order of speaking (Sunstein et al., 2014) to the jurors' status characteristics (Marder, 2005). Group polarization (Sunstein, 2000) would further shift jurors' tendencies from their initial leanings, making groups less stable — and, we believed, less consistent — adjudicators.

We find, indeed, that individuals change their minds more frequently after deliberating with a group. But to our surprise, we discover that the resulting decisions are no less consistent. A paired analysis between individuals working alone and the same individuals working in groups finds that

their consistency is within 6% of one another. Further, integrating the perspectives of all jurors is not as simple as conforming to a majority opinion; though we did affirm evidence of group polarization, minority jurors had their way more frequently than we expected.

Our results are ultimately optimistic for democracy, but not without some reservations. While group decisions are no less consistent than individual ones, consistency is a necessary, but ultimately insufficient, criterion for procedural justice. Designers of governance policies should be wary that, even with identities masked and most status characteristic erased, similar social dynamics managed to resurface in a quarter of all juries. A strong personality can be louder than outward signals, lending a vocal few disproportionate influence.

We also note that information integration within juries is of highly variable quality. From literature, we know that deliberation makes good teams great, and bad teams worse (Kerr, MacCoun, & Kramer, 1996); as a result, a team of biased decision makers may turn out "consistent," but only because deliberation reinforces heavily biased judgments. An open question, therefore, is finding ways to scaffold information integration, preventing less competent members from succumbing to groups' worst vulnerabilities.

Perhaps most interesting of all, we learned that individuals were spot-on at predicting their own consistency, but not the consistency of groups. Of all participants surveyed, 70.2% believed that individuals would make consistent judgments; the actual consistency turned out to be 65.9%. However, participants were incredibly pessimistic about the consistency of groups — just over half (51.6%) believed that groups would be consistent, but true group consistency turned out to be 64.8%. Our results thus suggest that deliberative democracy is not only consistent, but it is also, in some sense, underestimated. Though one might retort that this answer ought to have been clear from the beginning — the Anglo-American jury system must exist for a reason! — the survey results remind us that these findings were, in fact, not so obvious (until we learned the answer) (Watts, 2011).

Finally, the public's lack of confidence in jury consistency may also have implications on how jury decisions are perceived. Even when juries deliver consistent decisions, others may still *assume* that they are inconsistent. Thus, those that design deliberation-based governance ought to rightly ask: to whom should decisions be justified? Could consistent decisions still appear less fair?

Chapter 2

Related Work

Group consistency has not been directly measured. That is, no prior studies have investigated the consistency of repeated deliberations conducted by the same group. Previous experiments have instead focused on analyzing the social and informational influences that ultimately contribute to a decision outcome. This chapter will survey these known influences in the context of their potential impact on the consistency of both individuals and groups.

2.1 Influences on Individual Consistency

An individual's decision consistency is most heavily dependent on the individual's decision-making strategy, which is itself reliant on three major factors: (1) the features of the person (i.e., their cognitive ability and knowledge); (2) the features of the problem; and (3) the context that the individual is situated in (i.e., whether the individual feels accountable for making a particular decision). One model of this view is that individuals are *adaptive decision makers*, who, when confronted with overwhelming information and multiple (conflicting) goals and values, make decisions via available heuristics (Payne, Payne, Bettman, & Johnson, 1993). However, humans do not always adjudicate these tradeoffs consistently; for example, it is possible to set up a betting problem such that people would choose gamble 1 over gamble 2, yet bet more money on 2 than on 1 (Lichtenstein & Slovic, 1971). Furthermore, human judgements are subject to arbitrary extraneous factors; judges' parole decisions can vary depending on when they take their lunch break (Danziger, Levav, & Avnaim-Pesso, 2011).

In general, however, humans tend to predictably use a few decision-making heuristics. A study of Supreme Court justices found that, when examining the justices' individual decision-making, a handful of cues about the case could predict a substantial portion (up to 79% in the best case) of decisions correctly (Segal, 1986), suggesting that individual decision-makers tend to narrow down complex cases into a few essential pieces. Thus, if the essential pieces of two cases are the same, one

might expect between 60% to 80% consistency for individuals.

2.2 Influences on Group Consistency

Prior work in group consistency is broadly drawn from two areas: jury studies and social psychology. In the field of jury studies, there are five traditional research methods: archival analysis, post-trial interviews, shadow juries, field experiments, and mock jury experiments (MacCoun, 1987). These studies have contributed to our understanding of the influences that shape a jury's verdict, from the members' demographic representation to the strategies they use to deliberate. More broadly, because juries are a special type of group task, prior work in social psychology has also contributed to our understanding of the phenomena that influence group performance: conformity, cascade effects, and group polarization are among them.

One notable mock jury experiment is a 1998 study showing that jury verdicts cannot easily be predicted. In the study, the authors used actors to videotape different versions of a civil suit, with details altered in each version. Aggregated across the versions, the case was balanced overall (immediately after viewing the trial, 51% of participants favored the plaintiff). Following 120 mock jury experiments, the authors attempted to predict a jury's final verdict using the case information and jurors' demographic characteristics. They found that these features had only partial predictive power: the version of the case explained 57% of verdict preferences; adding weighted information about jurors' demographic data and predispositions (drawn from surveys) could explain up to 67% of verdict preferences (Diamond, Saks, & Landsman, 1998). However, features of deliberation explained by neither the content of the case nor the features of jurors accounts for fully 33% of verdict variation.

The remaining unexplained variation is deliberation's unique contribution. Indeed, theories of deliberative democracy argue that the ability to expand one's viewpoint, and to think from other perspectives, is precisely the epistemic justification for having deliberation in the first place (Habermas, 1992). The crucial point is how this variation changes the ultimate consistency of group deliberation: does it introduce new information that enables greater consistency, or does it introduce an element of randomness that makes jury decision-making less consistent? This section will provide a detailed view of the group decision-making process, examining four specific phenomena that potentially influence consistency: (1) Integrating (or failing to integrate) new information; (2) Using different decision-making strategies based on task type; (3) Amplifying errors and shifting to extremes; and (4) Using "fast and frugal" decision-making heuristics.

2.2.1 Integrating (or Failing to Integrate) New Information

Often a motivating reason for using a deliberative group is the ability to integrate information from multiple perspectives into a more representative whole. Participants in deliberation introduce overlooked facts, compelling analysis, and novel perspectives. Research on team-based learning

activities has found that, when a group is successful at integrating the knowledge from individual members, it outperforms its most proficient member in 97% of final projects. 40% of this increase in performance was unexplained by either the average performance or the best individual performance alone, suggesting that, in essence, the whole was more than the sum of its parts (Watson & Black, 1989). Similarly, in a 'survival task,' for which participants are required to rank a list of objects based on how useful they would be in a wilderness survival setting, groups routinely outperformed ($p \leq 0.01$) the average individual and tied with the top-performing individuals in the survival task (Miner Jr, 1984). Thus, the ability to effectively synthesize information enables groups to consistently arrive at accurate, high-quality decisions.

However, the process by which groups integrate information plays a substantial role in the quality of the outcome. For example, studies of juries have found that the structure of deliberation significantly shifts the final verdict: juries convicted more often when they evaluated complaints in descending order of seriousness (Devine, Clayton, Dunford, Seying, & Pryce, 2001). Similarly, *cascade effects* can play a role in altering the credibility of certain perspectives. Group members are far more likely to follow the perspective of the person who happens to speak first (Sunstein et al., 2014).

The person who speaks first is also, more likely than not, an individual with a more privileged background. Status often plays a substantial role in one's willingness to share knowledge. In one mock jury study, the sole African-American juror felt so excluded from the deliberations that he "literally left the table" (Marder, 2005). Moreover, juries tend to choose white, middle-class males of high status to serve as the foreperson (Marder, 2005; Hastie et al., 1983), which in turn enables this individual to wield disproportionate influence over the deliberation (Devine et al., 2001).

In fact, groups do a poor job of integrating knowledge from any kind of minority viewpoint, not just demographic minorities. If the number of individuals holding the minority view is too small compared to the majority faction size, minority faction members often choose not to speak up at all. In Hastie et al.'s mock jury trials, 5% of participants in 6-member juries remained silent throughout the deliberation. When the size of juries increased from 6 to 12, the percentage of passive individuals jumped to 20%. Finally, even when minority faction members do choose to speak up, their perspectives are discussed less frequently (Stasser & Titus, 1985). As a result of these factors, groups often do a poor job of integrating novel perspectives.

Another study, which compared pre-deliberation dispositions of jurors to their final verdicts, echoed these results. Most results of mock trials matched the pre-deliberation leanings of the majority faction, except when a critical mass of opposition voices (a 'viable minority') existed. A similar critical mass threshold has been explored in Asch's line-matching experiments, in which individuals became more willing to speak up after the minority faction size exceeded three (Asch, 1951). Thus, minorities were more influential in smaller juries (6-member rather than 12-member), since they accounted for a larger proportion of the overall group (Roper, 1980).

These findings have two implications for group consistency. First, if groups are effective at integrating new information, it suggests that they would be more consistent than individuals. Groups would likely uncover high-level analogues between similar scenarios and consistently arrive at accurate judgements. Second, if groups improperly integrate information fully — by ignoring minority viewpoints, or by changing their conclusions based on the arbitrary order of speaking — they are likely to be highly inconsistent. The question is whether the former tendency overrules the latter.

2.2.2 Using Different Decision-Making Strategies Based on Task Type

To understand which tendency — synthesizing information or succumbing to arbitrariness — is more likely to take place in a jury context, we next turn to studies of social influence. Moscovici and Faucheux (1972) theorize that there are three possible outcomes of deliberation: (1) Conformity (convergence to the majority position); (2) Innovation (acceptance of the minority view by the majority); and (3) Normalization (compromise). However, group deliberation outcomes vary among the three depending on how the question is phrased: when asked to prescriptively discuss what 'should' be the solution to a dilemma, groups tended to conform to majority perspectives. In contrast, groups were more open to normalization when asked to describe an objective truth, and the question is framed in terms of what would 'realistically' be done (Thorne et al., 2004).

This finding may explain why groups are at times effective at synthesizing information, and at other times vulnerable to following leaders and arbitrarily ignoring minority opinions. Work on deliberation accuracy has largely been focused on truth-based tasks (i.e., learning in group projects and the 'survival task'), which leads to normalization of knowledge from all contributors. The jury's role, however, is *both normative and objective*: though the jury is tasked to be an accurate fact-finder (Hastie et al., 1983), it must also make a normative judgement of whether the defendant should be condemned for the wrong attributed to them.

Empirical studies (Hackman & Katz, 2010; Hastie et al., 1983) have found that the initial verdict favored by members of the jury is typically also the final decision. This occurs in as many as 90% of cases (Devine et al., 2001), suggesting a strong tendency towards conformity over normalization.

However, textual analysis of Supreme Court decisions have found that members of the court are more "*integratively complex*" when the majority wins out; in other words, members of the majority faction tend to think more about trade-offs with alternatives than members of the minority. Though the Supreme Court is very much a special case when it comes to group decision-making (and its role should be in no way confused with that of a jury), the findings suggest that even when the majority prevails, the process is not so much simple conformity as it is a more normalizing one (Gruenfeld & Preston, 2000).

Thus, true to form, the mix of normative and objective elements in the decision-making task results in a mix of decision-making strategies. The stronger empirical evidence for conformity, however, suggests that juries treat their tasks more similarly to normative tasks than to objective

ones.

If juries normalize — that is, lean in the direction of the initial majority — this may increase their consistency relative to individuals. However, deliberation could obscure the true majority opinion: for example, the cascade effects described in the previous section might change the perceived credibility of certain answers. The next section describes how social information may amplify errors and cause jury opinions to shift to the extremes.

2.2.3 Amplifying Errors and Shifting to Extremes

In a jury context, each statement creates an *information externality* — a social signal about which sorts of judgements are generally reasonable to this particular group. Because of *reputational pressures* (that is, the desire to not be seen as deviant), individuals will tend to strongly conform to the perceived truth, even at the expense of honestly representing their own beliefs. Indeed, while there is a *polarization effect* within groups (that is, the group's opinion shifted more towards a more extreme direction after deliberation), private opinions often remain the same (Sunstein, 2000). Similarly, a large-scale survey of 3,500 actual jurors found that "over one-third of them would have reversed their jury's decision if they had been given sole control over the trial's outcome" (Son et al., 2019). In Son et al.'s mock jury experiments, jurors tended to increase the amount of recommended punishment after other jurors expressed a willingness to punish. Each additional punisher augmented an individual's rate of punishment.

Additionally, the 'wisdom of crowds' phenomenon (Surowiecki, 2005) has been shown to be inapplicable in cases of multiple-choice tasks — precisely like the sorts of "guilty/not guilty" verdicts that juries are required to make. By the Condorcet Jury Theorem (Condorcet, 1785), if an individual juror is 60% likely to rule incorrectly, a majority-rule jury is, by arithmetic alone, certain to amplify the error, and will have only a very small chance of a correct decision (Becker, 2019).

The impact of these effects on consistency is not known. Group polarization may increase consistency by repeatedly shifting the group conclusion in the same direction. However, the content of deliberation heavily shapes participants' perceptions of the norm. In a parallel universe, a participant who hid their private opinion in order to conform with others may choose to share their opinion instead — thus setting a new norm, and potentially causing the group to polarize in the other direction.

2.2.4 Using "Fast and Frugal" Decision-Making Heuristics

A final possible source of variability is the manner in which deliberation forces participants to consider their positions in detail in order to justify them to others, rather than taking a more 'fast and frugal' approach.

One study of individual jurors presented participants with text-based summaries of court cases. Participants' decisions were then compared to the true result from the court case. The authors

found that, when participants took longer to make decisions, their verdicts were, in fact, *less likely* to match the true result. The work suggests that fast, frugal decision processes can, in fact, allow decision-makers to filter out irrelevant signals and come to a quick and accurate decision (Curley, Murray, MacLean, & Laybourn, 2017). These findings supported research (Dhami & Ayton, 2001) which found that magistrates use, on average, only 1.1 cues when making bail decisions — a counter-intuitive finding, given the importance of such decisions.

One implication of this work could be that individuals are more consistent than groups, since, without the added pressure to thoroughly explain one's reasons to others, it is easier to rely on a fast and frugal decision-making heuristic. On the other hand, however, making decisions too quickly is ripe with unchecked error, and individual jurors may deliver ill-conceived and inconsistent results. Even expert judges make frequent mistakes; a simulation has found that improving jail-or-release decisions on the part of judges could reduce crime rates by 24.7% (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018).

2.3 Hypotheses

Our survey of related work indicates that the relevant influences in group decision-making generate mixed influences on consistency. Some, like conformity, suggest that juries will be quite consistent. Others, like group polarization and frugal decision-making, could be a potential influence in either direction. Still others, like social cascade effects, suggest that jury deliberation can be easily shaped by arbitrary forces, and therefore quite inconsistent. With regard to the consistency of jury deliberations, one might accurately say that the jury is still out.

A such, there is a gap in research investigating the overall directional influence of group deliberation on decision consistency. The current research seeks to fill this gap by directly investigating the consistency of juries via repeated online deliberations.

We hypothesize the following:

Juries will be less consistent than individuals.

More precisely, we split this hypothesis into three tiered sub-hypotheses — first, that groups will have a different level of decision-making consistency than individuals (H1) — and, specifically, will have a lower level of consistency than individuals (H2; H3). We pre-registered this hypothesis on 20 April 2020, prior to the collection of our data, at the following URL: `https://aspredicted.org/blind.php?x=9py7ab`

Hypothesis 1 (H1) *Juries and individuals have significantly different levels of decision-making consistency.*

Prior work in jury modeling confirm that the process of jury deliberation is not at all congruent

to the process of individual decision-making. For instance, the DeGroot Model (DeGroot, 1974) postulates that collective decisions are made by slowly revising beliefs, with each individual placing some weight on their previous belief and some weight on social information (Becker et al., 2019); another best-fit model predicts that jurors are susceptible to social influence, but become more rigid over time (Burghardt et al., 2019); a more traditional perspective is the "Story Model," in which jurors collectively build the narrative of events and determine the verdict from it (Hastie et al., 1983). By accounting for social information in some manner, we hypothesize that jury consistency will be different from that of individuals.

Hypothesis 2 (H2) *Juries' decisions will be significantly less consistent than that of individuals.*

Social information is a potential source of noise: the previous section has established that, even when juries attempt to converge to a majority opinion, the true majority opinion is obscured and contorted beneath information cascades and reputational pressures. Additionally, the margin for error is quite large. In a result found by creating parallel review processes for papers submitted to NIPS 2016, Shah, Tabibian, Muandet, Guyon, and Von Luxburg found support for the 'messy middle model' (Shah et al., 2018). 30% of papers submitted to NIPS 2016 could have, in a parallel universe, been decided another way. We hypothesize that, in these ambiguous cases, groups will be particularly vulnerable to social influence, and therefore will be more likely to make inconsistent decisions.

Hypothesis 3 (H3) *Even for juries composed of individuals who adjudicate consistently when alone, H2 is true.*

Anyone can find themselves pressured to change their mind in a group setting. Our final hypothesis, therefore, is that *even consistent individuals can form inconsistent groups.*

The next sections describe our methods for testing these group deliberation phenomena via an online jury platform. In our study, we manipulate participants' pseudonym identities within the online jury, such that participants reset prior social interactions and start anew. Thus, instead of measuring consistency by comparing pre-deliberation leanings to verdicts (Roper, 1980), by aggregating independent deliberations and measuring trends (Diamond et al., 1998), or by comparing the jury results to established "ground truth" answers (Curley et al., 2017), we aim to directly measure whether two independent instances of the same jury, making similar decisions, will lead to similar outcomes. Specifically, are decisions driven by the attributes of the case and decision-makers (held constant in our experiment), or by elements of randomness introduced via deliberation and social influence?

Chapter 3

Methods

3.1 Experimental Design

Our study was motivated by the central question: if the *same jury convenes twice, how will the consistency of its decision compare to the same individual convening twice?* To test this, we built an online platform that creates a virtual jury room, and assembled juries of 5–9 members. This mimics real-life juries of 6–8 members (Hastie et al., 1983; Marder, 2005).

The experiment compares two conditions: a group condition, in which all members of the jury discuss together and attempt to come to consensus, and an individual condition, in which an individual reads a case and decides alone. To account for the potential effect of deliberation time, all rounds were the same length. Moreover, we used self-elaboration to control for the varying depths of analysis that individuals would experience in the two conditions. Therefore, the main difference between the group and individual conditions was the social influence and presence of additional team members.

Each experiment consists of four rounds, two of which were randomly chosen to be individual, and two of which were randomly chosen to be group. All participants experienced all four rounds, and thus experienced both conditions twice. This allowed us to show that any differences between group and individual consistency were not be due to selecting two entirely different populations.

Additionally, the two individual rounds were assigned a pair of cases, which are known to have correlated outcomes. The two group rounds were similarly assigned a pair of cases with correlated outcomes (Figure 1.1).

The UI of the group and individual rounds were nearly identical (Figure 3.1; Figure 3.2). In order to visually signal that the task type had changed, the color of the side panel appeared magenta during group rounds and green during individual rounds.

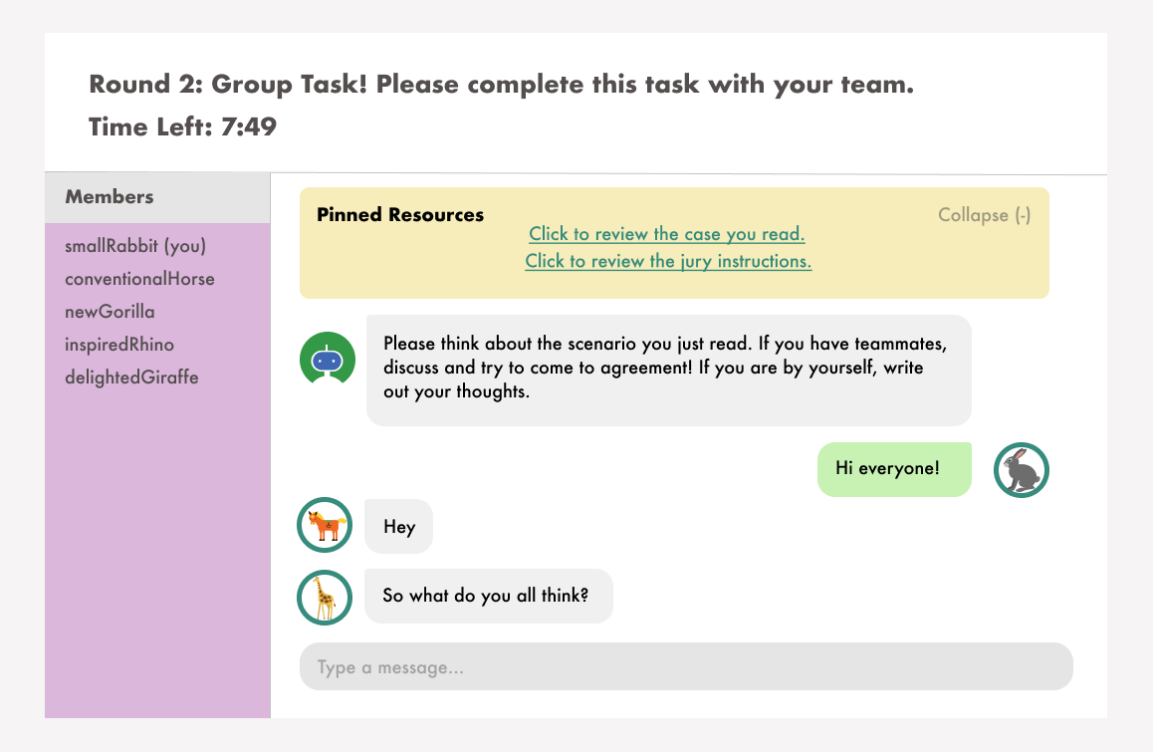

Figure 3.1: The interface for a group deliberation round, which is very similar to the interface for individual rounds; rather than work alone, participants chat with 4–8 other jury members.

3.1.1 Calculation of Consistency

Our core comparison is individual and group consistency. A key experimental decision involved determining how to measure consistency. Due to the prior inability to compare repeated interactions of groups, there is no agreed-upon definition of consistency. Definitions of consistency have ranged from "consistency with the actual decisions of juries" (Curley et al., 2017) to "consistency with one's predispositions" (Roper, 1980) to "consistency with predicted results" (Bone, Hey, & Suckling, 1999). Our novel approach of comparing groups to themselves therefore required its own standard of measurement.

What does it mean for a verdict to be consistent? Recall Hastie et al.'s (1983) description of the "hypothetical situation where the same case might be tried repeatedly by similar juries": a consistent verdict, therefore, is one that rules in favor of the same party each time a similar case is presented. Thus, for correlated pairs, consistency might be defined as having outcomes that align with the known correlated outcomes.

But the comparison between individuals and groups is, in fact, somewhat difficult to make. What counts as the group's outcome? One answer might be the result of a unanimous vote. However,

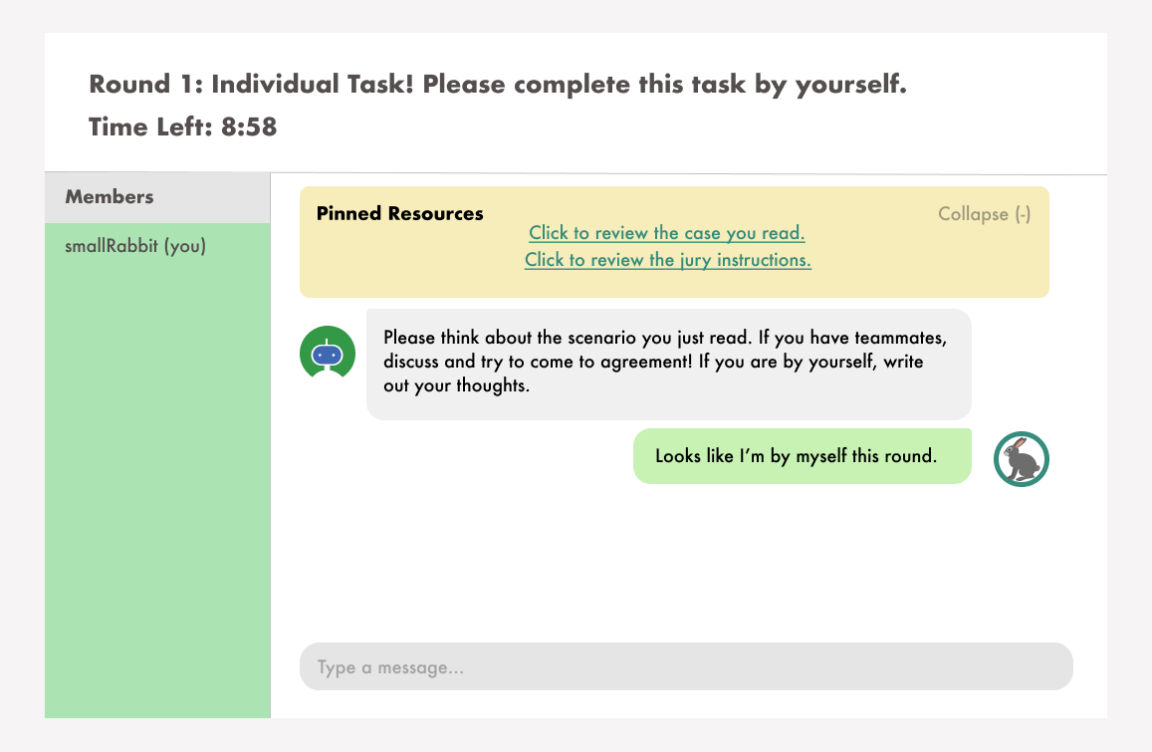

Figure 3.2: The interface for an individual round. Participants are alone in a chatroom, with only preset messages from a chatbot. A 'Pinned Resources' bar enables participants to review the case materials. The chatroom also displays the list of member pseudonyms, information about the round, and a countdown timer.

not all votes are unanimous: should a 4:2 majority vote be simply thrown out, or should it still be counted?

Furthermore, there is at least 5 times as much individual data as group data (since groups sizes range from 5–9). Measuring the difference between individuals and groups thus becomes a rather odd statistical comparison: the raw individual outcome is compared against some mathematically computed function of 5–9 aggregated outcomes. Apples are, in a sense, being compared to oranges.

There are thus several possible approaches to compare individual and group consistency, each with its own benefits and drawbacks.

Compare Groups to Individuals.

This comparison addresses our question most directly. Group consistency was computed via the *aggregate consistency* of all members in a team. Here, the group's outcome can be computed by either (1) simple majority vote, or (2) unanimous decisions (with non-unanimous decisions removed from the data). *Aggregate consistency* is then defined having the same group outcome in both

rounds.

The group's consistency can then be compared to individual consistency, which is calculated at the *participant level*: that is, an individual is consistent if their two in-round votes have the same outcome.

While this measure of consistency is the most straightforward, it compares two samples of very different sizes, and is blind to detailed changes in voting patterns between the group and individual conditions. Thus, two additional measures of consistency emerge.

Compare Groups to Nominal Groups.

One possible way to remove the disparity in sample size is to conduct a paired analysis between groups, treating individuals as "nominal groups" (that is, groups that never discuss in person). Since individuals could not have been asked to reach a unanimous conclusion with partners whom they never met, the outcome of a nominal group would be determined by a majority vote-based calculation. In a comparison between groups and nominal groups, then, we would compare whether the introduction of deliberation and social influence changes the aggregate consistency of the same individuals.

Compare Individuals to Individuals in Groups.

A second possible way to remove the disparity in sample size, while also revealing insights in individual voting patterns, is to calculate the participant-level consistency of individuals in group rounds, and conduct a paired analysis with the consistency of individuals while working alone. Such an analysis would have the potential to reveal whether individuals might to vote against their own interests and behave inconsistently while interacting in a group setting.

We ultimately argue that all three methods are valid ways to calculate the difference in individual and group consistency. Each one provides a distinct lens on the behaviors that we are interested in. Thus, the next section will report experimental results for all three metrics, with each revealing a different part of the story: an aggregate measure of consistency (comparing groups to individuals); a measure of whether individuals change their behavior while in groups (the paired analysis of individuals to individuals in groups); and a measure of the effect of deliberation in pushing teams towards consensus (the paired analysis between groups and nominal groups).

3.1.2 Pseudonym Masking

Of course, it is not enough to simply convene individuals and juries online. When the same group of people collaborate repeatedly, their second interaction builds upon the first (Morgan Jr, Salas, & Glickman, 1993), which means that their second interaction would not be independent of the first. In order to measure consistency directly, we use the *parallel worlds* technique of *one-way*

pseudonym masking that our prior work has established (Whiting et al., 2020). In each deliberation, jury members are assigned random pseudonyms consisting of an adjective and an animal name (e.g, 'inspiredDolphin', 'littleBear', 'spryElephant'). When the jury reconvenes, each participant's teammates are assigned new names, while the users' displayed names remain constant. Thus, 'inspiredDolphin' sees herself as 'inspiredDolphin' in all four rounds, and merely believes that she is working with new team members.

Additionally, when addressing other individuals by pseudonym, the system replaces references to pseudonyms with the names displayed in the participant's own view. For example, 'inspiredDolphin' may appear as 'smallPig' to another user. When the user types 'smallPig' or a near-misspelling into the chat, it is automatically corrected to 'inspiredDolphin' in the other user's view.

This technique temporarily hides the group's interaction history. Our prior work (Whiting, Blaising, et al., 2019) introduced this method and verified that participants do not realize that they are working with the same group again. In essence, one group deliberation becomes a "parallel universe" that enables us to measure just how commonly a decision would have gone another way.

3.1.3 Participants

Drawing on previous work (Whiting et al., 2020; Whiting, Blaising, et al., 2019), participants were sourced from Amazon Mechanical Turk. We selected only workers located in the United States (fulfilling the language requirement) and who had completed at least 100 tasks. We recruited 9 individuals for each jury. This procedure slightly over-recruited in order to account for the perils of online experimentation — 9 members insulated against inevitable Internet connectivity issues and other sources of potential drop-off. We discarded data where 4 or fewer jury members remained.

Each round of deliberation in the study lasted 7 minutes, with 2 minutes both before and after the deliberation period to answer surveys, and 4 minutes to read the case and jury instructions. Additionally, the task included 2 minutes of a final survey and 16 minutes of a warm-up activity, totalling 78 minutes for the entire study. Workers were paid via bonus at a rate of $15/hour, per standards of fair payment (Whiting, Hugh, & Bernstein, 2019).

Throughout the study, we make clear through written norms that participants are required to participate in every round in order to remain in the study. During each round, a chatbot reminds jurors to remain active. Subsequently, participants who contribute nothing in a round are deemed 'inactive' and removed. We only analyzed data from experiments in which at least 5 members were active throughout the deliberation, thus modeling a real-life jury of 6–8 members. This check ensures that we control for group dynamics, which are different for smaller groups than for larger ones (Driskell & Salas, 2006), as well as ensures ecological validity with real-life juries.

3.1.4 Manipulation Check

An important assumption of the experimental design is that the group does not detect that it has worked with one another twice over the course of the experiment, thus ensuring that the verdicts are independent. Otherwise, if a team repeats its prior social context, the verdict would be derivative of the previous interaction, and would confound our measurement of consistency. We build upon prior work (Whiting, Blaising, et al., 2019; Whiting et al., 2020), which uses the same one-way pseudonym masking technique to reconvene groups that had previously worked together, and implement the same manipulation check mechanism at the end of our experiment.

The manipulation check assumes that, if subjects realize that they are working with the same team, they will recognize each other in subsequent rounds. We therefore present a participant with the pseudonym of a teammate from one of the two group rounds. The participant is asked to recognize this individual's other pseudonym from a roster of participants from the second group round. They do this by selecting the paired name from a drop-down list and providing a brief justification.

Exclusion criteria

Samples in which the number of participants who successfully identified their teammate was greater than 2 standard deviations ($\sigma = 1.098$) above the mean (1.014) were considered excluded from the study for potentially seeing through the manipulation. Additionally, we manually inspected chat logs and removed samples where participants explicitly recognized each other.

Quantitative Confirmation of Manipulation Success

We confirmed that the remaining data passed the manipulation check by comparing the correct guess rate to the guess rate predicted by chance. The rate of correct guesses is calculated by c/N, where N is the number of participants in the group, and c is the number of correct guesses. The chance of randomly guessing the correct teammate in a round is given by $1/N$. We then confirmed, across all participants, that $\Sigma(c)/\Sigma(N)$ was not significantly different from the average value of $1/N$ for all samples.

3.1.5 Selection of Cases

Adapting Case Materials for the Internet

Prior work has conducted studies in which subjects were exposed to various versions of the same case, with some facts and details altered (Diamond et al., 1998). Diamond et al.'s study used two cases, both of which described a civil dispute between an employer and employee, who alleged that toxic materials in his working environment had led him to develop a lung condition. Details between the cases, such as the number of cigarettes smoked by the employee and the amount of toxic materials

present in the environment, were altered between the case versions, and both cases were balanced. Additionally, previous work in jury studies have demonstrated that text-based mediums of trial presentation are just as effective as videos of trials (Pezdek, Avila-Mora, & Sperry, 2010) and that the decision-making process is unaffected by ecological validity (Curley et al., 2017).

The process of adapting even text-based jury summaries for the Internet, however, faced another challenge: even "quick" jury studies last 50–90 minutes (Curley et al., 2017; Roper, 1980). To repeat this not once, but *four times* — twice for groups and twice for individuals — would require convening 5–9 participants for 6 hours. This would be infeasible, as it would be unrealistic to expect subjects to remain engaged online for so long: the risk of drop-off, technological glitches (i.e., loss of Internet connectivity), and the sheer exhaustion for subjects made this pathway untenable.

Our goal with selecting cases, therefore, required adapting jury materials for the Internet Age. We were inspired by the rise of online family dispute resolution (Casey & Wilson-Evered, 2012; Conley Tyler & McPherson, 2006), an area in which family problems will be eventually resolved via online-only proceedings (Hodson, 2019). Thus, we sought out cases in the realm of family disputes, for which online text delivery would be not only effective, but natural.

Generating Correlated Case Pairs

A key requirement of the study is that cases must form demonstrably similar pairs, for which we could expect consistent outcomes, while maintaining the balance of the cases. Finding cases that met these criteria required sourcing disputes from a sufficiently large and diverse pool. To do this, we turned to a platform at the heart of the Internet: Reddit. Reddit is recognized to host numerous active online communities, with intricate systems of discussion (Weninger, Zhu, & Han, 2013). Specifically, we scraped 23,055 posts from Reddit's 'Am I the Asshole?' (AITA) subreddit, or discussion board. Members of this discussion board describe a controversial personal action, after which the Reddit community adjudicates who is at fault. Although it also includes many frivolous and lighthearted posts, the AITA discussion board features a large number of incredibly detailed vignettes of family disputes. Themes include conflicts between family members, divorce, and custody of children, often involving issues of cultural, religious, and financial clash.

We filtered these 23,000+ posts by the number of Reddit votes they received, finding candidate cases that were well balanced (similar number of votes on each side) and which provided detailed, serious content of a dispute. The filter resulted in 56 candidate cases. These 56 cases were then pre-tested on Amazon Mechanical Turk with 136 participants, and 8 of the most balanced cases were isolated.

We then wrote 5 of these 8 cases from a different perspective, to create the 'opposite side' of the dispute. A wife's post about her husband, for instance, would be rewritten as a husband's post about his wife. This created a second variant of the same dispute, which would have a strong negative correlation in opinions with the outcome of the first dispute. That is, a 'consistent' decision would

likely rule in favor of the same party twice, and therefore have negatively correlated outcomes. The first decision would be in favor of the 'plaintiff,' and the second, from the opposite perspective, would rule in favor of the 'defendant.'

Additionally, we replaced the names of noun-entities in the rewritten cases to generate novelty, and thereby avoid a learning effect between cases. Three of the balanced cases were discarded because there was insufficient information to recreate the other side of the dispute.

We then confirmed that these cases would be valid pairs by collecting a further 128 datapoints of opinions on the cases. Further, we ensured that viewing the rewritten cases did not prime individuals' judgements on subsequent cases. In a pre-test with $N = 39$, we compared the decision consistency for each case in the case pair to an unrelated third case. The decision consistency to the third case were not significantly affected by the ordering of two cases in the pair ($p = .803$), which indicated that there was no discernible learning effect: participants' decisions were not influenced by the order in which the cases were evaluated.

Two pairs of cases stood out, with 65.4% and 68.2% consistency, respectively. These case pairs were also all relatively individually balanced, with balance of 50.4%, 44.5%, 41.4%, and 55.4%. These cases were ultimately chosen for the study. Additionally, a post-hoc analysis found that neither group consistency nor individual consistency was influenced by which of the two cases was presented. Group consistency was statistically indistinguishable across the two cases ($z = -0.003, p = 0.998$), as was individual consistency ($z = 0.164, p = 0.870$).

3.1.6 Jury Affordances

The jury deliberation platform has features designed to create virtual analogues of the real-world jury room. Participants in the virtual jury were required to undertake three key activities:

1. Reading case materials;
2. Deliberating;
3. Voting.

Reading case materials is the virtual equivalent of the trial experience, where jurors are presented with evidence, arguments for both sides, and jury instructions. In the real world, these activities are physically conducted in a separate room from the jury deliberation, and jurors are forbidden from discussing the trial while it is ongoing ("Handbook for Trial Jurors Serving in the United States District Courts", n.d.). In the online analogue, then, virtual jurors are taken into a *reading period*, where they are presented with text-based content about the trial, as well as jury instructions based on the the 2017 California Civil Jury Instructions ("CACI No. 5009. Predeliberation Instructions", 2017) (Figure 3.3).

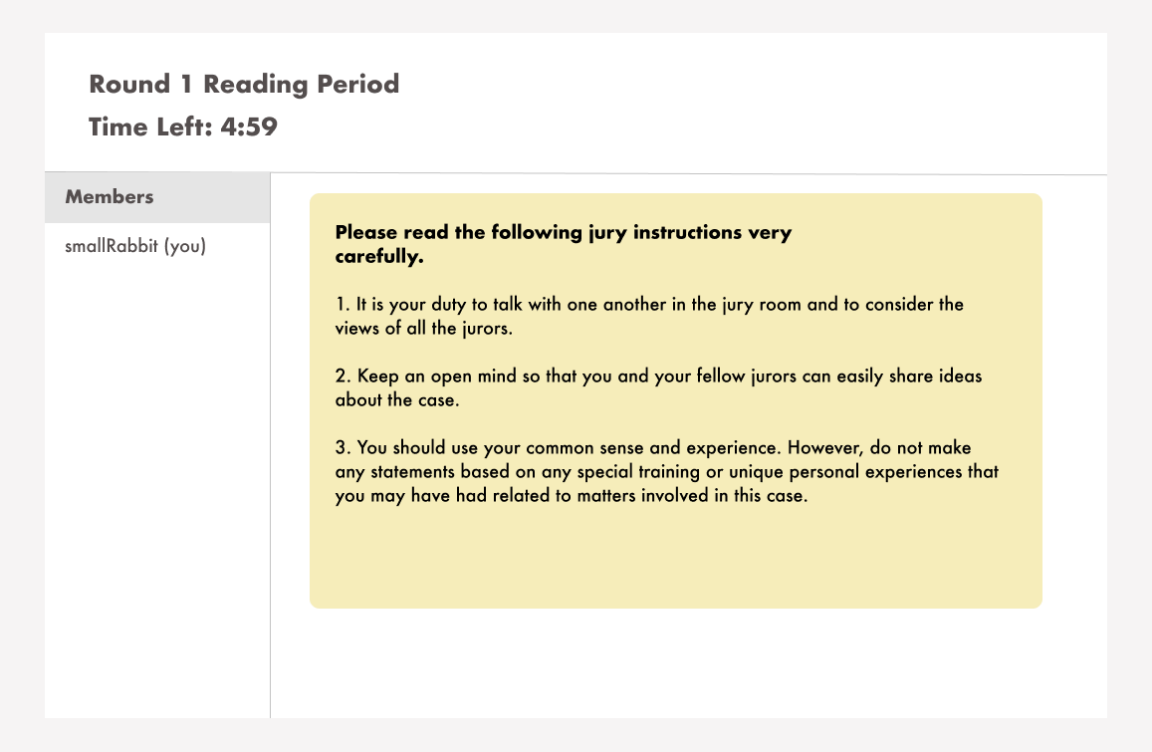

Figure 3.3: The interface for a reading period, where jurors read the case content on their own prior to entering the deliberation room. This mimics the trial period.

Participants are alone in the reading period, and, like offline juries, are unable to chat with any other participants. The timing of the reading period is calculated based on the average time required to read the material. The system supports multiple reading periods, so that the case and jury instructions can be communicated separately.

Once jurors have finished reading the cases and jury instructions, deliberation begins. Jurors are brought into a virtual chatroom (Figure 3.4). The chatroom's features are fairly standard, but features a dynamic voting mechanism at the top of the screen. When the round is 81% complete, a prompt appears, prompting jurors to vote and come to a conclusion. The late-round timing of the prompt was implemented based on the *evidence-driven jury deliberation style*, in which jurors begin with discussing the evidence and have an official vote only at the very end of deliberation (as opposed to beginning with an official vote and working to convince minority members in a targeted manner). This style of deliberation has been shown to be more robust in integrating a variety of opinions, and jurors participating in evidence-driven deliberation are more likely to rate the process as persuasive, open-minded, and serious (Hastie et al., 1983).

Given that the effects of social influence are of interest to us, we implemented a semi-transparent system in which the current vote percentages (i.e., 30% Defendant, 70% Plaintiff) are visible to

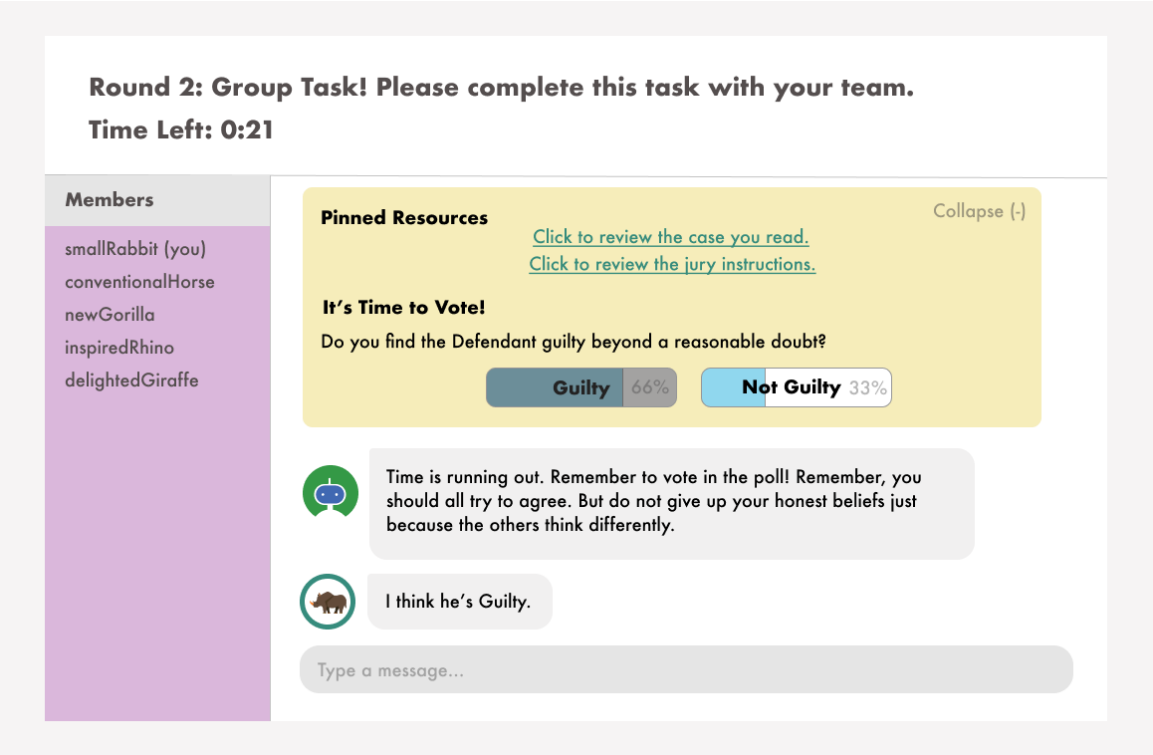

Figure 3.4: The interface during the voting period. A prompt appears, and jurors individually select their verdicts. The voting buttons display live poll results. Once a juror selects an option, the option is disabled; however, they are able to change their vote until the end of the round. Note that the prompt text here is merely an example.

all members of the jury, but the specific votes are not. Thus, the voting mechanism functionally implements a blind vote, in which jurors can view results but not details. Jurors wishing to undertake a more public vote are also able to do so by directly typing in the chat.

In this way, our platform captures several defining features of the jury experience. The next chapter details the results of our online jury experiment.

Chapter 4

Results

In our study, $N = 989$ subjects participated in 189 online juries. We removed juries in which fewer than 5 members were active, and those in which the number of participants who failed the manipulation check was greater than two standard deviations above the mean. 125 valid juries remained, with $N = 770$ participants completing all four rounds of the experiment. Group sizes ranged from 5 to 9 ($M = 6.136, \sigma = 1.073$).

Participants were 58.3% female and 41.2% male, with the remainder identifying as "Other." 74.7% of participants self-identified as White, 10.2% as Black or African-American, and 9.2% as Asian. 7.8% of participants self-identified as Hispanic or Latino. The average age was 36 years old, with a standard deviation of 11 years. 17.5% of participants had prior jury experience.

Passing of Manipulation Check

Within each team, the accuracy of answering the manipulation check question ($\mu = 0.166, \sigma = 0.150$) was, in fact, lower than chance ($\mu = 0.202, \sigma = 0.040$), and not significantly different ($p = 0.510$). Thus, because participants were unable to do better than random guessing in identifying former partners, we consider the use of one-way pseudonym masking to have successfully "erased history" and enabled juries to reconvene independently.

Qualitative Confirmation of Manipulation Success

To further ensure that participants had not detected the manipulation, we also manually inspected the justifications that participants provided during the manipulation check. We found that a small number of participants detected that a teammate was the same because of their polarizing beliefs:

> *I feel conventionalRabbit had some strong opinions about speaking the language of the country that people are in. I feel like i got that in the 3rd round too.*

If memory serves, likelyBear was one of the people who had a strong sense of familial obligation from round 2

However, many admitted that they were less focused on the pseudonyms than on the content of deliberation:

If I'm correct in recalling the names with their comments, I believe they were the same person. But I was focusing more on the comments and not the names of the people making them.

Finally, the vast majority of quotes indicated that they did not recall the names and had, in fact, guessed randomly:

I have no idea. I didn't notice that they were the same people they seemed quite different.

Just a guess. I didn't pay attention to names in the chat, just the remarks.

I honestly did not pay attention that much to names when arguing view points as I thought they would be randomly assigned anyways.

I really didnt know, Im just guessing here

These remarks strongly confirmed to us that the manipulation had been successful.

4.1 Are groups as consistent as individuals?

We hypothesized that groups would be less consistent in decision-making than individuals, because the process of accounting for social information introduces randomness in the decision-making process and causes groups to shift from their prior dispositions. The participants in our task agreed with our predictions: in a post-task survey, 78.1% of participants believed that individuals would be more consistent than groups. Further, only about half (51.6%) of participants believed that groups would be consistent at all, compared to over two-thirds (70.3%) of participants who believed that individuals deciding alone would be consistent.

And so, they — and we — would be surprised to find that the results tell a very different story.

4.1.1 Key Result: People are equally consistent, whether in groups or working alone.

2-Proportion Z Test

In our field experiments, we found that groups were 64.8% consistent, and individuals were 65.9% consistent (Figure 4.1). Our simplest statistical measure, a 2-proportion Z test, found this difference insignificant ($z = -0.229, p = 0.819$).

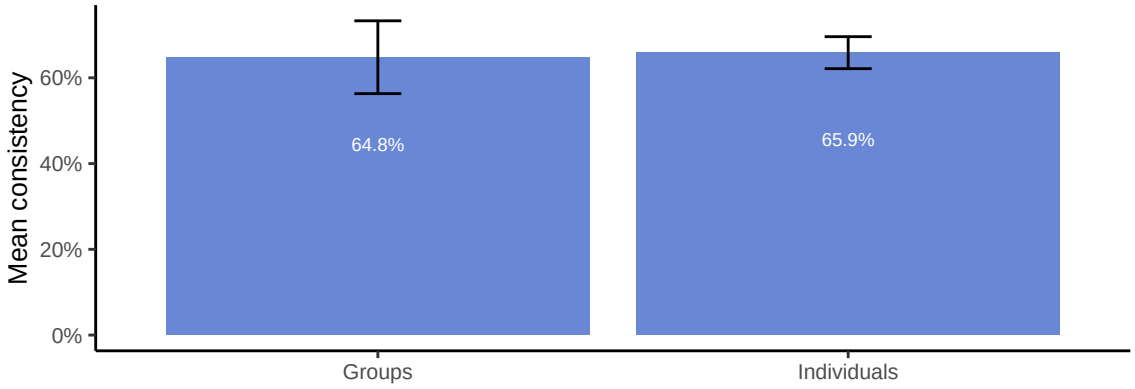

Figure 4.1: Consistency of groups and individuals, based on in-round voting. We see that the consistency levels are remarkably equivalent between the two.

Thus, the data does not support H1, H2, or H3, as group and individual consistency differences are too similar to be discernible. Indeed, a Fisher's Exact Test yielded $p = 0.918$, suggesting that, in fact, the opposite is true: individual and group decisions are equally consistent. This revelation shook our prior assumptions about the effects of group deliberation, and we conducted several additional statistical tests to validate these results.

Paired Individual Analysis: McNemar's Chi-squared Test and Bootstrapped Confidence Interval

We investigated the specific paired behavior of individuals within the individual versus group conditions. To do so, we conducted a McNemar's Chi-Squared Test with continuity correction. We satisfy all assumptions in the test, as our data involves dichotomous variables (consistent or inconsistent), one independent variable (group or individual), and two connected groups (the subjects in both conditions are the same).

The χ^2 value of McNemar's Test was 0.064, with $df = 1$ and $p = 0.800$. In other words, our data indicates that individual consistency does not change during group interactions. Further, using 5000 bootstrapped samples, we calculated a 95% confidence interval or the paired mean difference between individual consistency in groups and individual consistency while working alone. The resulting interval ranged from -0.050 to 0.061, indicating that the difference in participant-level consistency between the group and individual conditions is no more than 6%.

Figure 4.2 presents the results of the paired analysis between individuals in groups and individuals working alone — demonstrating that there are almost no differences between individual consistency while working in groups versus individual consistency while working alone.

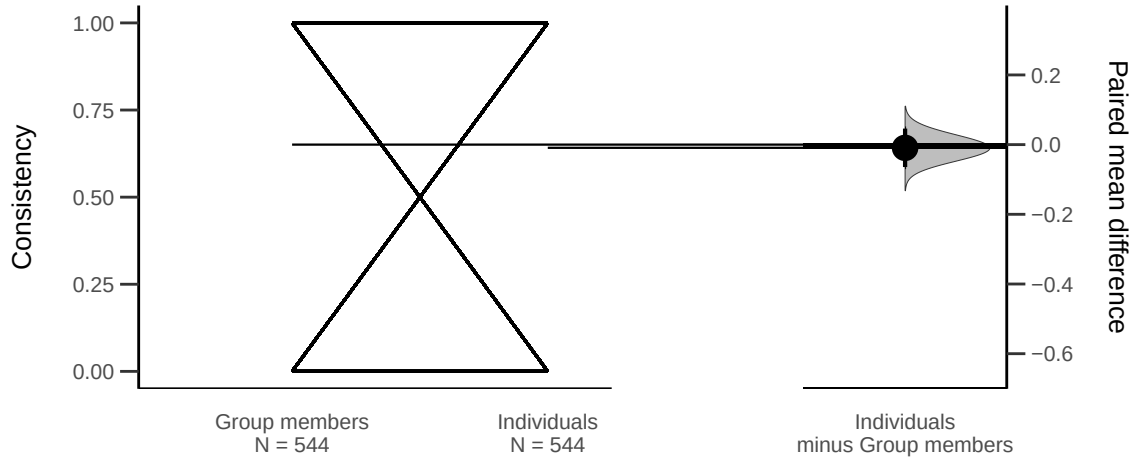

Figure 4.2: Paired analysis of individuals who worked in groups (group members) with the same individuals, working alone. We find the participants' behavior does not change significantly between the individual and group conditions.

4.2 Group Polarization and Conforming to Majority

Our results confirm prior work, which indicates that most deliberative groups arrive at the same conclusion that an initial majority of the group members were predisposed to support (Hastie et al., 1983; Devine et al., 2001; Thorne et al., 2004; Hackman & Katz, 2010). By comparing pre-survey results to during-deliberation poll results, we found that 80.0% of juries arrived at the same conclusion as that of the initial majority.

We compared this result to that of nominal teams, in which the votes of individuals who never collaborated are aggregated as a "team." We found that nominal teams support the initial majority in a much greater number of cases than true deliberating teams (86.0% of rounds). A Z test found a fairly substantial difference ($z = -1.786, p = 0.074$), but one that is ultimately not too surprising: without deliberation, most individuals rarely change their minds from their initial deliberation. Thus, the final "vote" tends to be identical with the initial one. The fact that conformity occurs in a smaller number of deliberating groups indicates that collaboration afforded minorities a greater opportunity to sway the ultimate verdict.

Crucially, deliberating teams demonstrate a substantial increase in the *strength of the majority* following deliberation; social proof, as well as mere exposure to different viewpoints, tends to shift groups more strongly in the direction of their prior leanings (Sunstein, 2000). Our data finds that, among nominal teams that voted for the same result as their pre-deliberation leaning, only 7.4% strengthen their majority (defined as having a greater majority faction size after deliberation than before). In contrast, nearly half (47.5%) of deliberating groups strengthen their majority, indicating a significant ($z = 9.212, p = 3.203 \times 10^{-20}$) group polarization effect, in addition to conforming to

initial leanings.

4.3 The Persistence of Opinions

How often do individuals change their minds?

While individuals make self-consistent decisions both within groups and alone, those working with groups tend to more frequently *change their minds* from their initial leanings. We find that 28.3% of individuals working in teams changed their minds from their initial predispositions, compared to only 8.1% of those working alone (these people, perhaps, managed to convince themselves to change their opinion via self-elaboration). The size of this difference is rather stark ($z = 10.307, p = 6.570 \times 10^{-25}$), indicating that group deliberation is effective in persuading some to change their opinions.

But precisely how many people were persuaded, and how persistent was this persuasion? To see this, we examined the number of dissenters in each group deliberation — that is, the number of individuals who voted against the majority. We find that a mean of 1.83 dissenters ($\sigma = 1.47$) per group round, indicating that consensus is by no means easy, and holdouts often remain.

Indeed, chat logs seem to reflect instances in which attempts to convert minority members failed:

> **youngBison (minority member)**: *I agree he is right with his feelings, but his tone and being hostile was over the top.*
>
> **niceMonkey**: *youngBison if you read the entire prompt, or listened to what we were saying, you would have known that he didn't say that.*
>
> **littlePanda**: *I think he was more saying what he wanted to say to her in his head. He didn't actually tell her that. I think he let his frustration out in the post.*
>
> **culturedRhino**: *niceMonkey FTW*
>
> **niceMonkey**: *thanks culturedRhino*
>
> **littlePanda**: *I feel we are all on the same track except for youngBison*

youngBison, the lone minority member, had previously repeated his position twice. He said nothing for 3 more minutes while his teammates attempted to persuade him. Then the deliberation ended, with **youngBison** presumably unmoved and ultimately outvoted.

One possible question is whether the time constraint had artificially limited the opportunity to persuade **youngBison**, or whether he would have remained unpersuaded. Certainly lengthening the conversation would have helped in many cases — but in others, faced by gridlock, and at times by unmoved silence, more time in the deliberation room appears unlikely to be effective. Thus, the question is whether those who were unpersuaded would have been helped with more time, or whether further deliberation would have only set them deeper in their beliefs.

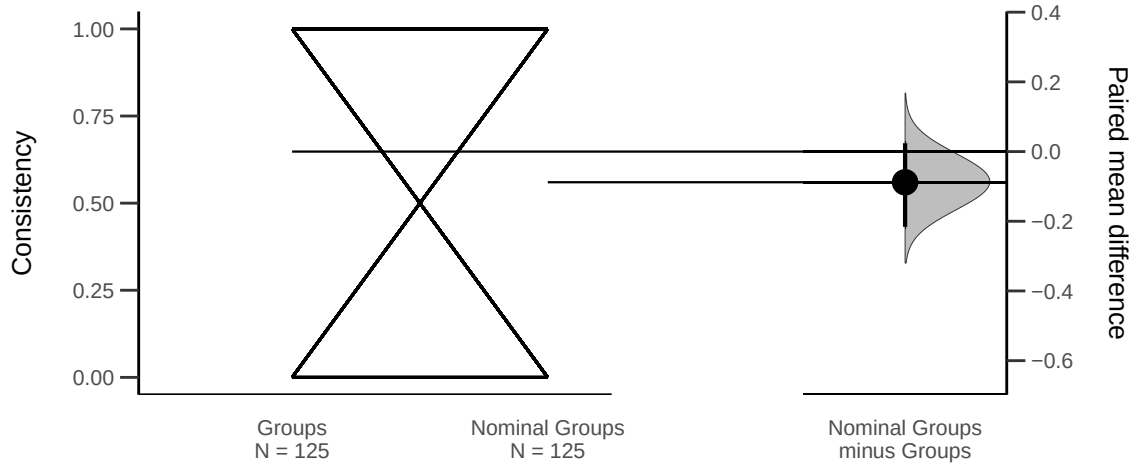

Figure 4.3: The paired analysis between groups and nominal groups (aggregated votes of individuals who worked alone). We find a slight leaning towards nominal groups being less consistent than deliberative groups, but the difference is not significant.

How authentic are jurors' beliefs?

Using post-deliberation survey data, we measured the authenticity of in-round beliefs. A vote is deemed inauthentic when an individual indicates a contradictory verdict in the post-round surveys. We find, in general, that most decisions were authentic, with only a few exceptions: on average, 0.2 individuals per round voted inauthentically, with $\sigma = 0.54$.

4.4 Group v. Nominal Consistency

Finally, we conducted a paired analysis between group and nominal consistency. Each group was paired with its nominal counterpart — that is, the aggregated opinions of the same individuals, acting alone (Figure 4.3). McNemar's Chi-squared test with continuity correction did not yield significant results ($\chi^2 = 1.639, p = 0.200$), and the bootstrapped Confidence Interval found that the mean difference in consistency ranged from -0.216 to 0.024 — meaning that nominal teams likely to be less consistent than their deliberating counterparts. This result weakly demonstrates the phenomena recorded in other metrics — the process of coming to consensus shifts groups to be a more consistent unit than nominal teams, which do not deliberate.

4.5 Does consistency change over time?

We tracked the consistency of individuals and groups at three points in time. pre-deliberation, during deliberation (the official vote), and post-deliberation. Each of the three votes represents

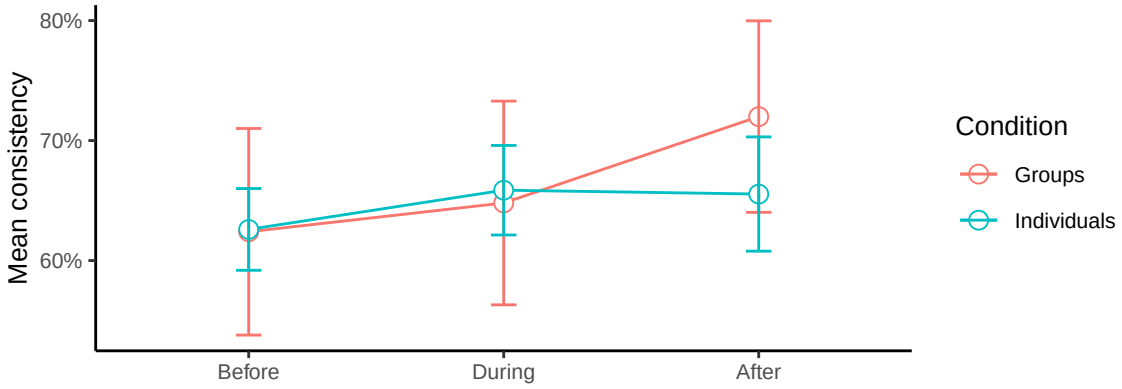

Figure 4.4: The consistency of individuals and groups plotted over the course of a round. Surveys are measured before, during, and after deliberation.

a snapshot of opinions in time. Pre-deliberation consistency examines the consistency of opinions prior to engaging in deliberation; the official vote represents the consistency of the immediate results of deliberation; and post-deliberation consistency examines opinions after subjects have voted and spent a few minutes reflecting on their experience.

Figure 4.4 indicates that group consistency has a clear trend of increasing over time. Indeed, as groups deliberate and come to consensus, their opinions become more and more consistent over time. Consistency of groups at the during deliberation is slightly but not significantly higher than the consistency of groups prior to deliberation ($z = 0.394$, $p = 0.347$); by the post-survey, a few minutes after deliberation, the consistency of group opinions becomes much higher than consistency pre-deliberation ($z = 1.617$, $p = 0.053$).

In a way, the post-deliberation survey results can be interpreted as the participants' final, authentic decisions for the round. The results may indicate that opinions take time to settle, particularly after participants have the opportunity to reflect on the results of deliberation. Prior findings suggest that group deliberation is viewed as more legitimate by participants (Fan & Zhang, 2020); in this case, perhaps this effect of deliberation also causes participants to view a group judgement with rose-colored lenses.

4.6 Change in Voting Patterns

We finally measured the extent to which individual voting patterns changed between the group and individual conditions. We conducted a Chi Squared analysis with $df = 3$, in which we compared the difference in voting frequencies across four categories: 'Plaintiff/Plaintiff,' 'Defendant/Defendant,'

'Plaintiff/Defendant,' and 'Defendant/Plaintiff.' This created a more granular view of how decision patterns shift between the individual and group conditions. For instance, if individuals more frequently voted 'Defendant/Defendant,' while groups more frequently vote 'Plaintiff/Plaintiff,' this distinction would not be visible in the consistency result. However, the shift in voting patterns might suggest that groups consistently come to different conclusions than individuals — an issue of substantial concern for those at trial. Conversely, a lack of statistical significance by this metric would indicate that individuals and groups are not only comparably consistent, but would also come to very similar decisions.

Our results show a relatively weak change in voting patterns based on in-round data ($\chi^2 = 2.938$, $p = 0.402$), suggesting that the effect of deliberation on in-round voting patterns is noisy, and therefore inconclusive. As the Discussion section will explore further, this weak signal is unsurprising given mixed literature in information integration theory.

Chapter 5

Discussion and Conclusion

5.1 Discussion

As we evaluate our own justice system and design new forms of conflict adjudication in the online space — from online dispute resolution (Hodson, 2019) to private digital juries (Ast, 2017) to content-moderating juries (Fan & Zhang, 2020) — it is crucial to understand the impact of group deliberation on the consistency of the decisions at stake. Perceived inconsistencies within the justice system can be a source of institutional distrust, inter-group tension, an an unwillingness to cooperate (van Prooijen et al., 2008). When it comes to group decision-making, the costs of getting it wrong are immense. The very fabric of our social institutions as we know them is at stake.

To an extent, the results of this study are hopeful: juries are just as consistent as individuals, and far more consistent than we had hypothesized. We find the results particularly interesting in light of the participants' own predictions. Recall that only half of our participants believed that groups would be consistent at all; most predicted that individuals would be more consistent than groups.

5.1.1 Changing Minds and Repeating Social Dynamics

An investigation into the individual survey responses reveals illuminating themes. Participants who answered that individuals would be more consistent highlighted the notion that individual opinions would be untainted by external influences:

> *You're more likely to change your opinions if you hear thoughts from others. If you're the only one thinking about a situation, you only have one POV influencing your thoughts.*
>
> *Individuals usually don't go back and forth with themselves after their first impression is made. They won't be challenged and made to expand their points without others.*

These quotes align with literature, supporting the notion that individuals will use their own decision-making heuristics (Payne et al., 1993), and will do so relatively consistently; we found that individuals were 65.9% consistent — a number that aligns well with prior work predicting individual decisions (Segal, 1986).

By contrast, participants who believed that group deliberation would be more consistent cited social influence as a consolidating mechanism:

> *I think that groups tend to be more consistent because they are all trying to agree.*
>
> *I think groups would be more consistent because someone may not want to upset the whole group or go against it.*
>
> *I think the peer pressure of a group will tend to keep someone's ideas [consistent].*

One subject even cited some of the motivating literature behind this work:

> *The wisdom of crowds applies, the crowd knows best.*

Our definition of consistency, it turns out, captures a unique relationship between these two predictions. Predicting that individuals are more consistent is correct, in a sense — unswayed by new perspectives, individuals were substantially less likely to change their minds from their initial predispositions (8%, compared to 28% for groups). Since group deliberation sways individuals from their initial predispositions, groups are "inconsistent" by this definition.

However, we defined consistency not as being aligned with one's predispositions, but as *making a similar judgment in a similar situation.* Under our definition, individuals remain consistent; those who view a similar scenario will apply the same heuristics and make the same judgment. And under this definition, we found that groups do the same — that group consistency is statistically indistinguishable from that of individuals.

So how is it that groups cause individuals to change their minds more frequently, yet they still consistently arrive at the same decision? Our paired analysis also found that, on a participant level, individuals' decision consistency in groups was within 6% of their decision consistency when acting alone. If deliberation causes individuals to change their minds, would we not have expected a greater difference in behavior from the paired analysis?

It turns out that the same individuals who were swayed before are likely to be swayed again. Nearly a quarter (23%) of all votes that changed between the pre-survey and official vote were repeated — that is, a participant changed their mind in both group rounds. Thus, one source of explanation for group consistency is that, when a group reconvenes, the same social dynamics repeat — even in the absence of the prior social context. Dominating jurors state their opinions strongly; more docile followers change their votes. In this way, group decisions are consistent (the voting patterns, indeed, are identical), and individual decisions are also consistent (the same people who changed their votes do so again).

Comments from the survey corroborate this effect, suggesting that some participants with strong personalities tended to repeat their stubborn manner:

> *I think with the group I was extra stubborn in my answers because I felt attacked and wanted to prove others wrong.*

Meanwhile, others reported a tendency to lie low and agree with the majority:

> *I felt personally that when I was working with the group I was more likely to want to sway my opinion to meet the groups beliefs.*
>
> *I think in groups, people feel pressured to not make waves and go with group think regardless of how they truly feel. Herd mentality, it's a thing. Peer pressure etc.*
>
> *I think once an individual has a answer or feeling in their mind, it is very difficult to move them off of it. As a group, I think people are more likely to go with the majority to keep harmony in the group.*

5.1.2 Integrating Information Unevenly

Repeating similar social dynamics is not sufficient to fully explain group consistency. Nor can the consistency be explained via a learning effect, as we had shown in a pre-test that participants' subsequent decisions are not influenced by the cases they viewed first ($p = .803$; see Section 3.1.5).

Thus, a second potential cause stems from information integration theory, which predicts that group deliberation causes jurors to be less dominated by personal dispositions and more focused on objective facts surfaced by the discussion (Kaplan & Miller, 1978). However, the literature on information integration is quite mixed; the process of integrating information is often complicated by arbitrary factors, from the order in which perspectives are stated aloud (Sunstein et al., 2014) to group members' status characteristics (Marder, 2005). Indeed, prior research has found that both the magnitude and direction of the effect are highly dependent on the quality of individual participants. Deliberation tends to amplify both the best and worst tendencies of individuals: a group comprised of generally unbiased individuals tends to become highly-performing and more objective, while a group comprised of poor and biased performers results in even worse group performance (Kerr et al., 1996).

Since our participant pool did not filter for individual competence, it is thus unsurprising that we found such a weak effect for in-round voting patterns between the two conditions. Future studies should attempt to separate the deliberation outcomes by individuals' level of competence, in order to investigate two open questions:

1. Are groups that are more competent in information integration also more consistent? Presumably, the individuals in these groups would be more effective in stripping away personal biases and identifying core facts. They seem likely to be more consistent as a result.

2. Are groups that are less competent in information integration also consistent? On the one hand, these groups are easily vulnerable to arbitrary features of the discussion; on the other, they may consistently revert to the same individual biases. Such groups could be more vulnerable to blindly following the foreperson — which may result in a low-quality, but nevertheless consistent, verdict.

5.1.3 Managing Minority Opinions

Our study reaffirmed prior theories that juries often tend to conform to the pre-existing majority (Thorne et al., 2004; Hackman & Katz, 2010), and that, after deliberation, the size of the majority tends to increase, following group polarization (Sunstein, 2000). However, we also observed that, compared to nominal teams, deliberating teams concluded in favor of the pre-existing majority less frequently (80%, as opposed 86% for nominal teams). Unlike prior work, which found that the majority nearly always won out (Hastie et al., 1983), and that juries conformed in 90% of cases (Devine et al., 2001), minority opinions in our experiment were more influential than expected.

While dissenters will always exist, particularly when adjudicating divisive cases, we find that a deliberative process is comparatively more effective at integrating minority voices. Although some, like **youngBison**, were outvoted in the jury deliberation, dissenters in the nominal condition were not only outvoted, but also siloed: their influence was limited to their votes alone, rather than extended by their persuasion.

Still, deliberation is imperfect. In line with prior work (Sunstein, 2000), some individuals voted with the majority but privately held different opinions. We found that 0.2 individuals per round ($\sigma = 0.447$) were members of the majority faction during the voting period, but expressed private disagreement in the post-deliberation survey. These results also echo participants' comments, expressing a desire to "not make waves" and prioritize group unity over personal beliefs.

5.1.4 Hope for democracy: Juries are much more competent than perceived.

Ultimately, our finding that jury decisions are consistent proves hopeful for democracy. The consistency of jury verdicts is statistically indistinguishable from those of individuals. Thus, tapping into the procedural benefits of juries, such as generating a sense of legitimacy in the decision-making process (Fan & Zhang, 2020; Noveck, 2004), adapting laws to better adhere to community norms (Flango, 2016), and serving as a space for participatory civic education (Marder, 2005) do not come at a cost of either accuracy (Miner Jr, 1984) or consistency.

However, we also find that, though jury decisions are consistent, *people do not expect them to be*. Indeed, the very individuals who participated in the juries did not seem to realize that they had themselves made consistent decisions on juries. Participants were far worse at predicting jury

consistency than they were at predicting individual consistency. The perception that juries are fundamentally more biased than individuals seems to be ingrained deeply into folk psychology: an early study by MacCoun and Tyler found that, despite general support for juries, the public perceives criminal juries as frequently prone to error (1988).

These perceptions have concrete implications for deciding whether to choose a trial by jury at all. In civil suits, for example, the perception of juror bias has driven litigants to choose bench trials over trial by jury, even when empirical studies have found no basis for such biases (Clermont & Eisenberg, 1991).

While juries are capable of consistent decisions, then, the perception of bias may lead some to shy away from jury decisions. Introducing jury decision-making into governance must therefore account for not only the factual consistency of verdicts, but also the ability to justify the decision-making process to skeptics who see jury decisions as fundamentally inconsistent.

5.2 Limitations

A major limitation in our work is the constraints that the online platform and repeated interaction format imposed on ecological validity. Notably, our jury deliberations were much briefer, and stripped of the many social signals of operating in a shared room with others. Typical jury deliberations occur over several days, during which members of the jury can slowly persuade one another. In-person juries have access to a richer set of polling procedures, including a nonverbal show of hands, a blind vote (which can be initiated and conducted repeatedly throughout the deliberation), a verbal go-around, and verbal expressions of dissent only (Hastie et al., 1983). Our online system does not support the full range these versatile actions, as the timing of the poll was preset, and there was no virtual equivalent of body language. Ultimately, however, we do not believe that the online format in any way invalidates our findings. Prior work has shown that online deliberation is an effective and powerful tool (Price, 2009); online juries have been adapted for mock trial use by lawyers (Marder, 2006), and online dispute resolution is becoming increasingly commonplace (Hodson, 2019).

Additionally, we note that our jury deliberations were very brief (only 7 minutes long); leading to three limitations in the results. First, as noted before, the time constraints may have artificially cut off deliberation or push groups towards conformity faster. Although visual inspection of chat logs shows high-quality deliberations, participants may have been influenced by the presence of the countdown timer to prioritize achieving consensus over thorough consideration.

Second, the early end to deliberations made it impossible to require unanimity within the jury, as one would realistically expect of a 6-member jury. While we instructed the jury to attempt to reach consensus, we did not computationally enforce consensus. Our results reveal that a large number of the resulting decisions were, in fact, not unanimous. However, our analysis indicates that our conclusions are not significantly altered by the unanimity of decisions: a Z test between individual

and group consistency, filtered only for cases in which both group decisions were unanimous (25 juries), was also statistically insignificant ($z = -0.8489, p = 0.396$). However, since the sample of unanimous juries is relatively small, further work is required to study the differences in consistency between unanimous and non-unanimous juries.

Third, in order to accommodate the brevity of deliberation time, the Reddit-sourced cases were also far less detailed and laden with fewer legal concepts than a typical court case. For a follow-up study, we have hired lawyers to summarize a variety of courtroom transcripts sourced from PACER. We intend to replicate our findings with the professionally-summarized court cases to alleviate concerns with the generalizability of our findings.

5.3 Conclusion

Consistency lies at the very heart of what makes the justice system just. When two individuals commit the same crime, yet are judged differently, outrage ensues. However, very little work has been done thus far to measure the consistency of group and individual judgments. Even as hundreds of thousands of cases are tried by jury each year, juries are merely assumed to be consistent.

This thesis proves that they are.

Using an online jury, we "erased history" by assigning participants one-way pseudonyms. This enabled jurors to effectively begin deliberation anew, and allowed us to directly compare a group's decision to its own verdict on a similar issue. Participants in the study experienced four rounds — two individual, two group — and adjudicated two paired cases. We compared the decision consistency of the two group rounds to that of the two individual rounds.

In all, groups were 64.8% consistent, and individuals were 65.9% consistent, a statistically indistinguishable difference. Though group deliberation often led individuals to sway from their initial judgements, verdicts across analogous cases was nevertheless consistent. And even as groups polarize more frequently, they manage to give minorities greater influence than in teams that merely voted without deliberation.

Still, juries are imperfect. Further work is required to understand how best to scaffold discussions in order to prevent juries from succumbing to their worst tendencies — amplifying existing biases; ignoring important information; being swayed by extraneous signals. These tendencies may make juries' decisions "consistent" in one sense, but certainly do not make their decisions any better.

That juries are consistent ultimately came as not only a surprise to us, but also cut against the collective predictions of all our participants. We found that, while participants in our online juries accurately judged individual consistency, the crowd's wisdom vastly underestimated the consistency of groups. Our findings had overturned conventional wisdom. Thus, the question becomes: even if juries are consistent, are they *perceived* to be consistent enough to be trusted?

Bibliography

Alschuler, A. W., & Deiss, A. G. (1994). A brief history of the criminal jury in the united states. *The University of Chicago Law Review*, *61*(3), 867–928.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes*, 295–303.

Ast, F. (2017). *Kleros, a protocol for a decentralized justice system.* Retrieved from `https://medium.com/kleros/kleros-a-decentralized-justice-protocol-for-the-internet-38d596a6300d`

Becker, J., Guilbeault, D., & Smith, N. (2019). Against voting? the crowd classification problem.

Bone, J., Hey, J., & Suckling, J. (1999). Are groups more (or less) consistent than individuals? *Journal of Risk and Uncertainty*, *18*(1), 63–81.

Burghardt, K., Rand, W., & Girvan, M. (2019). Inferring models of opinion dynamics from aggregated jury data. *PLOS ONE*, *14*(7). doi: https://doi.org/10.1371/journal.pone.0218312

Caci no. 5009. predeliberation instructions. (2017). Retrieved from `https://www.justia.com/trials-litigation/docs/caci/5000/5009/`

Casey, T., & Wilson-Evered, E. (2012). Predicting uptake of technology innovations in online family dispute resolution services: An application and extension of the utaut. *Computers in Human Behavior*, *28*(6), 2034–2045.

Clermont, K. M., & Eisenberg, T. (1991). Trial by jury or judge: Transcending empiricism. *Cornell L. Rev.*, *77*, 1124.

Condorcet, M. d. (1785). Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*.

Conley Tyler, M. H., & McPherson, M. W. (2006). Online dispute resolution and family disputes. *Journal of family studies*, *12*(2), 165–183.

Curley, L. J., Murray, J., MacLean, R., & Laybourn, P. (2017). Are consistent juror decisions related to fast and frugal decision making? investigating the relationship between juror consistency, decision speed and cue utilisation. *Medicine, Science and the Law*, *57*(4), 211–219.

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, *108*(17), 6889–6892.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*(345), 118–121.

Devine, D. J., Clayton, L. D., Dunford, B. B., Seying, R., & Pryce, J. (2001). Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, public policy, and law*, *7*(3), 622.

Dhami, M. K., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of behavioral decision making*, *14*(2), 141–168.

Diamond, S. S., Saks, M. J., & Landsman, S. (1998). Juror judgments about liability and damages: Sources of variability and ways to increase consistency. *DePaul L. Rev.*, *48*, 301.

Driskell, J. E., & Salas, E. (2006). Groupware, group dynamics, and team performance.

Fan, J., & Zhang, A. X. (2020, November). Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 chi conference on human factors in computing systems* (Vol. 3).

Flango, V. E. (2016). *Trends in state courts: Special focus on family law and court communications.* National Center for State Courts. Retrieved from `https://www.ncsc.org/~/media/Microsites/Files/Trends%202016/Revitalizing-the-Jury.ashx`

Glater, J. D. (2007). Race gap: Crime vs. punishment.

Gruenfeld, D. H., & Preston, J. (2000). Upending the status quo: Cognitive complexity in us supreme court justices who overturn legal precedent. *Personality and Social Psychology Bulletin*, *26*(8), 1013–1022.

Habermas, J. (1992). *Between facts and norms: Contributions to a discourse theory of law and democracy, trans. w. rehg.* Cambridge, MA: MIT Press.

Hackman, J. R., & Katz, N. (2010). Group behavior and performance.

Handbook for trial jurors serving in the united states district courts. (n.d.).

Hastie, R., Penrod, S., & Pennington, N. (1983). *Inside the jury.* Harvard University Press.

Hodson, D. (2019). The role, benefits, and concerns of digital technology in the family justice system. *Family Court Review*, *57*(3), 425–433.

Kaniovski, S., & Zaigraev, A. (2011). Optimal jury design for homogeneous juries with correlated votes. *Theory and decision*, *71*(4), 439–459.

Kaplan, M. F., & Miller, L. E. (1978). Reducing the effects of juror bias. *Journal of personality and social psychology*, *36*(12), 1443.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological review*, *103*(4), 687.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237–293.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of experimental psychology*, *89*(1), 46.

MacCoun, R. J. (1987). Getting inside the black box: Toward a better understanding of civil jury behavior.

MacCoun, R. J., & Tyler, T. R. (1988). The basis of citizen's perceptions of the criminal jury. *Law and Human Behavior*, *12*(3), 333–352.

Marder, N. S. (2005). *The jury process*. Foundation Press.

Marder, N. S. (2006). Cyberjuries: A new role as online mock juries. *U. Tol. L. Rev.*, *38*, 239.

Margolis, J., Huckaby, R. W., Odessey, B., & Hug, K. (Eds.). (2009). Anatomy of a jury trial. *eJournal USA*, *14*(7).

Miner Jr, F. C. (1984). Group versus individual decision making: An investigation of performance measures, decision strategies, and process losses/gains. *Organizational Behavior and Human Performance*, *33*(1), 112–124.

Morgan Jr, B. B., Salas, E., & Glickman, A. S. (1993). An analysis of team evolution and maturation. *The Journal of General Psychology*, *120*(3), 277–291.

Moscovici, S., & Faucheux, C. (1972). Social influence, conformity bias, and the study of active minorities. In *Advances in experimental social psychology* (Vol. 6, pp. 149–202). Elsevier.

Noveck, B. S. (2004). Democracy online. In P. M. Shane (Ed.), (pp. 21–46). Psychology Press. Retrieved from `http://dx.doi.org/10.1007/3-540-09456-9` doi: 10.1007/3-540-09237-4

Payne, J. W., Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge university press.

Pezdek, K., Avila-Mora, E., & Sperry, K. (2010). Does trial presentation medium matter in jury simulation research? evaluating the effectiveness of eyewitness expert testimony. *Applied Cognitive Psychology*, *24*(5), 673–690.

Price, V. (2009). Citizens deliberating online: Theory and some evidence. *Online deliberation: Design, research, and practice*, 37–58.

Roper, R. T. (1980). Jury size and verdict consistency:" a line has to be drawn somewhere"? *Law and Society Review*, 977–995.

Segal, J. A. (1986). Supreme court justices as human decision makers: An individual-level analysis of the search and seizure cases. *The Journal of Politics*, *48*(4), 938–955.

Shah, N. B., Tabibian, B., Muandet, K., Guyon, I., & Von Luxburg, U. (2018). Design and analysis of the nips 2016 review process. *The Journal of Machine Learning Research*, *19*(1), 1913–1946.

Son, J.-Y., Bhandari, A., & FeldmanHall, O. (2019). Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions. *Scientific reports*, *9*(1), 1–15.

Starr, S. B. (2015). Estimating gender disparities in federal criminal cases. *American Law and Economics Review*, *17*(1), 127–159.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology*, *48*(6),

1467.

Sunstein, C. R. (2000). Deliberative trouble? why groups go to extremes. *The Yale Law Journal*, *100*(71), 71-119.

Sunstein, C. R., Hastie, R., et al. (2014). Making dumb groups smarter. *Harvard Business Review*, *92*(12), 90–98.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Thorne, L., Massey, D. W., & Jones, J. (2004). "an investigation of social influence: Explaining the effect of group discussion on consensus in auditors' ethical reasoning. *Business Ethics Quarterly*, *14*(3), 525-551. doi: https://doi.org/10.5840/beq200414321

Totenberg, N. (2019). Supreme court sees 2 similar death penalty questions very differently. Retrieved from `https://www.npr.org/2019/03/30/708238203/supreme-court-sees-2-similar-death-penalty-questions-very-differently`

van Prooijen, J.-W., Gallucci, M., & Toeset, G. (2008). Procedural justice in punishment systems: Inconsistent punishment procedures have detrimental effects on cooperation. *British Journal of Social Psychology*, *47*(2), 311–324.

Watson, L. M. W., & Black, R. (1989). Realistic test of individual versus group decision making. *J. Applied Psychology*, *74*, 834–839.

Watts, D. J. (2011). *Everything is obvious:* once you know the answer*. Currency.

Weninger, T., Zhu, X. A., & Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the reddit community. In *2013 ieee/acm international conference on advances in social networks analysis and mining (asonam 2013)* (pp. 579–583).

Whiting, M. E., Blaising, A., Barreau, C., Fiuza, L., Marda, N., Valentine, M., & Bernstein, M. S. (2019). Did it have to end this way? understanding the consistency of team fracture. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–23.

Whiting, M. E., Gao, I., Xing, M., N'Godjigui, J. D., Nguyen, T., & Bernstein, M. S. (2020). Parallel worlds: Repeated initializations of the same team to improve team viability. *Proceedings of the ACM on Human-Computer Interaction*, *4*.

Whiting, M. E., Hugh, G., & Bernstein, M. S. (2019). Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the aaai conference on human computation and crowdsourcing* (Vol. 7, pp. 197–206).